Experimental Design under Network Interference *

Davide Viviano †

First Version: March, 2020
This version: October, 2021

Abstract

This paper studies the design of two-wave experiments in the presence of spillovers when the researcher's goal is to conduct precise inference on treatment effects. Consider a fully connected network, local dependence among individuals, and a general class of estimands, which encompasses average treatment and average spillover effects. The researcher optimizes over participants and treatment assignments to minimize the variance of the estimators of interest, using a first-wave (pilot) experiment to estimate the variance. I provide a statistical framework for the design of two-wave network experiments, illustrate the existence of a trade-off in the choice of the pilot study, and formally characterize the pilot's size relative to the main experiment. I derive guarantees for inference on treatment effects and regret guarantees on the variance obtained from the proposed design mechanism. Simulations illustrate the advantage of the method.

Keywords: Experimental Design, Spillovers, Two-wave experimentation, Causal Inference.
JEL Codes: C90.

*I acknowledge support of the Social Science Computing Facilities at UC San Diego and HPC@UC awards from the San Diego Super Computer Center. I thank Jelena Bradic, Graham Elliott, James Fowler, Brian Karrer, Karthik Muralidharan, Yixiao Sun, and Kaspar Wüthrich for helpful comments and discussion. All mistakes are mine.

†University of California, San Diego. Correspondence: dviviano@ucsd.edu.

1 Introduction

This paper studies the design of experiments in the presence of network interference for inference on treatment effects. We consider a setting where individuals are connected in a single network, which the researcher observes, and interact locally (i.e., between neighbors) in the network.[1] Network interference induces (i) spillovers across units and (ii) statistical dependence. Our goal is to obtain precise estimates of treatment effects. Differently from typical settings for clustered or saturation design experiments[2], no independent clusters are necessarily available. We assume that researchers have access to a pilot study (*first-wave experiment*) to estimate the variance and covariances between units and to optimally select participants and treatment assignments in the main experiment (*second-wave experiment*).[3]

Specifically, the following experimental protocol is considered: (1) researchers select a small sub-sample of individuals, and they conduct a pilot study; (2) using information from the pilot, they select *participants* and *treatment assignments* in the main experiment; (3) researchers collect information on the outcomes of the participating units. We consider a class of estimands of interest, which include as the main ones the (i) *overall* effect of treatment, the (ii) *direct* effect, the (iii) *spillover* effects and interactions of the latter two. For example, in the presence of a cash transfer program (Barrera-Osorio et al., 2011), we may be interested in the effect on recipients (i.e., direct effects), the effects on those non-recipients living close to the recipients (i.e., spillovers), and on the sum of these effects (i.e., overall effect). We consider a class of estimators linear in the observed outcomes.

Throughout the rest of our discussion, we refer to the proposed mechanism as Experiment under Local Interference (ELI). Our mechanism imposes the following

[1] This assumption is known as local interference (Manski, 2013), and it can be tested using, for instance, the framework in Athey et al. (2018). This is often assumed in practice (Egger et al., 2019; Dupas, 2014; Miguel and Kremer, 2004; Bhattacharya et al., 2013; Duflo et al., 2011) as well as in theoretical analysis (Forastiere et al., 2016; Leung, 2019; Sinclair et al., 2012).

[2] See e.g., Baird et al. (2018).

[3] Pilot studies are common practice and some examples include Karlan and Appel (2018); Karlan and Zinman (2008); DellaVigna and Pope (2018).

conditions: (i) interference and dependence is local and anonymous;[4] (ii) effects may be heterogeneous in summary statistics of the network structure, such as the number of neighbors or centrality measures having discrete support.[5]

We contribute to the literature with the first statistical framework for the design and inference of a two-wave experiment (pilot and main experiment) under network interference. We show that since individuals are dependent, the main experiment is unconfounded if individuals in the pilot and *their neighbors* are not participants in the main experiment, but not necessarily otherwise. This restriction on the choice of the pilot induces the following trade-off: a larger pilot guarantees more precise variance estimators, which are useful to design a better second-stage experiment, but it imposes stricter restrictions on the design of the main experiment. We select participants by solving a variation of the minimum cut problem in the network.

We study the choice of the pilot and main experiment also from a theoretical perspective. We characterize the rate of convergence of the difference between the variance of the two-wave experiment and the variance of the "oracle" experiment, which assigns treatment and participation indicators to minimize the *true* variance of the estimator (i.e., regret). The regret converges to zero at a rate that depends on (a) the ratio of the size of the first and second wave experiments, (b) the convergence rate of the estimated variance and covariance obtained from the pilot study, which depends inversely on the pilot's size, and (c) the maximum degree of the network. The first component illustrates the "price to pay" for the selection of a larger pilot study due to stricter restrictions on the main experiment. The second component is informative on the role of the pilot to obtain a precise first-stage estimator of the variances. We provide an explicit characterization of the rate and of the pilot's size that minimize the regret bound. Finally, we derive asymptotic properties of the estimator under the proposed design to conduct inference on treatment effects.

We conclude our discussion with a set of simulation results. We show that the

[4]These conditions encompass a large number of economic examples from the literature: spillovers in public policy programs (Muralidharan et al., 2017), cash transfer programs (Egger et al., 2019), health programs (Dupas, 2014), educational program (Duflo et al., 2011).

[5]In the presence of continuous centrality measures, discretization is necessary for the validity of the results.

proposed method significantly outperforms state-of-art competitors for estimating overall treatment effects as well as spillover and direct effects, especially in the presence of heteroskedastic variances and covariances. In the Appendix, we study extensions in the presence of partially observed networks, where only information on participants' neighbors is available.

This paper connects to the recent literature in statistics and econometrics, which studies design mechanisms in the presence of interference, but *without* using information from a pilot study for inference on treatment effects. Here, we show that the presence of the pilot study can be useful to improve precision. References include clustered experiments (Eckles et al., 2017; Taylor and Eckles, 2018; Ugander et al., 2013) and saturation design experiments (Baird et al., 2018; Basse and Feller, 2016; Pouget-Abadie, 2018), which often assume clustered observations. Additional references are Basse and Airoldi (2018b), who also assume Gaussian outcomes and lack of spillover effects; Wager and Xu (2019) who study sequential randomization for optimal pricing strategies under global interference, without discussing the problem of inference on treatment effects; Kang and Imbens (2016) who study encouragement designs, without focusing on the problem of variance-optimal design. Basse and Airoldi (2018a) discuss limitations of design-based causal inference under interference; Jagadeesan et al. (2017) and Sussman and Airoldi (2017) study the design of experiments for estimating direct treatment effects only, while this paper considers a more general class of estimands, which may include overall and spillover effects.

We relate to a large literature on experimental design in the *i.i.d.* setting for batch experiments, which can be divided into one-stage procedures (Harshaw et al., 2019; Kasy, 2016; Kallus, 2018; Barrios, 2014), and two-stage procedures (Bai, 2019; Tabord-Meehan, 2018). However, none of the above references study the problem under network interference.

We more broadly connect to the literature on treatment effects under network interference, which include Aronow and Samii (2017), Hudgens and Halloran (2008), Forastiere et al. (2016), Manski (2013), Leung (2019), Vazquez-Bare (2017), Athey et al. (2018), Goldsmith-Pinkham and Imbens (2013), Sävje et al. (2017), Ogburn et al. (2017), Kitagawa and Wang (2020) among others. None of these references

study the problem of experimental design (and variance optimal design).

Finally, in two separate papers, I study the problem of choosing whom to treat from different angles. Viviano (2019) provides small sample guarantees for policy targeting under interference, but without studying the problem of experimental design and inference. Viviano (2020) studies instead experimental design without focusing on the variance-optimal design and under a different set of assumptions where the network is unobserved.

The remainder of the paper is organized as follows: Section 2 introduces the problem; in Section 3, we discuss the design mechanism; in Section 4, we derive theoretical guarantees; Section 5 contains the numerical results and Section 6 concludes. The Appendix contains extensions and derivations.

2 Set up

In this section, we discuss the setup, model, and estimands.

2.1 Notation

We consider the following setting: N units are connected by a binary adjacency matrix $A, A_{i,j} \in \{0,1\}$, with $\mathcal{N}_i = \{j : A_{i,j} = 1\}$ denoting the neighbors of individual i. The adjacency matrix is observed by the researcher. Throughout our discussion, we will condition on A (i.e., A is non-random) unless otherwise specified. The researcher conducts two experiments, a pilot and the main experiment. For each unit $i \in \{1, \dots, N\}$ we denote

$$R_i = 1\Big\{i \text{ is in the main experiment}\Big\}, \quad P_i = 1\Big\{i \text{ is in the pilot experiment}\Big\},$$

respectively the participation indicator variable in the main experiment, and in the pilot, with $\sum_{i=1}^N R_i = n \in (n_1, n_2), \sum_{i=1}^N P_i = m$. Each unit i is associated with an outcome, pre-treatment observables, and binary assignment (Y_i, T_i, D_i), respectively. Here, T_i may depend on the network information (e.g., T_i denote the number

of neighbors, $T_i = |\mathcal{N}_i|$). The binary assignment is not reversible, i.e., whenever researchers assign $D_i = 1$ to pilot's units, their treatment status cannot be changed in the main experiment. See Section 3 for a comprehensive discussion. We denote

$$\mathbf{R} = \Big\{R_i, i \in \{1, \ldots, N\}\Big\}, \quad \mathbf{T}^R = \Big\{T_i, i : R_i = 1\Big\}, \quad \mathbf{D}^R = \Big\{D_i, i : R_i = 1 \text{ or } R_{\mathcal{N}_i} = 1\Big\},$$

the vector of selection indicators of each individual, observable statistics of participants, and the vector of treatment assignments of all individuals and their neighbors, respectively. Similarly, $\mathbf{P}$ denote the selection indicators of individuals in the pilot, while $\mathbf{T} = (T_1, \cdots, T_N)$ denotes the covariates of all the individuals in the population.

2.2 Model and Dependence

We consider the following outcome model.

Assumption 2.1 (Potential outcomes)**.**

$$Y_i = r\Big(D_i, \sum_{k \in \mathcal{N}_i} D_k, T_i, \varepsilon_i(\mathbf{D})\Big), \varepsilon_i(\mathbf{d}) | A, T_1, \cdots, T_N \sim \mathcal{P}, \text{ for all } i \in \{1, ..., N\}, \mathbf{d} \in \{0,1\}^N, \tag{1}$$

where the function $r(\cdot)$ and $\mathcal{P}$ are potentially unknown to the researcher, and $\varepsilon_i(\mathbf{d}) = \varepsilon_i(\mathbf{d}')$ for all $\mathbf{d}, \mathbf{d}' \in \{0,1\}^N$.

The above model assumes that (i) individuals only depend on neighbors' treatment assignments (and hence $\varepsilon_i(\mathbf{d})$ is a constant function in $\mathbf{d}$); (ii) the network affects the outcome variable through observables T_i only. Note that the assumption does not impose restrictions on the distribution of $\mathbf{D}$ which is left unspecified. Since $\varepsilon_i(\mathbf{d})$ is a constant function in $\mathbf{d}$, throughout or discussion, we will write unobservables as ε_i, omitting its argument.

Throughout the rest of our discussion, we denote

$$\mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big)\Big| T_i = l\Big] = m(d, s, l), \tag{2}$$

the conditional mean given $T_i = l$, and fixing the individual and neighbors' treatment

assignments to be respectively (d, s). The model follows similarly to Leung (2019) whenever $T_i = |\mathcal{N}_i|$ denotes the number of neighbors.[6]

Example 2.1. Sinclair et al. (2012) study spillover effects for political decisions within households. The authors propose a model of the form

$$Y_i = \mu + \tau_1 D_i + \tau_2 1\Big\{\sum_{j\in\mathcal{N}_i} D_j \geq 1\Big\} + \tau_3 1\Big\{\sum_{j\in\mathcal{N}_i} D_j \geq |\mathcal{N}_i|/2\Big\} + \tau_4 1\Big\{\sum_{j\in\mathcal{N}_i} D_j = |\mathcal{N}_i|\Big\} + \varepsilon_i, \tag{3}$$

where $\mathcal{N}_i$ denotes the element in the same household of individual i. The model captures the effect for individual i being treated and at least one, half and all of the other units in the household being treated. □

Example 2.2. Consider the following equation

$$Y_i = \mu + \tau_1 D_i + \tau_2 \sum_{k\in\mathcal{N}_i} D_k \Big/ |\mathcal{N}_i| + \varepsilon_i.$$

Then the above assumption holds with $T_i = |\mathcal{N}_i|$. □

We allow the observables (T_i, D_i) to have arbitrary dependence, and, hence, we do not impose conditions on their dependence. We instead impose restrictions on the dependence structure of unobservables ε_i. Namely, we assume that unobservables exhibit one-degree dependence. We leave extensions to higher-order degree dependence to Appendix A.2.

Assumption 2.2 (Dependence)**.** Assume that for all $i \in \{1, ..., N\}$, almost surely

$$\Big\{\varepsilon_i, \{\varepsilon_k\}_{k\notin\mathcal{N}_j, j\in\mathcal{N}_i}\Big\} \perp \{\varepsilon_j\}_{j\notin\mathcal{N}_i} \Big| A, T_1, \cdots, T_N$$
$$(\varepsilon_i, \varepsilon_j) =_d (\varepsilon_{i'}, \varepsilon_{j'}) \Big| A, T_1, \cdots, T_N \text{ for all } (i, j, i', j') : i \in \mathcal{N}_j, i' \in \mathcal{N}_{j'}, T_i = T_{i'}, T_j = T_{j'}.$$

The above condition states that unobservables of non-adjacent neighbors are mutually independent. The condition also imposes that individuals are dependent if

[6] Examples in applications include Muralidharan et al. (2017); Sinclair et al. (2012); Cai et al. (2015); Duflo et al. (2011), among others.

(i, j) are neighbors. It also allows the joint distribution of unobservables to depend on observables (T_i, T_j).

Remark 1 (Higher order dependence)**.** Extensions to higher order dependence of degree M, reads as follows

$$\left\{\varepsilon_i, \{\varepsilon_k\}_{k \notin \cup_{u=1}^{M} \mathcal{N}_j^u, j \in \cup_{u=1}^{M} \mathcal{N}_i^u}\right\} \perp \{\varepsilon_j\}_{j \notin \cup_{u=1}^{M} \mathcal{N}_i^u} \Big| A, T_1, \cdots, T_N, \tag{4}$$

where $\mathcal{N}_i^u$ denotes the set of neighbors of degree u. In this case, unobservables that are not adjacent by at least M edges are independent. All our results extend to this setting, as discussed in Section A.2. □

2.3 Estimands

We consider a *class* of estimands of interest which may capture the direct, spillover, or overall effects of the treatments. We introduce the estimands below.

Definition 2.1 (Estimands)**.** For some known weights $v(l)$, consider a class of estimands

$$\mathcal{C} = \Big\{\tau_v(d, s, d', s') \text{ for some } d, d' \in \{0, 1\}, s, s' \in \mathbb{Z}, v : \mathbb{Z} \mapsto [0, 1]\Big\},$$
$$\text{where } \tau_v(d, s, d', s') = \sum_{l \in \mathcal{T}} v(l)\Big[m(d, s, l) - m(d', s', l)\Big].$$

Definition 2.1 defines a class of estimands of interest (which may contain more than one estimand). These estimands define the difference of potential outcomes' conditional expectations, averaged by some weights $v(l)$. Here, potential outcomes are exposed to individual treatments (d, d') and neighbors' treatments (s, s'). We average over the heterogeneity $T_i = l$ using some arbitrary weights $v(l)$, which are known to the researcher. We provide two examples below.

Example 2.3 (Direct and Overall Treatment Effect)**.** Let $T_i = |\mathcal{N}_i|$ denote the number of neighbors. Then $\Big[m(1, l, l) - m(0, 0, l), m(1, 0, l) - m(1, 0, l)\Big]$ define the effect of treating all individuals compared to the one of treating none, and the direct

effect of treating one single unit for those individuals with l many friends, respectively. By letting $v(l) = \sum_{i=1}^{N} 1\{|\mathcal{N}_i| = l\}/N$,

$$\sum_{v=0}^{\infty} v(l)\Big[m(1,l,l) - m(0,0,l)\Big], \quad \sum_{v=0}^{\infty} v(l)\Big[m(1,0,l) - m(0,0,l)\Big]$$

denote their expected value, averaged over the number of friends. □

In the following condition, we impose a restriction on the estimands and assume that $\tau_v(d,s,d',s')$ takes a linear representation in the outcomes' conditional expectations. This is formalized below.

Assumption 2.3. For any $(\mathbf{R}, \mathbf{D}^R, \mathbf{T}^R), \tau_v(d,s,d',s') \in \mathcal{C}$, there exists a known function $w_N \in \mathcal{W}_N$ (which may depend on v, d, s, d', s'), such that

$$\begin{aligned} &\tau_v(d,s,d',s') = \Gamma_n(w_N), \\ &\text{where } \Gamma_n(w_N) = \frac{1}{n}\sum_{i=1}^{N} R_i w_N\Big(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R\Big) m\Big(D_i, \sum_{k\in\mathcal{N}_i} D_k, T_i\Big). \end{aligned} \tag{5}$$

Assumption 2.3 states that we can write each target estimand as a weighted average of individual conditional expectations $m(\cdot)$, where the weights depend on treatment assignments, selection indicators, and covariates. Two leading examples illustrate settings where the assumption holds.

Example 2.4 (Difference in conditional expections)**.** Let $T_i = |\mathcal{N}_i|$ and consider the following class of weights:

$$w_N\Big(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R\Big) = \begin{cases} \frac{I_i(d,s,l)}{\sum_{j:R_j=1} I_j(d,s,l)/n} - \frac{I_i(d',s',l)}{\sum_{j:R_j=1} I_j(d,s,l)/n} \text{ if } R_i = 1 \\ 0 \text{ otherwise.} \end{cases}$$

where $I_i(d,s,l) = 1\Big\{D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s, T_i = l\Big\}$. Then $\Gamma_n(w_N) = m(d,s,l) - m(d',s',l)$, and similarly for any weighted combination, weighted by $v(l)$. □

Example 2.5 (Linear model)**.** Let $T_i = |\mathcal{N}_i|$. Consider the following *vector* of

weights:

$$\mathbf{w}_N(i,.) = \begin{cases} \left(\frac{1}{n}\sum_{i:R_i=1}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\mathbf{X}_i, \text{ if } R_i = 1 \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

where, $\mathbf{X}_i = \left(1, D_i, \sum_{k\in\mathcal{N}_i} D_k/|\mathcal{N}_i|\right)$. By further assuming that $m(d,s,l) = \mu + d\beta + \gamma s/l$, we have $\left[\Gamma_n(\mathbf{w}_N^{(2)}), \Gamma_n(\mathbf{w}_N^{(3)})\right] = (\beta, \gamma)$, where β is the direct effect of the treatment and γ the spillover effect of treating all the neighbors. □

Assumption 2.3 shows that our setting applies to (generic) estimands from linear models. Our setting also allows for non-linearities (Example 2.4), with the weights taking a difference in means between individuals exposed to different treatments. Under Assumption 2.3, we will equivalently refer to the class of estimands as

$$\left\{\Gamma_n(w_n), w_n \in \mathcal{W}_N\right\},$$

indexed by the weights w_N.

3 Two-wave Experiment

In this section, we discuss the experimental protocol in Algorithm 1 (pilot) and Algorithm 2 (main experiment). We can summarize the algorithm as follows

- Researchers observe A and $T_1, \cdots, T_N$ for all individuals;
- Researchers select a subset of individuals and collect information

$$\left\{i \in \{1,\cdots,N\} : P_i = 1\right\}, \quad \left[P_i\left(Y_i, D_i, T_i, D_{\mathcal{N}_i}\right)\right]_{i=1}^{N}$$

who participate in the pilot study. They collect outcomes and treatment assignments and neighbors assignments of all individuals in the pilot. The treatments assigned to the neighbors of the individuals in the pilot who are not in the pilot are constant at zero. Researchers choose individuals in the pilot with the least number of connections to the remaining units, under the constraint that some

individuals also have their neighbors selected in the pilot study (see below for a discussion).

- Researchers use information from the pilot study to select the participants (i.e., indicators R_i) and the treatment assignments D_i for individuals participating in the experiment and their neighbors. Researchers do not select in the main experiment the participants in the pilot *and their neighbors*. The treatment D_i for those units who do not participate in the experiment (and pilot) is assumed to be constant at zero, and the treatment assignment to the pilot units remains unchanged;

- Researchers collect information

$$\left[R_i\Big(Y_i, D_i, T_i, D_{j\in\mathcal{N}_i}, \mathcal{N}_i\Big)\right]_{i=1}^{N};$$

- For each weight (and estimand) $w_N \in \mathcal{W}_N$ as in Assumption 2.3, researchers estimate treatment effects as follows:

$$\widehat{\Gamma}_n(w_N) = \frac{1}{n}\sum_{i=1}^{N} R_i w_N\Big(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R\Big) Y_i. \tag{7}$$

  The weights $w_N(.)$ are functions of *observable* characteristics, i.e., treatment assignments, selection indicators, individual observables T_i, and outcomes of the participants.

The complete algorithm is illustrated in Algorithm 1, 2, which we discuss in details below.

3.1 Unconfoudedness and Selection of the Pilot

Before discussing the details of Algorithm 1 and 2, we note that the most important aspect of Algorithm 2, is that individuals in the pilot study and their neighbors are not selected in the main experiment. In this section, we show why this condition is imposed and how it plays an important role in selecting the pilot study.

Algorithm 1 Pilot study

Require: A, $T_1, \cdots, T_N$ observed by researchers, constants $\delta, \bar{C}$.

1: Select the participants in the pilot study

$$\begin{aligned}(P_1, \cdots, P_N) \in \arg \min_{(p_1,\cdots,p_N)\in\{0,1\}^N} \sum_{i=1}^{N} \sum_{j\in\mathcal{N}_i} p_i(1-p_j), \\ \text{such that } \sum_{i=1}^{N} p_i = m, \sum_{i=1}^{N} p_i \sum_{j\in\mathcal{N}_i} p_j \geq \delta.\end{aligned} \tag{8}$$

2: Assign treatments to the units in the pilot as

$$D_i \sim \begin{cases} \text{Bern}(1/2) \text{ if } P_i = 1, \\ 0 \text{ otherwise.} \end{cases}$$

3: Collect information from the pilot: $\left[P_i\left(Y_i, D_i, T_i, D_{\mathcal{N}_i}\right)\right]_{i=1}^{N}$.

4: Estimate $\left(\sigma^2(\cdot), \eta(\cdot)\right)$ and return the estimator $\left(\widehat{\sigma}_p(\cdot), \widehat{\eta}_p(\cdot)\right)$.

return P, $\left(\widehat{\sigma}_p(\cdot), \widehat{\eta}_p(\cdot)\right)$.

Specifically, if we were to select pilot units or their neighbors in the main experiment, we may lead to confounded experimentation. To gain further intuition, consider Figure 1. In the figure, the set of pilot units includes the vertices N4, N5, N6. Researchers may use their outcomes for the design of the main experiment. As a result, the treatment assignment mechanism is dependent on the unobservables of the pilot units. However, since the pilot units are statistically dependent on their neighbors (N7), selecting N7 may depend on the treatment assignment mechanism and unobservables in the main experiment and confound the experiment.

Formally, let us define

$$\mathcal{J} = \left\{i \text{ and } \mathcal{N}_i : P_i = 1\right\} \tag{9}$$

the set of units after excluding individuals in the pilot study and the corresponding

neighbors. Then the experiment is unconfounded if it satisfies the following restrictions.

Proposition 3.1 (Sufficient Conditions for Unconfounded Experiment)**.** *Suppose that*

(a) $\varepsilon_{i\in\{1,\cdots,N\}\setminus\mathcal{J}} \perp \left(\mathbf{D}^R, \mathbf{R}\right) \Big| A, T_{1,\cdots,N}, \mathbf{P}$*;*

(b) $\varepsilon_{i\in\{1,\cdots,N\}} \perp \mathbf{P} \Big| A, T_1, \cdots, T_N$*;*

(c) $R_i = 0$ *for all* $i \in \mathcal{J}$*.*

Then under Assumption 2.1, 2.2

$$\mathbb{E}\left[\hat{\Gamma}_n(w_N) \Big| \mathbf{D}^R, \mathbf{R}, A, T_1, \cdots, T_N, \mathbf{P}\right] = \Gamma_n(w_N).$$

The proof is in the Appendix. Proposition 3.1 provides sufficient conditions that guarantee that the estimator is unbiased, *conditional* on the treatment assignment mechanism.

The first condition states that unobservables in the set $\{1, \cdots, N\} \setminus \mathcal{J}$ are independent of treatment assignments and selection indicators. The second condition states that the choice of the pilot units is randomized, depending on information $A, T_{1,\cdots,N}$ only. The third condition states that the units participating in the experiment have not been previously selected in the pilot *and* they are not the neighbors of units in the pilot.

Condition (a) in Proposition 3.1 is imposed on all units, except those units in the pilot study and *their neighbors*. Condition (c) states that the participants in the main experiment are neither units participating in the pilot nor their neighbors.

Proposition 3.1 also provides insights on the choice of the pilot units. In particular, the choice of the pilot must take into account two facts: (i) under Proposition 3.1, participants in the pilot study can be selected based on network information only; (ii) the larger the set $\mathcal{J}$ (i.e., pilot study and the neighbors of the units in the pilot), the stricter the constraint imposed on the second-wave experiment.

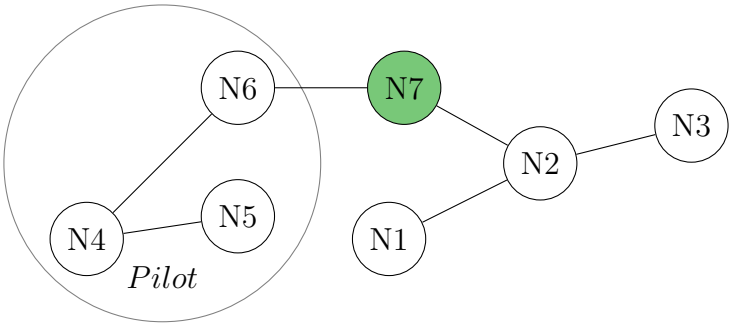

Figure 1: Example of network. In such a setting, under one-degree dependence, N7 does not satisfy the validity condition since it connects to the pilot study, which is used for the randomization of treatments and indicators.

We can now study Algorithm 1 and 2. Algorithm 1 chooses units in the pilot to minimize $|\mathcal{J}|$. It also imposes one additional restriction: participants in the pilot must have some *neighbors* in the pilot study. The latter restriction guarantees that we can identify and estimate covariances between individuals from the pilot study. We then randomize treatments in the pilot study exogenously.

The optimization problem (Equation (8)) reads like a variation of the min-cut problem in a graph: we find a set of units that are "well" separated from the rest under constraints on the number of units and their neighbors be included in the pilot study. The optimization can be solved using mixed-integer quadratic programming (MIQP).

The following corollary holds.

Corollary. *Let Assumption 2.1 hold. Then the two-wave experiment constructed with Algorithm 1 and Algorithm 2 satisfies the conditions in Proposition 3.1.*

The above corollary illustrates that the proposed algorithm returns unbiased estimators of treatment effects.

Given the selected individuals in the pilot and the treatment assignments, we use information from the pilot study to estimate the variance of the estimator of interest. We are interested in estimating

$$\mathrm{Var}\Big(Y_i\Big|A, D_i, T_i, D_{\mathcal{N}_i}, \mathbf{P}\Big), \quad \mathrm{Cov}\Big(Y_i, Y_j\Big|A, D_i, D_j, D_{\mathcal{N}_i}, D_{\mathcal{N}_j}, T_i, T_j, \mathbf{P}\Big)$$

which denote the variance and covariance, conditional on the adjacency matrix, individual type and pilot choice. The following lemma permits us to identify and estimate the above variance functions.

Lemma 3.2. *Suppose that Assumption 2.1, 2.2 hold and the pilot is chosen as in Algorithm 1. Then for all individuals* $(i,j) \in \mathbf{P}$,

$$\mathrm{Var}\Big(Y_i \Big| A, D_i, T_i, D_{\mathcal{N}_i}, \mathbf{P}\Big) = \sigma^2\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k\Big)$$

$$\mathrm{Cov}\Big(Y_i, Y_j \Big| A, D_i, D_j, D_{\mathcal{N}_i}, D_{\mathcal{N}_j}, T_i, T_j, \mathbf{P}\Big) = \begin{cases} \eta\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k, T_j, D_j, \sum_{k\in\mathcal{N}_j} D_k\Big) & \text{if } i \in \mathcal{N}_j \\ 0 \text{ otherwise.} \end{cases}$$

for some functions $\sigma^2(\cdot), \eta(\cdot)$.

The proof is in the Appendix. Lemma 3.2 guarantees identification (and estimation) of the variance and covariance pilot's participants. We denote $\Big(\widehat{\sigma}_p^2, \widehat{\eta}_p\Big)$, the variance and covariance function estimated from the pilot study. We impose high-level conditions on the rate of convergence of these functions in Assumption 4.3.

3.2 Design of the Main Experiment

Next, we discuss the main experiment (Algorithm 2). We define the conditional variance of the estimators of interest, conditional on the treatment assignment and the underlying network below.

$$V_N\Big(w_N; A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R, \mathbf{P}\Big) = \mathrm{Var}\Big(\frac{1}{1^\top \mathbf{R}} \sum_{i:R_i=1} w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R) Y_i \Big| A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R, \mathbf{P}\Big). \tag{10}$$

Given the selection of the pilot study $\mathbf{P}$, the design of the *main* experiment (i.e., second-wave experiment) minimizes the worst case over the class of estimands (indexed by $w_N \in \mathcal{W}_N$) estimated variance $\hat{V}_{N,\hat{\sigma}_p,\hat{\eta}_p}$. Here, $\hat{V}_{N,\hat{\sigma}_p,\hat{\eta}_p}$ denotes the plug-in estimator of the variance using the estimated functions $\widehat{\sigma}_p, \widehat{\eta}_p$ obtained from the pilot

study. Formally, letting $n = 1^\top \mathbf{R}$,

$$\hat{V}_{N,\hat{\sigma}_p,\hat{\eta}_p}(w_N) = \frac{1}{n^2} \sum_{i:R_i=1} w_N^2(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R) \hat{\sigma}_p^2\Big(T_i, D_i, \sum_{k \in \mathcal{N}_i} D_k\Big)$$
$$+ \frac{1}{n^2} \sum_{i:R_i=1} \sum_{j \in \mathcal{N}_i} R_j w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R) w_N(j, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R) \hat{\eta}_p\Big(T_i, D_i, \sum_{k \in \mathcal{N}_i} D_k, T_j, D_j, \sum_{k \in \mathcal{N}_j} D_k\Big). \tag{11}$$

Here, the estimated variance takes into account the (heteroskedastic) variance component $\hat{\sigma}_p^2$ of each unit, estimated using information from the pilot, and the covariances between neighbors.[7]

The optimization problem is in Equation (14). The minimization is with respect to two sets of choice variables: the participation indicators and the treatment assignments. Intuitively, different participants have heteroskedastic variances and a different number of neighbors (and covariances). The optimization problem selects between n_1 and n_2 individuals. We also restrict the weights to be finite and smaller than some arbitrary constant $\bar{C}$. This guarantees that the optimization problem does not return ill-posed solutions. For example, for the weights corresponding to the difference in means as in Example 2.4, the bound on the weights implies that at least some individuals are treated, and some are not.

Additional constraints may be included: for example, only some units can participate in the experiments, corresponding to constraints on $R_i = 0$ for some of the units. An alternative constraint is to impose $D_i \times R_i \geq D_i$. This constraint imposes that those units which are not selected as participants have treatment assignments equal to zero. This is omitted for brevity only, and our results also hold in these scenarios.[8]

The constraint in Equation (14) illustrates the *trade-off* in the selection of the pilot study: the larger the pilot study, the more precise the estimator of the variance. However, the larger the pilot study, the larger the set $\mathcal{J}$ and therefore, the more

[7] In the presence of higher-order interference, we also add additional components which depend on higher-degree neighbors as discussed in Appendix.

[8] Also, note that an alternative specification of the objective function consists of minimizing a weighted combination of the variances of each estimator. Formally, minimizing $\sum_{w_N \in \mathcal{W}_N} u(w_N) V_N(w_N; A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R, \mathbf{P})$ for some given weights $u(w_N)$. The proposed mechanism and all our results directly extend also to this setting.

stringent the constraint imposed in the above optimization procedure.

Finally, Algorithm 2 also returns estimates of the variance and covariances $\hat{\sigma}, \hat{\eta}$ using information from the main experiment only. Such estimates will be used for inference on treatment effects. Variance and covariances can be identified in the main experiment, as illustrated in the following lemma.

Lemma 3.3. *Suppose that Assumption 2.1, 2.2 hold. Consider an experimental design in Algorithm 2, with pilot chosen as in Algorithm 1. Then for all units participating in the main experiment:*

$$\begin{aligned}
&\text{Var}(Y_i|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}_S, \mathbf{P}) = \sigma^2\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k\Big). \\
&\text{Cov}(Y_i, Y_j|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}_S, \mathbf{P}) = \begin{cases} \eta\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k, T_j, D_j, \sum_{k\in N_j} D_k\Big) \text{ if } i \in \mathcal{N}_j \\ 0 \text{ otherwise} \end{cases}
\end{aligned} \tag{12}$$

for some functions $\sigma^2(.), \eta(.)$.

Given Lemma 3.3, we can identify and estimate the variance and covariance between units parametrically or non-parametrically. The estimators from the main experiment were defined as $\widehat{\sigma}, \widehat{\eta}$. Inference uses the plug-in estimator of the variance which uses the estimated $(\hat{\sigma}, \hat{\eta})$ obtained from the main experiment $\hat{V}_{N,\widehat{\sigma}^2,\widehat{\eta}}(w_N)$, and formal properties are discussed in Section 4.2.

Remark 2 (Choosing the sample size)**.** We are left to answer one question: "how should we select the number of participants in the main experiment?". In this case, we consider the problem where the number of treated units is internalized in the decision problem, for the variance not being larger than a pre-specified threshold $\beta(w_N)$. The optimization problem in such a case takes the following form.

$$\begin{aligned}
\min_{\mathbf{R},\mathbf{D}^R} \sum_{i=1}^{N} R_i \quad &\text{such that } (i)\beta_\alpha(w_N) \geq \hat{V}_{N,\hat{\sigma}_p,\hat{\eta}_p}(w_N; A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}, \mathbf{P}), \\
&(ii) R_i = 0 \quad \text{for all } i \in \mathcal{J}.
\end{aligned} \tag{13}$$

Remark 3 (Partial Network Information)**.** In Appendix A.1, we design an experiment that only uses partial information on the network structure. In particular, instead of observing the entire adjacency matrix, it observes some entries of A and imputes the remaining entries with a model. The main assumption is that individuals are organized into at least two disconnected components (clusters), the pilot study is constructed using one cluster, and the main experiment using the other cluster. □

Algorithm 2 Main Experiment

Require: A, $T_1, \cdots, T_N$ observed by researchers, $\left(\mathbf{P}, \widehat{\sigma}_p(\cdot), \widehat{\eta}_p(\cdot)\right), n_1, n_2$.

1: Select participants and treatment assignments as follows

$$(\mathbf{D}^R, \mathbf{R}) \in \arg\min_{\mathbf{r},\mathbf{d}^r} \max_{w_N \in \mathcal{W}_N} \widehat{V}_{N,\widehat{\sigma}_p,\widehat{\eta}_p}\left(w_N; A, \mathbf{d}^r, \mathbf{r}, \mathbf{T}^r, \mathbf{P}\right),$$
$$\text{s.t. } \mathbf{1}^\top \mathbf{r} \in [n_1, n_2], \mathbf{r}_j = 0 \text{ for all } j \in \mathcal{J}, \max_i \left| w_n\left(i, \mathbf{d}^r, \mathbf{r}, \mathbf{T}^R\right)\right| \leq \bar{C}, \tag{14}$$

where $\widehat{V}_{N,\widehat{\sigma}_p,\widehat{\eta}_p}(\cdot)$ is the plug-in estimator of the variance in Equation (11), and $\mathcal{J}$ as in Equation (9).

2: Collect information $\left[R_i\left(Y_i, D_i, T_i, D_{j\in\mathcal{N}_i}, \mathcal{N}_i\right)\right]_{i=1}^N$;

3: Estimate $\Gamma_n(w_N)$ as in Equation (7).

4: Estimate $\hat{\sigma}, \hat{\eta}$ using observations in the main experiment, and construct the variance of the estimator as $\widehat{V}_{N,\hat{\sigma},\hat{\eta}}(w_n; \cdot)$

return $\widehat{\Gamma}_n(w_N), \widehat{V}_{N,\hat{\sigma},\hat{\eta}}(w_n; \cdot)$ for each $w_n \in \mathcal{W}_N$.

4 Theoretical Analysis and Inference

In this section, we study theoretical guarantees and asymptotic inference. First, we introduce additional notation. Recall that m, n denote the size of the pilot and main experiment. Let $\mathcal{N}_{\max} = \max_{i\in\{1,\cdots,N\}} |\mathcal{N}_i|$ the maximum degree. We say that $a \lesssim b$ if $a \leq \bar{C}' b$ for a finite constant $\bar{C}'$.

4.1 Regret Analysis and Pilot's Size

We study how the worst-case variance of the estimator obtained from the two-stage experiment in Section 3 compares to the variance obtained from the oracle experiment, where the variance and covariance function are *known* and where all units, also the ones in the pilot experiment, may participate in the main experiment. Formally, the variance of the oracle method reads as follows:

$$\begin{aligned}\mathcal{V}_N = &\min_{\mathbf{r},\mathbf{d}^r} \max_{w_N \in \mathcal{W}_N} \mathrm{Var}\Big(\frac{1}{1^\top \mathbf{r}} \sum_{i=1}^{N} r_i w_N(i, \mathbf{d}^r, \mathbf{r}, \mathbf{T}^r) Y_i \Big| A, \mathbf{d}^r, \mathbf{r}, \mathbf{T}^r\Big), \\ &\text{such that } \mathbf{1}^\top \mathbf{r} = n_2.\end{aligned} \tag{15}$$

The oracle experiment minimizes the *true* variance and it does not impose any condition on the units in the pilot and their neighbors not participating in the main experiment. We assume that the oracle experiment treats exactly n_2 many individuals, which is the largest number of participants allowed in the feasible experiment in Equation (14) (instead of a number between (n_1, n_2)). From an inspection of the proof, we observe that, we may also impose a lower bound equal to $n_1 + |\mathcal{J}|$ and assume that $\mathbf{1}^\top \mathbf{r} \in [n_1 + |\mathcal{J}|, n_2]$ without affecting our results.[9]

We define the *regret* as the difference between the variance under the oracle solution of the optimization problem against the variance evaluated at the estimated treatment assignment.

$$\mathcal{R}_N = \max_{w_N \in \mathcal{W}_N} \mathrm{Var}\Big(\frac{1}{n} \sum_{i=1}^{N} R_i w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R) Y_i \Big| A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}^R\Big) - \mathcal{V}_N \tag{16}$$

where $\mathbf{R}, \mathbf{D}^R$ solve the two-wave experiment in Equation (14). We study the be-

[9] In this alternative scenario, the upper bound matches the upper bound in the empirical design discussed in Equation (14), namely the same maximum number of participants is considered for the two cases. On the other hand, we impose that for the oracle experiment $\sum_{i=1}^{N} r_i/n_1 \geq 1 + |\mathcal{J}|/n_1$, which exceeds the lower bound on the original design in Equation (14) by a factor $|\mathcal{J}|/n_1$. In the asymptotic regime, where the size of the pilot experiment is assumed to grow at a slower rate than the number of participants in the main experiment, $|\mathcal{J}|/n_1 \lesssim \mathcal{N}_{\max} m/n_1 = o(1)$, for the sample size n_1 being large enough and larger than the maximum degree and pilot's size.

havior of $n_2\mathcal{R}_N$, after approriately rescaling by the maximum sample size, since, otherwise we might expect that each component in the right-hand side of Equation (16) converges to zero.

Before stating our main theorem, we impose the following conditions.

Assumption 4.1. Assume that for some $\xi > 0$, the following hold:

$$\sup_{d,s,l} \left|\hat{\sigma}_p(d,s,l) - \sigma(d,s,l)\right| \lesssim m^{-\xi}, \sup_{d,s,l,d',s',l'} \left|\hat{\eta}_p(d,s,l,d',s',l') - \eta(d,s,l,d',s',l')\right| \lesssim m^{-\xi}.$$

Assumption 4.1 characterizes the convergence rate of the variance and covariance functions. For instance, the rate of convergence is $m^{-1/2}$ for parametric estimators of these functions. In the following assumption we impose moment conditions.

Assumption 4.2 (Moment and Distributional Conditions)**.** Suppose that the following holds:

(A) $Y_i \in [-M, M]$ where $M < \infty$;

(B) $\mathcal{N}_{\max}^2 / n_1^{1/2} = o(1)$.

Assumption 4.2 imposes the following conditions: (a) the outcome is bounded; (b) the maximum degree grows at a rate slower than $n_1^{1/4}$.

Theorem 4.1. *Under Assumption 2.1, 2.2, 4.1, 4.2, assuming that* $n_2/n_1 = \alpha \in (1, C), C < \infty$ *for* (n_1, n_2) *such that* $n_1 \geq \mathcal{N}_{\max} m \Big/ (\alpha - 1)$,

$$n_2 \mathcal{R}_N \lesssim \mathcal{N}_{\max} m^{-\xi} + \frac{\mathcal{N}_{\max}^2 m}{n_1}. \tag{17}$$

Theorem 4.1 characterizes the difference between the variance of the experiment with a pilot study against the variance of the oracle experiment with known variance and covariance functions. The theorem illustrates a trade-off: the size of the pilot experiment plays two contrasting effects on the upper bound for the regret: (i) the larger the size of the pilot experiment, the smaller the estimation error; (ii) the larger the size of the pilot, the stronger the constraints imposed in the optimization

algorithm, and therefore the larger the regret with respect to the oracle assignment mechanism. We can now state the following corollary.

Corollary. *Suppose that the conditions in Theorem 4.1 hold, with* $\xi = 1/2$ *(i.e., parametric rate). Then for* $m \asymp (n_1/\mathcal{N}_{\max})^{2/3}$*, we have* $n_2\mathcal{R}_N \lesssim \mathcal{N}_{\max}^{4/3} n_1^{-1/3}$*. Therefore under the above conditions,* $n_2\mathcal{R}_N \to_{a.s.} 0$*.*

The above corollary is the first result that formally characterizes the pilot's size with respect to the main experiment. The corollary shows that the pilot's size should be equal to the size of the main experiment divided by the maximum degree to the power of two-thirds. This result has important practical implications: it provides guidance on the choice of the pilot's size relative to the main experiment.

4.2 Asymptotic Inference

In the following lines, we derive the asymptotic properties of the estimator conditional on the treatment assignments. Inference depends on the estimated variance.

One necessary condition for the validity of the above estimator is uniform consistency of the estimator of the conditional variance $\hat{\sigma}^2(.)$ and covariance function $\hat{\eta}$. On the other hand, in the presence of a growing maximum degree, such a condition is not sufficient since the second component may have arbitrarily many elements. Therefore, validity may require some additional conditions on the network. Here, we require that the number of highly connected individuals represent a relatively small portion of the sample.

Assumption 4.3. Assume that the following holds:

(i) there exist a finite $L < \infty$ such that

$$L : \Big|\{i : |\mathcal{N}_i| > L\}\Big| \leq n^{3/4}\bar{C}',$$

for some universal constants $L, \bar{C}' < \infty$;

(ii) $\sup_{l,d,s} |\sigma^2(l,d,s) - \hat{\sigma}^2(l,d,s)| = o_p(1)$,
$\sup_{l,d,s,l',d',s'} |\eta(l,d,s,l',d',s') - \hat{\eta}(l,d,s,l',d',s')| = o_p(1)$;

(iii) $\mathcal{W}_N$ is finite dimensional;

(iv) for all $w_N \in \mathcal{W}_N$, $n_1 V_N(w_N) > 0$ almost surely.

Condition (i) states that the number of "influential nodes", namely the number of individuals with a large degree (larger than some finite L), grows at a slower rate than the sample size. Condition (ii) assumes that consistent variance and covariance estimators are available to the researcher. Condition (iii) assumes that the number of estimands if finite dimension, and (iv) that the variance is larger than zero after appropriately rescaling (to avoid degenerate convergence in distribution). We can state the next theorem.

Theorem 4.2. *Suppose that Assumption 2.1, 2.2, 4.2, 4.3 hold. Then for all* $w_N \in \mathcal{W}_N$,

$$\frac{\sqrt{n}(\hat{\Gamma}_n(w_N) - \Gamma_n(w_N))}{\sqrt{n\hat{V}_{N,\hat{\sigma},\hat{\eta}}(w_N)}} \to_d \mathcal{N}(0,1). \tag{18}$$

The proof of the theorem is contained in the Appendix. The above theorem establishes asymptotic normality for a general class of linear estimators. The rate of convergence of the estimator depends on the variance component $V_N(w_N)$. Whenever $nV_N(w_N) = O(1)$, the estimator achieves the parametric $\sqrt{n}$ convergence rate. Asymptotic properties of estimator for network data have been discussed in a variety of contexts (e.g., Ogburn et al. (2017); Chin (2018)). Here, we derive the asymptotic result conditionally on the assignment mechanism. Also, we impose restrictions not on the size of the maximum degree but on the number of individuals having a growing degree, allowing the maximum degree to be arbitrarily large.

5 Numerical Studies

In this section, we collect simulation results. Throughout this section, we set $T_i = |\mathcal{N}_i|$. We consider the following functional form for the variance and covariance

functions:

$$\sigma(l,d,s) = \mu + \beta_1 d + \frac{s\beta_2}{\max\{l,1\}}, \quad \eta(l,d,s,l',d',s') = \sqrt{\sigma(l,d,s) \times \sigma(l',d',s')}\alpha. \quad (19)$$

The variance depends on the individual treatment status and on the percentage of treated neighbors. The covariance instead is chosen using the Cauchy-Swartz inequality with α being the equivalent of the intra-cluster correlation in the presence of clustered networks (Baird et al., 2018). Notice that $\alpha \in [-1, 1]$. Similarly to simulations in Baird et al. (2018), we choose $\alpha = 0.1$. We choose $\mu = 0.5$ and we collect results for parameters β_1 and β_2 in $(0, 0)$, $(0.5, 0.5)$, $(0.5, 1)$. We denote each case respectively homoskedastic, "small heteroskedasticity", "large heteroskedasticity". In the Appendix we discuss results for a broader choice of parameters. Using the same exposure mapping in the simulations in Eckles et al. (2017), we choose the following specification of the outcome model

$$Y_i = D_i\gamma_1 + \frac{\sum_{k\in\mathcal{N}_i} D_k}{|\mathcal{N}_i|}\gamma_2 + \varepsilon_i. \quad (20)$$

We choose $\gamma_1 = 0.5, \gamma_2 = 1$. We remark that the choice of such coefficients does not affect the variance of the estimator (conditional on $\mathbf{D}^R$).

5.1 Simulated and Real-World Networks

In a first set of simulations, we generate data from an Erdős-Rényi graph with $P(A_{i,j} = 1) = 2/n$ and an Albert-Barabasi graph. For the latter, we first draw $n/5$ edges according to Erdős-Rényi graph with probabilities $p = 2/n$, and second, we draw connections of the new nodes sequentially to the existing ones with probability equal to the number of connection of each pre-existing node divided by the overall number of connections in the graph. We evaluate the methods over 200 data sets. For the simulated networks, we consider a graph with $N = 800$ and where the number of participants selected in the main experiment by the proposed method is *no larger than* the sample (i.e., we impose the constraint $1^\top \mathbf{R} \leq 400$).

In the second set of simulations, we evaluate results using the adjacency matrix from Cai et al. (2015). We consider two different adjacency matrices obtained from this study: the "weak" network, where two individuals are connected if either indicated the other as a friend, the "strong" network where two individuals are connected if both individuals indicate the other as a friend. The weak network presents a dense structure, whereas the strong network presents a sparse structure. We consider the adjacency matrix to be the matrix obtained from the first five villages, which counts in total $N = 832$, and we constraint the number of maximum participants selected by the proposed method to be 416 (i.e., $1^\top \mathbf{R} \leq 400$).

5.2 Methods

We evaluate the *proposed method*, with complete knowledge of the adjacency matrix and with a pilot study containing 70 units. Estimation of the variance and covariances is performed using a quadratic program with a positivity constraint on the variance function. In the estimation, we impose constraints on the estimated parameter for α being in $[0, 0.3]$. Such estimation problem reflects correct prior but imperfect knowledge of researchers on a positive correlation among neighbors, which often occurs in applications (Baird et al., 2018), and full incomplete knowledge of the parameters of the variance function. We solve the optimization problem over treatment assignment and participation indicators using non-linear mixed-integer programming.

In the case of a real-world network, we also consider the *proposed method with partially observed network* (see Appendix A.1). We estimate the variance and covariances, selecting 70 units for a pilot study from the sixth village. For such a method, the network in the main village is only partially observed before the randomization of the experiment. We consider the case where only the *sub-block* of the adjacency matrix of the first 200 individuals out of the 832 individuals is observable to the researcher before randomization. We impute missing edges using a simple Erdős-Rényi model, with a uniform prior on the probability of connections. The model is clearly wrongly specified in the real-world scenario. However, it shows that even such simple

specification may lead to improvements. We solve the optimization problem by alternating a Monte-Carlo step for estimating the variance over the unobserved edges and the optimization step over treatment assignments and participation indicators.

We compare to a set of competitors, where the number of participants either equals the number of participants in the main experiment, or it equals the *sum* of the number of participants in the main experiments and the number of units used in the pilot study. We consider the following **competing methods**: (ii) the 3-ϵ net *graph clustering* method with 400 participants discussed in Ugander et al. (2013); (iii) the 3-ϵ net *graph clustering* method with 470 participants, denoted as *Clustering+*, and three different saturation designs. Since saturation design methods are not directly applicable in the presence of a fully connected network, we consider *extensions* of saturation designs, where we combine the ϵ-net clustering discussed in Ugander et al. (2013), with the saturation design mechanism (Baird et al., 2018). We consider several alternative specifications (iv) *Saturation1*, with 400 participants, with uniform probability assignment across the estimated clusters; (v) the *Saturation1+*, having 470 participants and being as Saturation1; (vii) *Saturation2+*, with 470 participants, selects the saturation probabilities and the percentage of clusters for each probability of minimizing the sum of the standard errors of the treatment and spillover effect, with intracluster correlation equals to the true α and with the variance of the individual error set to be homoskedastic; (iix) *Saturation3+*, with 470 participants, instead minimizes the sum of the standard errors of treatment effects, spillover effects as well as on the slope effects as defined in Baird et al. (2018). However, we remark that saturation designs may have a poor performance in this particular case since they are not directly applicable in scenarios where (i) the network is not clustered; (ii) the variance is unknown to the researcher. Finally, we consider *Random Assignment +*, which selects at random 470 participants and assign equal probabilities treatments.[10] All competitors, with the exception of the random assignment mechanism, uses *complete* information of the network structure.

[10]Since the method in Jagadeesan et al. (2017) is only valid for direct effects, but not spillovers and overall effects, such method is not a suitable competitor in these simulations.

5.3 Results

We collect results for the real-world network in Table 1, where we report the variance of the estimator. Each column corresponds to different values of the coefficients (β_1, β_2). The left-hand side panel collects results for the network with strong ties, and the right-hand side panel collects results for the network with weak ties. Results show that the proposed method with the pilot study on real-world network simulations, significantly *outperforms uniformly* any competitor under *any design*. The improvement is significantly larger as the values of the coefficients increase, i.e., in the presence of heteroskedasticity.

In the presence of the partially observed network, the only valid competitor to the proposed method is the random allocation. In such a case, we observe that the proposed method significantly outperforms the random allocation strategy uniformly. Such behavior suggests the benefits of using the proposed method, even when little information is known about the network, and a simple and possibly misspecified modeling strategy is used for the network.

In the left panel of Figure 2 we report the percentage decrease in the sample size of the units necessary to achieve the same level of variance of the ELI method, when using the best competitor against the ELI method, for simulations on the real-world network. The "unobserved network" case in the panel compares the ELI method with partially observed network to the random allocation only. The number of units used by the ELI method is given by the sum of participants and the size of the pilot study. We observe that the proposed methodology requires between twenty and forty percent fewer participants to achieve the same level of precision.

In the right panel of Figure 2, we report the variance in the log-scale of the proposed method (in blue) against the competitor with the lowest median variance, which randomizes using the sum of participant and units in the pilot study. We consider a fully observed network where the network is simulated, as discussed above.

In the heteroskedastic case, we observe that the proposed method outperforms *uniformly* any competitor, and the improvement with respect to the competitors increases for a larger degree of heteroskedasticity. In the homoskedastic case (i.e.,

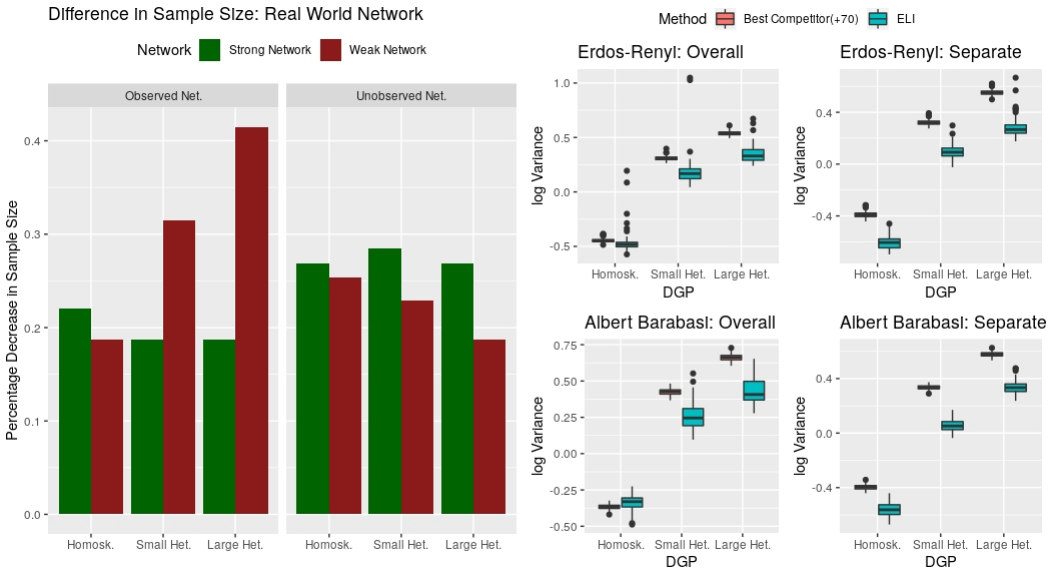

Figure 2: In the left panel, we report the percentage decrease in the number of units necessary to achieve the same level of variance between the best performing competitor and the ELI method, using the simulations with the real-world network. The case denoted as "Unobserved network"compares the random allocation to the ELI method with the partially observed network. In the right panel of Figure 2, we report the variance in the log-scale of the proposed method (in blue) against the competitor with the lowest median variance, which randomizes using the sum of participants selected by ELI and units in the pilot study.

$(\beta_1, \beta_2) = (0, 0)$), we observe the same behavior with *one single* exception, corresponding to estimating the overall effect under the Albert-Barabasi network. In such a case, the only method that outperforms the proposed procedure is graph clustering algorithms, with 70 *more* participants in the main experiment than the proposed method. In all remaining cases, the proposed method outperforms any competitor, including those that contain 70 more participants. Such behavior reflects the benefit of conducting a small pilot study before the main experiment, especially in the presence of heteroskedastic variances. Since, in this setting, we do not consider the presence of a separate cluster, as in the real-world network analysis, results for the partially observed network are not computed for simulated networks.

Table 1: Variance for estimating the overall effect, using data originated from Cai et al. (2015), using the first five villages as the population of interest $N = 832$). Each column corresponds to a different design, for different values of the coefficients (β_1, β_2). "ELI" corresponds to the proposed method, where 400 participants from the 832 potential participants are sampled in the main experiment, and a pilot study with 70 units is used. The second row corresponds to the proposed method, where only the first sub-block with the first 200 observations is observable from the main experiment, and a pilot of 70 units from the sixth village is used. Methods with a + use 470 participants in the main experiment, and without a +, such methods use 400 participants in the main experiment. All competitors, with the exception of the random allocation (Random All+), exploit full knowledge of the network structure.

	Strong			Weak		
Overall Effect	(0,0)	(0.5,0.5)	(0.5,1)	(0,0)	(0.5,0.5)	(0.5,1)
ELI	0.551	1.134	1.367	0.769	1.442	1.668
ELI - Unobserved Net	0.857	1.710	2.171	1.934	3.928	5.014
Random All+	1.107	2.249	2.876	2.430	4.827	6.127
Graph Clustering+	0.694	1.591	2.038	0.874	1.830	2.345
Saturation1+	0.913	1.985	2.513	1.523	3.143	3.866
Graph Clustering	0.793	1.847	2.420	0.989	2.104	2.623
Saturation1	1.059	2.259	2.940	1.736	3.603	4.482
Saturation2+	0.719	1.669	2.104	0.944	1.973	2.418
Saturation3+	0.931	2.171	2.772	1.700	3.844	4.829

	Strong			Weak		
Treatment and Spill	(0,0)	(0.5,0.5)	(0.5,1)	(0,0)	(0.5,0.5)	(0.5,1)
ELI	0.491	1.028	1.299	0.790	1.525	1.822
ELI - Unobserved Net	0.589	1.263	1.619	1.598	3.060	3.846
Random All+	0.641	1.431	1.882	1.813	3.580	4.477
Graph Clustering+	0.864	2.147	2.600	1.838	3.528	4.431
Saturation1+	0.652	1.500	1.942	1.403	2.807	3.569
Graph Clustering	0.999	2.491	3.001	2.165	4.022	5.302
Saturation1	0.760	1.755	2.286	1.654	3.283	4.068
Saturation2+	0.773	1.900	2.371	1.516	2.986	3.724
Saturation3+	0.801	1.910	2.449	2.231	4.202	5.155

6 Conclusions

In this paper, we have introduced a novel method for designing experiments under interference. We propose a design that selects treatment assignments and participation indicators to minimize the variance of the final estimator. We propose the first statistical framework for two-wave experiments with interference and derive regret guarantees.

We considered designs where the complete network information is available to the researchers. In the Appendix and simulations, we show how our setting extends to partially observed networks. Our numeric findings suggest that imputing the network may lead to improvements in the variance of the estimators. We leave for future research a comprehensive theoretical analysis of the scenario with the partially observed network and model selection in this case.

This paper makes two key assumptions: interactions are anonymous, and interference propagates to the neighbors only. Future research should address the question of design under general interactions and interference propagating on the entire network. Exploring the effect of the network topology and different exposure mappings on the performance of the design mechanisms remains an open research question.

References

Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics 11*(4), 1912–1947.

Athey, S., D. Eckles, and G. W. Imbens (2018). Exact p-values for network interference. *Journal of the American Statistical Association 113*(521), 230–240.

Bai, Y. (2019). Optimality of matched-pair designs in randomized controlled trials. *Available at SSRN 3483834*.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018). Optimal design of experiments in the presence of interference. *Review of Economics and Statistics 100*(5), 844–860.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Barrera-Osorio, F., M. Bertrand, L. L. Linden, and F. Perez-Calle (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. *American Economic Journal: Applied Economics 3*(2), 167–95.

Barrios, T. (2014). Optimal stratification in randomized experiments. *Manuscript, Harvard University*.

Basse, G. and A. Feller (2016). Analyzing multilevel experiments in the presence of peer effects. *arXiv preprint arXiv 1608*.

Basse, G. W. and E. M. Airoldi (2018a). Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology 48*(1), 136–151.

Basse, G. W. and E. M. Airoldi (2018b). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika 105*(4), 849–858.

Bhattacharya, D., P. Dupas, and S. Kanaya (2013). Estimating the impact of means-tested subsidies under treatment externalities with application to anti-malarial bednets. Technical report, National Bureau of Economic Research.

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2017). Using aggregated relational data to feasibly identify network structure without network data. Technical report, National Bureau of Economic Research.

Cai, J., A. De Janvry, and E. Sadoulet (2015). Social networks and the decision to insure. *American Economic Journal: Applied Economics 7*(2), 81–108.

Charnes, A. and W. W. Cooper (1962). Programming with linear fractional functionals. *Naval Research logistics quarterly 9*(3-4), 181–186.

Chin, A. (2018). Central limit theorems via stein's method for randomized experiments under interference. *arXiv preprint arXiv:1804.03105*.

DellaVigna, S. and D. Pope (2018). Predicting experimental results: who knows what? *Journal of Political Economy 126*(6), 2410–2456.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review 101*(5), 1739–74.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica 82*(1), 197–228.

Eckles, D., B. Karrer, and J. Ugander (2017). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference 5*(1).

Egger, D., J. Haushofer, E. Miguel, P. Niehaus, and M. W. Walker (2019). General equilibrium effects of cash transfers: experimental evidence from kenya. Technical report, National Bureau of Economic Research.

Forastiere, L., E. M. Airoldi, and F. Mealli (2016). Identification and estimation of treatment and interference effects in observational studies on networks. *arXiv preprint arXiv:1609.06245*.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264.

Harshaw, C., F. Sävje, D. Spielman, and P. Zhang (2019). Balancing covariates in randomized experiments using the gram-schmidt walk. *arXiv preprint arXiv:1911.03071*.

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

Jagadeesan, R., N. Pillai, and A. Volfovsky (2017). Designs for estimating the treatment effect in networks with interference. *arXiv preprint arXiv:1705.08524*.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(1), 85–112.

Kang, H. and G. Imbens (2016). Peer encouragement designs in causal inference with partial interference and identification of local average network effects. *arXiv preprint arXiv:1609.04464*.

Karlan, D. and J. Appel (2018). *Failing in the field: What we can learn when field research goes wrong*. Princeton University Press.

Karlan, D. S. and J. Zinman (2008). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review 98*(3), 1040–68.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis 24*(3), 324–338.

Kitagawa, T. and G. Wang (2020). Who should get vaccinated? individualized allocation of vaccines over sir network. *arXiv preprint arXiv:2012.04055*.

Leung, M. P. (2019). Treatment and spillover effects under network interference. *Available at SSRN 2757313*.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23.

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217.

Muralidharan, K., P. Niehaus, and S. Sukhtankar (2017). General equilibrium effects of (improving) public employment programs: Experimental evidence from india. Technical report, National Bureau of Economic Research.

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

Paluck, E. L., H. Shepherd, and P. M. Aronow (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences 113*(3), 566–571.

Pouget-Abadie, J. (2018). *Dealing with Interference on Experimentation Platforms*. Ph. D. thesis.

Ross, N. et al. (2011). Fundamentals of stein's method. *Probability Surveys 8*, 210–293.

Sävje, F., P. M. Aronow, and M. G. Hudgens (2017). Average treatment effects in the presence of unknown interference. *arXiv preprint arXiv:1711.06399*.

Sinclair, B., M. McConnell, and D. P. Green (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science 56*(4), 1055–1069.

Sussman, D. L. and E. M. Airoldi (2017). Elements of estimation theory for causal effects in the presence of network interference. *arXiv preprint arXiv:1702.03578*.

Tabord-Meehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv:1806.05127*.

Taylor, S. J. and D. Eckles (2018). Randomized experiments to detect and estimate social influence in networks. In *Complex Spreading Phenomena in Social Systems*, pp. 289–322. Springer.

Ugander, J., B. Karrer, L. Backstrom, and J. Kleinberg (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 329–337. ACM.

Vazquez-Bare, G. (2017). Identification and estimation of spillover effects in randomized experiments. *arXiv preprint arXiv:1711.02745*.

Viviano, D. (2019). Policy targeting under network interference. *arXiv preprint arXiv:1906.10258*.

Viviano, D. (2020). Policy choice in experiments with unknown interference. *arXiv preprint arXiv:2011.08174*.

Wager, S. and K. Xu (2019). Experimenting in equilibrium. *arXiv preprint arXiv:1903.02124*.

A Extensions

A.1 Design with Partial Network Information

In this section, we now consider the case where the researcher has access to partial network information only. The main assumption is that individuals are organized into at least two disconnected components (clusters), and the pilot study is constructed using information from one of the two clusters only. The experimental protocol is discussed below.

Experimental Protocol:

1. Pilot study: Researchers select a random sample of individuals $\mathbf{P}$, which is assumed to be disconnected from all other eligible units, and assign treatments $\mathbf{D}^P$ to the units in the pilot. This sample may be collected from a disconnected component of the network, which we denote as $\mathcal{C}$.[11] Researchers observe

$$\Big[P_i\Big(Y_i, D_i, T_i, D_{\mathcal{N}_i}\Big)\Big], \quad i \in \mathcal{C} : (i, \mathcal{N}_i) \cap \{1, \cdots, N\} = \emptyset,$$

where $\{1, \cdots, N\}$ denotes the elegible sample for the main experiment. Researchers estimate $\hat{\sigma}_p, \hat{\eta}_p$ from the pilot study (note that the pilot study must contain for some individuals also their neighbors to be able to estimate $\eta(\cdot)$).

2. Survey: researchers collect network information of a random subset of individuals $i \in \{1, ..., N\}$, namely

$$\tilde{A} = \Big(\mathcal{N}_i, i \in \mathcal{S} \subset \{1, \cdots, N\}\Big), \quad \mathbf{T} = \Big(T_1, \cdots, T_N\Big).$$

[11] Examples include a village (Banerjee et al., 2013), school (Paluck et al., 2016) or region (Muralidharan et al., 2017).

3. Main Experiment: researchers select the participants and the corresponding treatment assignments in the main experiment. They minimize the posterior variance

$$\min_{\mathbf{r},\mathbf{d}^r} \max_{w_N \in \mathcal{W}_N} \mathbb{E}\Big[\hat{V}_{N,\hat{\sigma}_p,\hat{\eta}_p}(w_N; A, \mathbf{d}^r, \mathbf{r}, \mathbf{T}, \mathbf{P})\Big|\tilde{A}, \mathbf{T}\Big], \quad s.t. : \sum_{i=1}^{N} r_i = n, \quad r_i = 0 \text{ for all } i \in \mathcal{C}. \tag{21}$$

The posterior variance is estimated with respect to a prior distribution of $(A, T_1, \cdots, T_N) \sim \mathcal{P}$.

4. Second survey and analysis: researchers collect information

$$\Big[R_i\Big(Y_i, D_i, D_{j\in\mathcal{N}_i}, \mathcal{N}_i\Big)\Big]$$

for each participant, and construct $\hat{\Gamma}_n(w_N)$ as in Equation (7). We also estimate $\hat{\sigma}, \hat{\eta}$ using information from the units sampled in the main experiment and their neighbors. Inference follows similarly to Section 4.2.

The experiment consists of four steps: a pilot study, where network information is available to the researcher, a first survey that collects partial network information, the design, and the analysis. Note that the estimator $\hat{\Gamma}_n(w_N)$ and the variance and covariance functions $\hat{\eta}, \hat{\sigma}$ can be estimated without observing the entire network, but it suffices the outcomes, covariates and treatment assignments of individuals in the main experiment and their neighbors. In the main experiment, we optimize the posterior expectation of the variance, fixing the estimated variance and covariance functions obtained from the main pilot, and averaging over the edges' distributions. The optimization can be solved by alternating a Monte Carlo simulation, and the optimization through mixed-integer programming, which we show in simulations, performs well in practice.[12] Under the above protocol, the following holds.

[12]The imputation problem can also be solved in a fully Bayesian fashion by imposing a prior distribution also on potential outcomes and taking the posterior density over the outcomes. However, a full derivation of a hierarchical model goes beyond the scope of this paper, and we leave this for

Proposition A.1. *Let Assumption 2.1, 2.2 hold. Then the experimental design in the current section satisfies the restrictions in Proposition 3.1.*

Example A.1. Consider the following Erdős-Rényi model:

$$\{A_{i,j}\}_{j>i} \sim_{i.i.d} \text{Bern}(p), \quad p \sim \mathcal{U}(0,1). \tag{22}$$

Assume in addition that $A_{i,j} = A_{j,i}$ and $A_{i,i,} = 0$. The model assumes that each individual connects with independent probabilities. Such probabilities are modeled based on a uniform prior. Suppose we observe edges of a subset of individuals $\tilde{n}$. Then we obtain that

$$P(A_{i,j} = 1|\tilde{A}) \sim \begin{cases} \delta_1 \text{ if } \tilde{A}_{i,j} = 1 \\ \delta_0 \text{ if } \tilde{A}_{i,j} = 0 \\ \text{Beta}(\alpha, \beta) \text{ if } \tilde{A}_{i,j} \text{ is missing} \end{cases} \tag{23}$$

where δ_c denotes a point-mass distribution at c and

$$\alpha = \sum_{u>v} \tilde{A}_{u,v} 1\{\tilde{A}_{u,v} \in \{0,1\}\} + 1, \quad \beta = \tilde{N} - \sum_{u>v} \tilde{A}_{u,v} 1\{\tilde{A}_{u,v} \in \{0,1\}\} + 1 \tag{24}$$

$\tilde{N}$ is the number of observed connections. □

Example A.2. Following Breza et al. (2017), we can consider a model of the form

$$P(A_{i,j} = 1|\nu_i, \nu_j, z_i, z_j, \delta) \propto \exp\Big(\nu_i + \nu_j + \delta \text{dist}(z_i, z_j)\Big), \tag{25}$$

where ν_i denotes individual fixed effect, z_i denotes a position in some latent space and δ is an hyper-parameter of interest. □

future research.

A.2 Higher-Order Dependence

In this section, we relax the local dependence assumption and consider the general case where unobservables exhibit M-dependence. We denote $\mathcal{N}_i^M$ the set of individuals connected to individual i by exactly M edges.

We replace Assumption 2.1 with the following conditions.

Assumption A.1 (Model under Higher-Order Dependence)**.** Assume in addition that for all $i \in \{1, ..., N\}$,

$$\begin{aligned}
&\left\{\varepsilon_i, \{\varepsilon_k\}_{k \notin \cup_{u=1}^M N_j^u, j \in \cup_{u=1}^M \mathcal{N}_i^u}\right\} \perp \{\varepsilon_j\}_{j \notin \cup_{u=1}^M \mathcal{N}_i^u} \Big| A, \mathbf{T} \quad a.s. \\
&(\varepsilon_i, \varepsilon_j) =_d (\varepsilon_{i'}, \varepsilon_{j'}) \Big| A, \mathbf{T} \text{ for all } (i, j, i', j') : i \in \mathcal{N}_j, i' \in \mathcal{N}_{j'}, T_i = T_{i'}, T_j = T_{j'}, \\
&\mathcal{N}_{\max} < c,
\end{aligned} \tag{26}$$

for a universal constal $c < \infty$.

Assumption A.1 states the following: (i) unobservables are independent whenever they are distant by more than M edges; (ii) the joint distribution of two unobservables given the adjacency matrix is the same, whenever covariates are the same, and these unobservables are connected by the same number of edges. In addition, the assumption states that the maximum degree is uniformly bounded.[13]

The second condition is the experimental restriction. We define

$$\tilde{\mathcal{H}} = \{1, \cdots, N\} \setminus \left\{i : P_i = 1, \cup_{u=1}^M \mathcal{N}_i^u\right\}. \tag{27}$$

The set $\tilde{\mathcal{H}}$ denotes all individuals in the population of interest, after excluding the pilot units and the neighbors of the pilot units up to the Mth degree. The following restriction is imposed.

Assumption A.2 (Experimental Restriction)**.** Let the following hold:

[13] After a quick inspection of the derivations contained in Appendix E, the reader may observe that such condition can be replaced by assuming that the maximum degree of the sampled units and their neighbors up to order M scales at a rate slower than $n^{1/4}$.

(A) $\varepsilon_{i\in\tilde{\mathcal{H}}} \perp \left(\mathbf{D}^R, \mathbf{R}\right) \Big| A, \mathbf{T}, \mathbf{P}$;

(B) $\varepsilon_{1,\cdots,N} \perp \mathbf{P} \Big| A, \mathbf{T}$;

(C) $R_j = 0$ for all $j \in \left\{i : P_i = 1, \cup_{u=1}^{M} \mathcal{N}_i^u\right\}$.

The following theorem guarantees unbiased of the estimator for any design satisfying the above restrictions.

Theorem A.2. *Under Assumption 2.1, Assumption A.1, A.2*

$$\mathbb{E}\left[\hat{\Gamma}_n(w_n) \Big| \mathbf{P}, A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}\right] = \Gamma_n(w_n).$$

The design minimizes the estimated variance from the pilot study under the restrictions in Assumption A.2. The variance takes the following form:

$$\begin{aligned}
nV_N(w_N) = &\frac{1}{n}\sum_{i:R_i=1} w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T})\mathrm{Var}(Y_i|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) \\
&+ \frac{1}{n}\sum_{i:R_i=1}\sum_{j\in\mathcal{N}_i^1} R_j w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) w_N(j, \mathbf{D}^R, \mathbf{R}, \mathbf{T})\mathrm{Cov}(Y_i, Y_j|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) \\
&+ \frac{1}{n}\sum_{i:R_i=1}\sum_{j\in\mathcal{N}_i^2} R_j w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) w_N(j, \mathbf{D}^R, \mathbf{R}, \mathbf{T})\mathrm{Cov}(Y_i, Y_j|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) \qquad (28) \\
&+ \cdots \\
&+ \frac{1}{n}\sum_{i:R_i=1}\sum_{j\in\mathcal{N}_i^M} R_j w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) w_N(j, \mathbf{D}^R, \mathbf{R}, \mathbf{T})\mathrm{Cov}(Y_i, Y_j|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}).
\end{aligned}$$

Therefore, the variance sums over the covariances of each individual and her neighbors up to the Mth degree. Notice now that the variance and each covariance component is identified with information containing neighbors' assignments and outcomes up to the M degree, where each covariance component depends on the distance of

unit i from element j. Formally, we obtain that the following holds.

$$\text{Var}(Y_i|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) = \text{Var}\Big(r(D_i, \sum_{k\in\mathcal{N}_i} D_k, T_i, \varepsilon_i)\Big| D_i, \sum_{k\in\mathcal{N}_i} D_k, T_i\Big) = \sigma^2\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k\Big), \tag{29}$$

which guarantees identifiability of the variance function. Similarly, for a given $j \in \mathcal{N}_i^u$ we have

$$\begin{aligned}\text{Cov}(Y_i, Y_j|A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) &= \text{Cov}\Big(r(D_i, \sum_{k\in\mathcal{N}_i} D_k, T_i, \varepsilon_i), r(D_j, \sum_{k\in\mathcal{N}_j} D_k, T_j, \varepsilon_j)|A, \mathbf{D}^R, \mathbf{T}\Big) \\ &= \text{Cov}\Big(r(D_i, \sum_{k\in\mathcal{N}_i} D_k, T_i, \varepsilon_i), r(D_j, \sum_{k\in\mathcal{N}_j} D_k, T_j, \varepsilon_j)|j \in \mathcal{N}_i^u, \mathbf{D}^R\Big) \\ &= \eta_u\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k, T_j, D_j, \sum_{k\in\mathcal{N}_j} D_k\Big).\end{aligned} \tag{30}$$

The above expression states that under the above condition the covariance between two individuals, whose shortest path between such two individual is of length u is a function which only depends on (a) the length of the path, (b) the treatment assignment of each of these two individuals, (c) the treatment assignments of the corresponding neighbors, (d) the network statistics T_i and T_j of these two individuals.

The design of the experiment consists in minimizing the variance under the restriction in Assumption A.2.

The following theorem allows for inference.

Theorem A.3. *Suppose that Assumption 2.1, A.1, A.2 hold. Then for $V_N(w_n)$ as defined in Equation (28), with $n_1 V_N(w_n) > 0$,*

$$\frac{\sqrt{n}(\hat{\Gamma}_n(w_N) - \Gamma_n(w_n))}{\sqrt{nV_N(w_n)}} \to_d \mathcal{N}(0, 1). \tag{31}$$

The proof of the theorem is contained in Appendix E. The variance can then be estimated consistently via plug in of each estimated covariance function.

B Additional Tables

We collect results of the simulated network in Table 2, and Table 3. Each table reports the variance averaged over two-hundred replications. Each design corresponds to a different set of parameters (β_1, β_2), which can be found at the top of the table.

Table 2: Variance of the overall effect (sum of spillover and treatment effects). 200 replications. Each column corresponds to different values of the coefficient. Panel at the top collects results for the Erdős-Rényi graph and at the bottom for the Albert-Barabasi graph.

ER	(0,0)	(0, 0.5)	(0,1)	(0, 1.5)	(0.5,0.5)	(0.5,1)	(0.5,1.5)	(1,1.5)
ELI	0.624	0.929	1.194	1.420	1.200	1.415	1.637	1.855
Random All+	1.162	1.740	2.315	2.891	2.329	2.905	3.479	4.068
Graph Clust+	0.640	0.991	1.343	1.694	1.361	1.713	2.063	2.434
Saturation1+	0.908	1.378	1.849	2.316	1.859	2.330	2.801	3.282
Graph Clust	0.767	1.188	1.607	2.029	1.631	2.051	2.471	2.916
Saturation1	1.090	1.654	2.217	2.781	2.231	2.794	3.358	3.932
Saturation2+	0.679	1.047	1.416	1.783	1.430	1.800	2.169	2.550
Saturation3+	0.993	1.587	2.177	2.771	2.178	2.771	3.364	3.954

AB	(0,0)	(0, 0.5)	(0,1)	(0, 1.5)	(0.5,0.5)	(0.5,1)	(0.5,1.5)	(1,1.5)
ELI	0.714	0.909	1.278	1.566	1.294	1.548	1.595	2.035
Random All+	1.144	1.714	2.284	2.851	2.299	2.874	3.482	4.028
Graph Clust+	0.693	1.098	1.503	1.908	1.531	1.938	2.060	2.773
Saturation1+	0.936	1.435	1.936	2.434	1.950	2.451	2.800	3.464
Graph Clust	0.837	1.325	1.811	2.299	1.845	2.333	2.471	3.338
Saturation1	1.132	1.733	2.336	2.934	2.354	2.955	3.358	4.179
Saturation2+	0.732	1.152	1.572	1.992	1.594	2.015	2.169	2.882
Saturation3+	1.091	1.762	2.433	3.103	2.425	3.096	3.364	4.425

Table 3: Maximum variance between estimator of the direct treatment and spillover effect. 200 replications. Each column corresponds to different values of the coefficient. Panel at the top collects results for the Erdős-Rényi graph and at the bottom for the Albert-Barabasi graph.

ER	(0,0)	(0, 0.5)	(0,1)	(0, 1.5)	(0.5,0.5)	(0.5,1)	(0.5,1.5)	(1,1.5)
ELI	0.545	0.782	1.091	1.308	1.102	1.321	1.579	1.864
Random All+	0.676	1.037	1.400	1.763	1.379	1.740	2.101	2.441
Graph Clust+	1.224	1.674	2.120	2.568	2.601	3.046	3.497	4.424
Saturation1+	0.678	1.036	1.395	1.756	1.409	1.769	2.128	2.501
Graph Clust	1.496	2.038	2.585	3.129	3.173	3.717	4.259	5.397
Saturation1	0.825	1.262	1.698	2.136	1.715	2.150	2.588	3.037
Saturation2+	0.969	1.393	1.820	2.247	2.053	2.478	2.901	3.564
Saturation3+	0.930	1.474	2.016	2.562	1.930	2.473	3.013	3.474

AB	(0,0)	(0, 0.5)	(0,1)	(0, 1.5)	(0.5,0.5)	(0.5,1)	(0.5,1.5)	(1,1.5)
ELI	0.571	0.792	1.081	1.359	1.059	1.399	1.574	1.879
Random All+	0.672	1.057	1.443	1.830	1.398	1.782	2.101	2.510
Graph Clust+	0.984	1.383	1.784	2.184	2.192	2.594	3.495	3.809
Saturation1+	0.676	1.060	1.444	1.829	1.453	1.837	2.127	2.613
Graph Clust	1.204	1.689	2.175	2.661	2.678	3.163	4.261	4.638
Saturation1	0.827	1.294	1.763	2.233	1.773	2.239	2.587	3.189
Saturation2+	0.859	1.262	1.665	2.066	1.902	2.307	2.904	3.350
Saturation3+	0.984	1.590	2.196	2.805	2.107	2.713	3.015	3.834

C Identification

Proof of Proposition 3.1

Consider all $\mathbf{D}^R$ such that $D_i = d, \sum_{k \in \mathcal{N}_i} D_k = s$, and all A such that $T_i = l$. To derive the result we want to show that

$$\mathbb{E}\Big[Y_i \Big| D_i = d, \sum_{k \in \mathcal{N}_i} D_k = s, T_i = l, \mathbf{D}^R, A, \mathbf{R}, \mathbf{T}, \mathbf{P}\Big] = m(d, s, l) \tag{32}$$

for all those units in the sample (i.e., $R_i = 1$).

Notice first that under Assumption 2.1,

$$\begin{aligned} &\mathbb{E}\Big[Y_i \Big| D_i = d, \sum_{k \in \mathcal{N}_i} D_k = s, T_i = l, \mathbf{D}^R, A, \mathbf{R}, \mathbf{T}, \mathbf{P}\Big] = \\ &\mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big) \Big| D_i = d, \sum_{k \in \mathcal{N}_i} D_k = s, T_i = l, \mathbf{D}^R, A, \mathbf{R}, \mathbf{P}\Big]. \end{aligned} \tag{33}$$

Observe now that under the conditions in Proposition 3.1, since participants are *not* units in the pilot study, and since $\varepsilon_i(\mathbf{d})$ is a constant function in $\mathbf{d}$, we have that the following holds:

$$\mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big) \Big| D_i = d, \sum_{k \in \mathcal{N}_i} D_k = s, T_i = l, \mathbf{D}^R, A, \mathbf{R}, \mathbf{T}, \mathbf{P}\Big] = \mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big) \Big| T_i = l, A, \mathbf{T}\Big]. \tag{34}$$

Under Assumption 2.1, since $\varepsilon_i \perp (A, \mathbf{T})$, the proof completes.

Proof of Theorem A.2

The proof follows similarly to the previous proof. Consider all $\mathbf{D}^R$ such that $D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s$, and all A such that $T_i = l$. Notice first that under Equation (1),

$$\begin{aligned}&\mathbb{E}\Big[Y_i\Big|D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s, T_i = l, \mathbf{D}^R, A, \mathbf{R}, \mathbf{T}, \mathbf{P}\Big] = \\ &\mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big)\Big|D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s, T_i = l, \mathbf{T}, \mathbf{D}^R, A, \mathbf{R}, \mathbf{P}\Big].\end{aligned} \tag{35}$$

Observe now that under Assumption A.2, since participants are *not* units in the pilot study and their neighbors up to the Mth degree, we have that the following holds:

$$\mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big)\Big|D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s, T_i = l, \mathbf{D}^R, A, \mathbf{R}, \mathbf{T}, \mathbf{P}\Big] = \mathbb{E}\Big[r\Big(d, s, l, \varepsilon_i\Big)\Big|A, \mathbf{T}, T_i = l\Big]. \tag{36}$$

Under Equation (1), since $\varepsilon_i \perp (A, \mathbf{T})$, the proof completes.

Proof of Lemma 3.2

For all individuals i selected in the pilot, we can write

$$\begin{aligned}\mathrm{Var}\Big(Y_i\Big|A, D_i = d, T_i = l, \mathcal{D}_{\mathcal{N}_i} = \mathbf{s}, \mathbf{P}\Big) &= \mathrm{Var}\Big(r(D_i, \sum_{k\in\mathcal{N}_i} D_k, T_i, \varepsilon_i)\Big|A, D_i = d, T_i = l, \mathcal{D}_{\mathcal{N}_i} = \mathbf{s}, \mathbf{P}\Big) \\ &= \mathrm{Var}\Big(r(d, 1^\top\mathbf{s}, l, \varepsilon_i)\Big|A, T_i = l, \mathbf{P}\Big) \\ &= \mathrm{Var}\Big(r(d, 1^\top\mathbf{s}, l, \varepsilon_i)\Big|\mathbf{P}\Big) \\ &= \mathrm{Var}\Big(r(d, 1^\top\mathbf{s}, l, \varepsilon_i)\Big) \\ &= \sigma^2(d, 1^\top\mathbf{s}, l).\end{aligned}$$

This complete the claim for the variance. The first equation follows from Assumption 2.1. The second equation follows from the fact that treatments are randomized exogenously and $\varepsilon_i(\mathbf{D})$ is constant in $\mathbf{D}$. The third equation follows from the fact that $\varepsilon_i \perp (\mathbf{T}, A)$. The last equation follows from the fact that $\mathbf{P}$ is chosen based on

$\mathbf{T}, A$ only. The analysis of the covariances follows similarly. Namely, following the same steps as before,

$$\begin{aligned}&\mathrm{Cov}\Big(Y_i, Y_j\Big|A, D_i = d, D_j = d', D_{\mathcal{N}_i} = \mathbf{s}, D_{\mathcal{N}_j} = \mathbf{s}', T_i = l, T_j = l', \mathbf{P}\Big) = \\ &\mathrm{Cov}\Big(r(d, 1^\top \mathbf{s}, l, \varepsilon_i), r(d', 1^\top \mathbf{s}', l', \varepsilon_j)\Big|A, T_i = l, T_j = l', \mathbf{P}\Big)\end{aligned} \tag{37}$$

where we used the exogeneity of the treatment assignment and the fact that $\varepsilon_i(\mathbf{d})$ is a constant function in $\mathbf{d}$. Using Assumption 2.2, we have

$$(37) = \begin{cases} \mathrm{Cov}\Big(r(d, 1^\top \mathbf{s}, l, \varepsilon_i), r(d', 1^\top \mathbf{s}', l', \varepsilon_j)\Big|j \in \mathcal{N}_i, \mathbf{P}\Big) \text{ if } j \in \mathcal{N}_i \\ 0 \text{ otherwise.} \end{cases}$$

Using the second condition in Assumption 2.2, and the fact that $\mathbf{P}$ is assigned using information from A only, we can write

$$\mathrm{Cov}\Big(r(d, 1^\top \mathbf{s}, l, \varepsilon_i), r(d', 1^\top \mathbf{s}', l', \varepsilon_j)\Big|j \in \mathcal{N}_i\Big) = \mathrm{Cov}\Big(r(d, 1^\top \mathbf{s}, l, \varepsilon_i), r(d', 1^\top \mathbf{s}', l', \varepsilon_j)\Big|j \in \mathcal{N}_i\Big)$$

which completes the proof.

Proof of Lemma 3.3

The proof follows similarly to the one of Lemma 3.2.

Consider all $\mathbf{D}^R$ such that $D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s$, and all A such that $T_i = l$, $\mathbf{R} : R_i = 1$. Under Assumption 2.1,

$$\begin{aligned}&\mathrm{Var}\Big(Y_i\Big|D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s, \mathbf{D}^R, \mathbf{R}, A, T_i = l, R_i = 1, \mathbf{T}, \mathbf{P}\Big) \\ =&\mathrm{Var}\Big(r(d, s, l, \varepsilon_i)\Big|D_i = d, \sum_{k\in\mathcal{N}_i} D_k = s, \mathbf{D}^R, \mathbf{R}, A, T_i = l, R_i = 1, \mathbf{T}, \mathbf{P}\Big).\end{aligned} \tag{38}$$

Under Assumption 2.2, since $R_i = 0$ for all those units not being in the pilot study

and their neighbors, it follows that

$$\begin{aligned}&\mathrm{Var}\Big(r(d,s,l,\varepsilon_i)\Big|D_i=d,\sum_{k\in\mathcal{N}_i}D_k=s,\mathbf{D}^R,\mathbf{R},A,T_i=l,R_i=1,\mathbf{P},\mathbf{T}\Big)\\&=\mathrm{Var}\Big(r(d,s,l,\varepsilon_i)\Big|\mathbf{T},A\Big).\end{aligned}\tag{39}$$

Under Assumption 2.1, since $\varepsilon_i \perp (A,\mathbf{T})$, the proof of the first part completes.

For the covariance component, the same reasoning follows. Consider $(i,j) : R_i = R_j = 1$. Then we write

$$\begin{aligned}&\mathrm{Cov}\Big(Y_i,Y_j\Big|D_i=d,D_j=d',\sum_{k\in\mathcal{N}_i}D_k=s,\sum_{k\in N_j}D_k=s',\mathbf{D}^R,\mathbf{R},A,T_i=l,T_j=l',\\&R_i=1,R_j=1,\mathbf{T},\mathbf{P}\Big)=\\&\mathrm{Cov}\Big(r(d,s,l,\varepsilon_i),r(d',s',l',\varepsilon_j)\Big|D_i=d,D_j=d',\sum_{k\in\mathcal{N}_i}D_k=s,\sum_{k\in N_j}D_k=s',\mathbf{D}^R,\\&\mathbf{R},A,T_i=l,T_j=l',R_i=1,R_j=1,\mathbf{T},\mathbf{P}\Big).\end{aligned}\tag{40}$$

By Assumption 2.2, since $\mathbf{D}^R,\mathbf{R}$ depends on $(A,\mathbf{T})$ and unobservables which are not connected to (i,j), $\varepsilon_i \perp \mathbf{D}^R,\mathbf{R}$, and we obtain that Equation (40) equals

$$\mathrm{Cov}\Big(r(d,s,l,\varepsilon_i),r(d',s',l',\varepsilon_j)\Big|A,T_i=l,T_j=l',\mathbf{T},\mathbf{P}\Big).\tag{41}$$

The covariance is zero if two individuals are not neighbors. In such a case the lemma trivially holds. Therefore, consider the case where individuals are neighbors. Then by Assumptionn 2.1 the above equation simplifies to

$$\mathrm{Cov}\Big(r(d,s,l,\varepsilon_i),r(d',s',l',\varepsilon_j)\Big).$$

which completes the proof.

D Proof of Theorem 4.1

Preliminaries First, recall that weights $w(i;\cdot)$ are equal to zero for those units not in the experiment. Throughout the proof we will be omitting the superscript $\mathbf{D}^R$ and write $\mathbf{D}$ the assignments in the main experiment. For arbitrary $\mathbf{D}^*, \mathbf{R}^*$, we denote

$$\hat{V}_{n,p}(\mathbf{D}^*, \mathbf{R}^*) = \max_{w_N \in \mathcal{W}_N} \hat{V}_{n,\hat{\sigma}_p,\hat{\eta}_p}(w_N; \mathbf{D}^*, \mathbf{R}^*, \mathbf{T}, A),$$

the maximum conditional variance over $\mathcal{W}_N$ with estimated covariance and variance function obtained from the pilot experiment and $V_N(\mathbf{D}^*, \mathbf{R}^*)$, the population counterpart. For notational convenience we refer to $w_N(i,.)$, omitting the last arguments, whenever clear from the context. Let

$$(\tilde{\mathbf{D}}^R, \tilde{\mathbf{R}}) \in \arg \min_{\mathbf{r},\mathbf{d}^r, n_1 \leq \sum_{i=1}^N r_i \leq n_2, r_j = 0 \forall j \in \mathcal{J}} V_N(\mathbf{d}^r, \mathbf{r}), \tag{42}$$

the optimal assignments for *known* variance and covariance function and *constraint* on the pilot units as in Algorithm 2. Denote $\mathbf{D}, \mathbf{R}$ the assignments that solve the experimenter problem in Equation (14). Whenever clear from the context, we define

$$\sigma(i, \mathbf{D}) = \sigma(T_i, \mathbf{D}_i, \sum_{k \in \mathcal{N}_i} \mathbf{D}_k) \tag{43}$$

and similarly for $\eta(i, j, \mathbf{D})$. Finally, note that under Assumption 4.2, since Y is uniformly bounded, also $\sigma^2(\cdot), \eta(\cdot)$ are uniformly bounded by a universal constant (the bound for $\eta(\cdot)$ directly follows from the Cauchy-Swartz inequality).

Preliminary upper bound First observe that $|\mathcal{J}| \leq \mathcal{N}_{\max} m$. Since $n_1 > \mathcal{N}_{\max} m/(n_2/n_1 - 1) \geq |\mathcal{J}|/(n_2/n_1 - 1)$ we have that the constraint $1^\top r = n_2$ is a stricter constraint than $n_1 + |\mathcal{J}| \leq 1^\top r \leq n_2$. We can therefore write

$$
\begin{aligned}
\mathcal{R}_N &= V_N(\mathbf{D},\mathbf{R}) - \min_{\mathbf{r},\mathbf{d}^r,\sum_{i=1}^N r_i = n_2} V_N(\mathbf{d},\mathbf{r}) \\
&\leq V_N(\mathbf{D},\mathbf{R}) - \min_{\mathbf{r},\mathbf{d}^r,n_1+|\mathcal{J}|\leq\sum_{i=1}^N r_i\leq n_2} V_N(\mathbf{d}^r,\mathbf{r}) \\
&\leq V_N(\mathbf{D},\mathbf{R}) - \min_{\mathbf{r},\mathbf{d}^r,n_1+|\mathcal{J}|\leq\sum_{i=1}^N r_i\leq n_2} V_N(\mathbf{d}^r,\mathbf{r}) + V_N(\tilde{\mathbf{D}},\tilde{\mathbf{R}}) - \hat{V}_{n,p}(\mathbf{D},\mathbf{R}) + \hat{V}_{n,p}(\mathbf{D},\mathbf{R}) - V_N(\tilde{\mathbf{D}},\tilde{\mathbf{R}}) \\
&\leq \underbrace{\Big(V_N(\mathbf{D},\mathbf{R}) - \hat{V}_{n,p}(\mathbf{D},\mathbf{R})\Big)}_{(i)} + \underbrace{\Big(\hat{V}_{n,p}(\tilde{\mathbf{D}},\tilde{\mathbf{R}}) - V_N(\tilde{\mathbf{D}},\tilde{\mathbf{R}})\Big)}_{(ii)} + \underbrace{V_N(\tilde{\mathbf{D}},\tilde{\mathbf{R}}) - \min_{\mathbf{r},\mathbf{d}^r,n_1+|\mathcal{J}|\leq\sum_{i=1}^N r_i\leq n_2} V_N(\mathbf{d}^r,\mathbf{r})}_{(iii)}.
\end{aligned}
\tag{44}
$$

The last bound follows from the fact that $\hat{V}_{n,p}(\tilde{\mathbf{D}},\tilde{\mathbf{R}}) \geq \hat{V}_{n,p}(\mathbf{D},\mathbf{R})$, since $(\tilde{\mathbf{D}},\tilde{\mathbf{R}})$ and $(\mathbf{D},\mathbf{R})$ satisfy the same set of constraints, and $(\mathbf{D},\mathbf{R})$ minimizes $\hat{V}_{n,p}(\mathbf{D},\mathbf{R})$. We study each component separately.

Component (i) **and** (ii) We can write

$$
\begin{aligned}
(i) \leq \max_{w_N\in\mathcal{W}_N} \frac{1}{(1^\top\mathbf{R})^2}\sum_{i=1}^N w_N^2(i)R_i\Big(\sigma^2(i,\mathbf{D}) - \hat{\sigma}_p^2(i,\mathbf{D})\Big) \\
+ \frac{1}{(1^\top\mathbf{R})^2}\sum_{i=1}^N\sum_{j\in\mathcal{N}_i} w_N(i)w_N(j)R_iR_j\Big(\eta(i,j,\mathbf{D}) - \hat{\eta}_p(i,j,\mathbf{D})\Big).
\end{aligned}
\tag{45}
$$

Therefore, we obtain

$$
\begin{aligned}
(i) \leq \max_{w_N\in\mathcal{W}_N} \underbrace{\Big|\frac{1}{(1^\top\mathbf{R})^2}\sum_{i=1}^N w_N(i)R_i\Big(\sigma^2(i,\mathbf{D}) - \hat{\sigma}_p^2(i,\mathbf{D})\Big)\Big|}_{(I)} \\
+ \max_{w_N\in\mathcal{W}_N} \underbrace{\Big|\frac{1}{(1^\top\mathbf{R})^2}\sum_{i=1}^N\sum_{j\in\mathcal{N}_i} w_N(i)w_N(j)R_iR_j(\eta(i,j,\mathbf{D}) - \hat{\eta}_p(i,j,\mathbf{D}))\Big|}_{(II)}.
\end{aligned}
\tag{46}
$$

The above term satisfies

$$(46) \lesssim \mathcal{N}_{\max} \sup_{d,s,l,d',s',l'} \Big(\eta(d,s,l,d',s',l') - \hat{\eta}_p(d,s,l',d',s',l')\Big) \Big/ n_1 + \sup_{d,s,l} \Big(\sigma(l,d,s) - \hat{\sigma}_p(l,d,s)\Big) \Big/ n_1. \tag{47}$$

The same reasoning also applies to the term (ii). Therefore, we can write

$$(i) + (ii) \lesssim \mathcal{N}_{\max} m^{-\xi} / n_1.$$

(iii) **Part 1: Lower bound for** $\min V_n(\mathbf{d}, \mathbf{r})$ Finally, consider the term (iii). As a first step, we provide a lower bound to $\min_{\mathbf{r},\mathbf{d}^r, n_1 + |\mathcal{J}| \le \sum_{i=1}^N r_i \le n_2} V_N(\mathbf{d}^r, \mathbf{r})$.

We can write

$$\begin{aligned}
&\min_{\mathbf{r},\mathbf{d}^r, n_1 + |\mathcal{J}| \le \sum_{i=1}^N r_i \le n_2} V_N(\mathbf{d}^r, \mathbf{r}) = \min_{\mathbf{r},\mathbf{d}^r, n_1 + |\mathcal{J}| \le \sum_{i=1}^N r_i \le n_2} \max_{w_N \in \mathcal{W}_N} \\
&\Big(\frac{1}{(\sum_{i=1}^N r_i)^2} \sum_{i \in \mathcal{J}_2} r_i w_N^2(i, \mathbf{d}^r, \mathbf{r}) \sigma^2(i, \mathbf{d}^r, A) + \sum_{j \in \mathcal{N}_i \setminus \mathcal{J}} r_i r_j w_N(i, \mathbf{d}^r, \mathbf{r}) w_N(j, \mathbf{d}^r, \mathbf{r}) \eta(i, j, \mathbf{d}^r, A) \\
&+ \frac{1}{(\sum_{i=1}^N r_i)^2} \sum_{i \in \mathcal{J}} r_i w_N^2(i, \mathbf{d}^r, \mathbf{r}) \sigma^2(i, \mathbf{d}^r, A) + \sum_{j \in \mathcal{N}_i} r_i r_j w_N(i, \mathbf{d}^r, \mathbf{r}) w_N(j, \mathbf{d}^r, \mathbf{r}) \eta(i, j, \mathbf{d}^r, A) \Big) \\
&\ge (A) + (B)
\end{aligned} \tag{48}$$

where

$$\begin{aligned}
(A) = &\min_{\mathbf{r},\mathbf{d}^r, n_1 + |\mathcal{J}| \le \sum_{i=1}^N r_i \le n_2} \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i=1}^N r_i)^2} \Big(\sum_{i \in \mathcal{J}_2} r_i w_N^2(i, \mathbf{d}^r, \mathbf{r}) \sigma^2(i, \mathbf{d}^r, A) \\
&+ \sum_{j \in \mathcal{N}_i \setminus \mathcal{J}} r_i r_j w_N(i, \mathbf{d}^r, \mathbf{r}) w_N(j, \mathbf{d}^r, \mathbf{r}) \eta(i, j, \mathbf{d}^r, A) \Big) \\
(B) = &\min_{\mathbf{r},\mathbf{d}^r, n_1 + |\mathcal{J}| \le \sum_{i=1}^N r_i \le n_2} \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i=1}^N r_i)^2} \Big(\sum_{i \in \mathcal{J}} r_i w_N^2(i, \mathbf{d}^r, \mathbf{r}) \sigma^2(i, \mathbf{d}^r, A) \\
&+ \sum_{j \in \mathcal{N}_i} r_i r_j w_N(i, \mathbf{d}^r, \mathbf{r}) w_N(j, \mathbf{d}^r, \mathbf{r}) \eta(i, j, \mathbf{d}^r, A) \Big)
\end{aligned}$$

$$\mathcal{J}_2 = \{1, \cdots, N\} \setminus \{i, \mathcal{N}_i : P_i = 1\}, \quad \mathcal{J} = \{i, \mathcal{N}_i : P_i = 1\}.$$

(*iii*) **Part two: lower bound decomposed into two groups $\mathcal{J}_2, \mathcal{J}$** We now analyze each component in the right hand side of Equation (48). Notice now that the following term

$$\begin{aligned}(B) &\geq \min_{\mathbf{r},\mathbf{d}^r, n_1+|\mathcal{J}|\leq\sum_{i=1}^N r_i\leq n_2} \frac{1}{(\sum_{i=1}^N r_i)^2}\sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{N}_i} r_i r_j w_N(i,\mathbf{d}^r,\mathbf{r}) w_N(j,\mathbf{d}^r,\mathbf{r})\eta(i,j,\mathbf{d}^r) \\ &\geq -\bar{C}|\mathcal{J}|\max_{i\in\mathcal{J}}|\mathcal{N}_i|/(n_1+|\mathcal{J}|)^2\end{aligned} \tag{49}$$

since the second moment are bounded by Assumption 4.2, for a universal constant $\bar{C} < \infty$. Therefore, the following holds:

$$(48) \geq (A) - \bar{C}|\mathcal{J}|\max_{i\in\mathcal{J}}|\mathcal{N}_i|/(n_1+|\mathcal{J}|)^2. \tag{50}$$

(*iii*) **Part 3: Lower bound to** (A) Next, we provide a lower bound to (A). Observe that in (A) neither the variance nor the covariance component of units which are not in $\mathcal{J}$ appears. Instead, the decision variables of all units in such set affects the objective function only through the constraint and the denominator. The following step is to consider the optimization problem with a slacker constraint, whose objective function is a lower bound of the above objective function. Since $r_i \in \{0,1\}$ we have that the constraint

$$n_1 + |\mathcal{J}| \leq \sum_{i=1}^N r_i = \sum_{i\in\mathcal{J}_2} r_i + \sum_{i\in\mathcal{J}} r_i \leq n_2 \tag{51}$$

is a stricter constraint than

$$n_1 \leq \sum_{i\in\mathcal{J}_2} r_i \leq n_2 \tag{52}$$

since $|\mathcal{J}| \geq \sum_{i\in\mathcal{J}} r_i \geq 0$. Therefore, the following inequality holds:

$$(A) \geq \min_{\mathbf{r},\mathbf{d}^r, n_1 \leq \sum_{i\in\mathcal{J}_2} r_i \leq n_2} \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i=1}^N r_i)^2} \Big(\sum_{i\in\mathcal{J}_2} r_i w_N^2(i,\mathbf{d}^r,\mathbf{r})\sigma^2(i,\mathbf{d}^r) + \sum_{j\in\mathcal{N}_i\setminus\mathcal{J}} r_i r_j w_N(i,\mathbf{d}^r,\mathbf{r}) w_N(j,\mathbf{d}^r,\mathbf{r})\eta(i,j,\mathbf{d}^r)\Big). \tag{53}$$

In the above expression we relaxed the constraint, by allowing the decision variable for units in $\mathcal{J}$ to be unconstrained. Since $r_i, i \in \mathcal{J}$ affect the above expression only through the denominator, we have

$$(A) \geq \min_{\mathbf{r},\mathbf{d}^r, n_1 \leq \sum_{i\in\mathcal{J}_2} r_i \leq n_2} \max_{w_N \in \mathcal{W}_N} \Big(\frac{1}{(\sum_{i\in\mathcal{J}_2} r_i + |\mathcal{J}|)^2} \sum_{i\in\mathcal{J}_2} r_i w_N^2(i,\mathbf{d}^r,\mathbf{r})\sigma^2(i,\mathbf{d}^r) + \sum_{j\in\mathcal{N}_i\setminus\mathcal{J}} r_i r_j w_N(i,\mathbf{d}^r,\mathbf{r}) w_N(j,\mathbf{d}^r,\mathbf{r})\eta(i,j,\mathbf{d})\Big) := (j). \tag{54}$$

Therefore, we can write

$$(iii) \leq V_N(\tilde{\mathbf{D}},\tilde{\mathbf{R}}) - \min_{\mathbf{r},\mathbf{d}^r, n_1 \leq \sum_{i\in\mathcal{J}_2} r_i \leq n_2} \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i\in\mathcal{J}_2} r_i + |\mathcal{J}|)^2} \Big(\sum_{i\in\mathcal{J}_2} r_i w_N^2(i,\mathbf{d}^r,\mathbf{r})\sigma^2(i,\mathbf{d}^r) + \sum_{j\in\mathcal{N}_i\setminus\mathcal{J}} r_i r_j w_N(i,\mathbf{d}^r,\mathbf{r}) w_N(j,\mathbf{d}^r,\mathbf{r})\eta(i,j,\mathbf{d}^r)\Big) + \bar{C}|\mathcal{J}| \max_{i\in\mathcal{J}} |\mathcal{N}_i|/(n_1+|\mathcal{J}|)^2. \tag{55}$$

(iii) **Part 4: Upper bound to $V_N(\tilde{\mathbf{D}},\tilde{\mathbf{R}})$** Consider now the right-hand side in Equation (54). Observe, that we can write

$$(j) = \min_{\mathbf{r},\mathbf{d}^r, n_1 \leq \sum_{i\in\mathcal{J}_2} r_i \leq n_2, r_i = 0 \forall i\in\mathcal{J}} \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i\in\mathcal{J}_2} r_i + |\mathcal{J}|)^2} \Big(\sum_{i\in\mathcal{J}_2} r_i w_N^2(i,\mathbf{d}^r,\mathbf{r})\sigma^2(i,\mathbf{d}^r) + \sum_{j\in\mathcal{N}_i\setminus\mathcal{J}} r_i r_j w_N(i,\mathbf{d}^r,\mathbf{r}) w_N(j,\mathbf{d}^r,\mathbf{r})\eta(i,j,\mathbf{d}^r)\Big),$$

where we added the condition $r_i = 0 \forall i \in \mathcal{J}$. The reason is because the expression does not depend on the selection indicators r_i of those individuals $i \in \mathcal{J}$, i.e., of the

pilot units and their neighbors. Observe that $\mathbf{d}^r$ may still depend on the treatments assigned to the neighbors of individuals in the pilot but not themselves in the pilot. Instead, the treatment assigned to the pilot units does not affect the objective function, which only depends on variance and covariance of individuals which are far apart from the pilot by at least one neighbor. Define

$$\mathbf{D}^{**}, \mathbf{R}^{**} \in \arg \min_{\mathbf{r}, \mathbf{d}^r, n_1 \leq \sum_{i \in \mathcal{J}_2} r_i \leq n_2, r_i = 0 \forall i \in \mathcal{J}} \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i \in \mathcal{J}_2} r_i + |\mathcal{J}|)^2} \Big(\sum_{i \in \mathcal{J}_2} r_i w_N^2(i, \mathbf{d}^r, \mathbf{r}) \sigma^2(i, \mathbf{d}^r) + \sum_{j \in \mathcal{N}_i \setminus \mathcal{J}} r_i r_j w_N(i, \mathbf{d}^r, \mathbf{r}) w_N(j, \mathbf{d}^r, \mathbf{r}) \eta(i, j, \mathbf{d}^r) \Big).$$

Observe that by construction $V_N(\tilde{\mathbf{D}}, \tilde{\mathbf{R}}) \leq V_N(\mathbf{D}^{**}, \mathbf{R}^{**})$, since $\tilde{\mathbf{D}}, \tilde{\mathbf{R}}$ minimize $V_N(\cdot)$, and since $(\mathbf{D}^{**}, \mathbf{R}^{**})$ satisfy the constraints in Equation (42). Therefore, we can write

$$(55) \leq V_N(\mathbf{D}^{**}, \mathbf{R}^{**}) - \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i \in \mathcal{J}_2} R_i^{**} + |\mathcal{J}|)^2} \Big(\sum_{i \in \mathcal{J}_2} R_i^{**} w_N^2(i, \mathbf{D}^{**}, \mathbf{R}^{**}) \sigma^2(i, \mathbf{D}^{**}) + \sum_{j \in \mathcal{N}_i \setminus \mathcal{J}} R_i^{**} R_j^{**} w_N(i, \mathbf{D}^{**}, \mathbf{R}^{**}) w_N(j, \mathbf{D}^{**}, \mathbf{R}^{**}) \eta(i, j, \mathbf{D}^{**}) \Big) + \bar{C} |\mathcal{J}| \max_{i \in \mathcal{J}} |\mathcal{N}_i| / (n_1 + |\mathcal{J}|)^2.$$

Conclusion By simple algebra, and using the same argument for the weights used for (i), we obtain,

$$\begin{aligned}
&V_N(\mathbf{D}^{**}, \mathbf{R}^{**}) - \max_{w_N \in \mathcal{W}_N} \frac{1}{(\sum_{i \in \mathcal{J}_2} R_i^{**} + |\mathcal{J}|)^2} \Big(\sum_{i \in \mathcal{J}_2} R_i^{**} w_N^2(i, \mathbf{D}^{**}, \mathbf{R}^{**}) \sigma^2(i, \mathbf{D}^{**}) \\
&+ \sum_{j \in \mathcal{N}_i \setminus \mathcal{J}} R_i^{**} R_j^{**} w_N(i, \mathbf{D}^{**}, \mathbf{R}^{**}) w_N(j, \mathbf{D}^{**}, \mathbf{R}^{**}) \eta(i, j, \mathbf{D}^{**}) \Big) \\
&\leq 2 \max_{w_n \in \mathcal{W}_n} \Big| \Big(\frac{1}{(\sum_{i \in \mathcal{J}_2} R_i^{**} + |\mathcal{J}|)^2} - \frac{1}{(\sum_{i \in \mathcal{J}_2} R_i^{**})^2} \Big) \Big(\sum_{i \in \mathcal{J}_2} R_i^{**} w_N^2(i, \mathbf{D}^{**}, \mathbf{R}^{**}) \sigma^2(i, \mathbf{D}^{**}) \\
&+ \sum_{j \in \mathcal{N}_i \setminus \mathcal{J}} R_i^{**} R_j^{**} w_N(i, \mathbf{D}^{**}, \mathbf{R}^{**}) w_N(j, \mathbf{D}^{**}, \mathbf{R}^{**}) \eta(i, j, \mathbf{D}^{**}) \Big) \Big|
\end{aligned}$$

by Assumption 4.2 (bounded outcome)

$$\begin{aligned}(iii) &\leq \bar{C} n \mathcal{N}_{\max} \frac{n|\mathcal{J}| + |\mathcal{J}|^2}{(\sum_{i=1}^N R_i^{**})^4} + \bar{C}|\mathcal{J}| \max_{i \in \mathcal{J}} |\mathcal{N}_i|/(n_1 + |\mathcal{J}|)^2 \\ &\leq \bar{C} \mathcal{N}_{\max} \frac{n^2|\mathcal{J}| + n|\mathcal{J}|^2}{\alpha^4 n^4} + \bar{C}|\mathcal{J}|\mathcal{N}_{\max}/(n_1 + |\mathcal{J}|)^2\end{aligned} \tag{56}$$

for a universal constant $\bar{C} < \infty$. Notice now that $|\mathcal{J}| \leq (1 + |\mathcal{N}_{\max}|) \times m$ which completes the proof.

E Inference

Lemma E.1. *(Ross et al., 2011) Let $X_1, ..., X_n$ be random variables such that $\mathbb{E}[X_i^4] < \infty$, $\mathbb{E}[X_i] = 0$, $\sigma^2 = \text{Var}(\sum_{i=1}^n X_i)$ and define $W = \sum_{i=1}^n X_i/\sigma$. Let the collection $(X_1, ..., X_n)$ have dependency neighborhoods $\mathcal{N}_i$, $i = 1, ..., n$ and also define $D = \max_{1 \leq i \leq n} |\mathcal{N}_i|$. Then for Z a standard normal random variable, we obtain*

$$d_W(W, Z) \leq \frac{D^2}{\sigma^3} \sum_{i=1}^n \mathbb{E}|X_i|^3 + \frac{\sqrt{28} D^{3/2}}{\sqrt{\pi}\sigma^2} \sqrt{\sum_{i=1}^n \mathbb{E}[X_i^4]}, \tag{57}$$

where d_W denotes the Wesserstein metric.

Theorem E.2. *Suppose that Assumption 2.1, 4.2 hold, and for all $w_N \in \mathcal{W}_N$, $nV_N(w_N) > 0$. Then for all $w_N \in \mathcal{W}_N$,*

$$\frac{\sqrt{n}(\hat{\Gamma}_n(w_N) - \Gamma_n(w_N))}{\sqrt{nV_N(w_N)}} \to_d \mathcal{N}(0, 1). \tag{58}$$

Proof of Theorem E.2. We prove asymptotic normality after conditioning on the sigma algebra $\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{T}, \mathbf{P})$. Notice that unbiasdness holds by Proposition 3.1. Next, we show that Y_i for all $i : R_i = 1$ are locally dependent, given $\sigma(A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}, \mathbf{P})$. To show this, it suffices to show that

$$\{\varepsilon_i\}_{i:R_i=1} \Big| \sigma(A, \mathbf{D}^R, \mathbf{R}, \mathbf{T}, \mathbf{P})$$

are locally dependent, i.e., form a local dependency graph as described in Ross et al. (2011). Define $\mathcal{H} = \{1, \cdots, N\} \setminus \mathcal{J}$.

The argument is the following. Under Assumption 2.2, unobservables are locally dependent given the adjacency matrix A and covariates $\mathbf{T}$. Since $\mathbf{P}$ only depends on A, it is also follows that unobservables are locally dependent given $(A, \mathbf{T}, \mathbf{P})$ That is,

$$\varepsilon_{1,\cdots,N} \Big| \sigma(A, \mathbf{P}, \mathbf{T})$$

are locally dependent. Consider now the distribution of all unobservables in the set $\mathcal{H}$, given $A, \mathbf{P}, \mathbf{T}$. Here, unobservables are mutually independent on $\mathbf{D}^R, \mathbf{R}$, given $\sigma(A, \mathbf{P}, \mathbf{T})$ and $\mathcal{H}$ is measurable with respect to $\mathbf{P}$. Therefore,

$$\varepsilon_{i \in \mathcal{H}} \Big| \sigma(A, \mathbf{P}, \mathbf{D}^R, \mathbf{R}, \mathbf{T})$$

are locally dependent. Since $\{i : R_i = 1\} \subseteq \mathcal{H}$ the local dependence assumption of unobservables in such a set holds conditional on $A, \mathbf{P}, \mathbf{D}^R, \mathbf{R}, \mathbf{T}$ for such units.

Recall that by Assumption 2.1

$$Y_i = r\Big(D_i, \sum_{k \in \mathcal{N}_i} D_k, T_i, \varepsilon_i\Big). \tag{59}$$

Therefore, given $\sigma(A, \mathbf{P}, \mathbf{D}^R, \mathbf{R}, \mathbf{T})$ outcomes $Y_{\{1,\cdots,n\}}$ are locally dependent. Let

$$X_i := \frac{1}{\sqrt{V_N(w_N)}} w_N(i, \mathbf{D}^R, \mathbf{R}, \mathbf{T}) \Big(Y_i - m(D_i, \sum_{k \in \mathcal{N}_i} D_k, T_i)\Big). \tag{60}$$

Notice that by Proposition 3.1, we have

$$\mathbb{E}[X_i | \sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})] = 0. \tag{61}$$

To prove the theorem we invoke Lemma E.1. In particular, we observe that for

$Z \sim \mathcal{N}(0,1)$, we have

$$\sup_{x\in\mathbb{R}}\left|P\Big(\sum_{i:R_i=1} X_i \le x\Big|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})\Big) - \Phi(x)\right| \le c\sqrt{d_{W|\sigma(\mathbf{D}^R,A,\mathbf{R},\mathbf{P},\mathbf{T})}(\sum_{i:R_i=1} X_i, Z)}. \tag{62}$$

where $d_{W|\sigma(\mathbf{D}^R,A,\mathbf{R},\mathbf{P},\mathbf{T})}(\sum_{i:R_i=1} X_i, Z)$ denotes the Wesserstein metric taken with respect to the conditional marginal distribution of $\sum_{i:R_i=1} X_i$ given $\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})$ and $\Phi(x)$ is the CDF of a standard normal distribution, and $c < \infty$ is a universal constant. To apply Lemma E.1 we take $\sigma^2 = 1$ since X_i already contains the rescaling factor defined in Lemma E.1. In addition, since $nV_N(w_N)$ is strictly bounded away from zero we obtain under Assumption 4.2

$$\mathbb{E}[X_i^4|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})] \le \bar{C}\frac{1}{n^2}, \quad \mathbb{E}[X_i^3|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})] \le \bar{C}\frac{1}{n^{3/2}}. \tag{63}$$

Therefore, the condition in Lemma E.1 are satisfied. Then we obtain

$$\begin{aligned} &d_{W|\sigma(\mathbf{D}^R,A,\mathbf{R},\mathbf{P},\mathbf{T})}(\sum_{i:R_i=1} X_i, Z) \le \mathcal{N}_{\max}^2 \sum_{i:R_i=1} \mathbb{E}[|X_i|^3|\mathbf{D}^R, \mathbf{R}, A, \mathbf{P}, \mathbf{T}] \\ &+ \frac{\sqrt{28}\mathcal{N}_{\max}^{3/2}}{\sqrt{\pi}}\sqrt{\sum_{i:R_i=1}\mathbb{E}[X_i^4|\mathbf{R}, A, \mathbf{D}^R, \mathbf{P}, \mathbf{T}]} \\ &\le \frac{\mathcal{N}_{\max}^2}{n^{1/2}}\bar{C} + \frac{\sqrt{28}\mathcal{N}_{\max}^{3/2}}{\sqrt{\pi n}}\bar{C} \end{aligned} \tag{64}$$

for a universal constan $\bar{C} < \infty$. Since $\mathcal{N}_{\max}^2/n^{1/2} = o(1)$, we obtain

$$\sup_{x\in\mathbb{R}}\left|P\Big(\sum_{i:R_i=1} X_i \le x\Big|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})\Big) - \Phi(x)\right| \le \sqrt{\frac{\mathcal{N}_{\max}^2}{n^{1/2}}\bar{C} + \frac{\sqrt{28}\mathcal{N}_{\max}^{3/2}}{\sqrt{\pi n}}\bar{C}} = o(1) \tag{65}$$

where the latter result is true since the conditions in Lemma E.1 are satisfied pointwise for any $w_N \in \mathcal{W}_N$ and by the property of the Wesserstein metric. To prove that the result also holds unconditionally, we may notice that for some arbitrary measure

μ_N,

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\Big|\int P\Big(\sum_{i:R_i=1} X_i \le x\Big|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})\Big)d\mu_N - \Phi(x)\Big| \\
&\le \sup_{x\in\mathbb{R}}\int \Big|P\Big(\sum_{i:R_i=1} X_i \le x\Big|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})\Big) - \Phi(x)\Big|d\mu_N \qquad (66)\\
&\le \int \sup_{x\in\mathbb{R}}\Big|P\Big(\sum_{i:R_I=1} X_i \le x\Big|\sigma(\mathbf{D}^R, A, \mathbf{R}, \mathbf{P}, \mathbf{T})\Big) - \Phi(x)\Big|d\mu_N = o(1).
\end{aligned}$$

This concludes the proof. □

Corollary. *Theorem A.3 holds.*

Proof. The proof follows similarly to the above theorem with an important modification. We observe that the variables X_i in Equation (60) do not follow a dependence graph since they exhibit M degree dependence. Instead, we construct a graph where two individuals are connected if they are connected by at least M edges in the original graph. In such a graph, the variables X_i as defined in Equation (60) satisfy the local dependence assumption in Lemma E.1. In order for the lemma to apply, we need to show that the maximum degree of such a graph, denoted as $\bar{\mathcal{N}}_M^2$ satisfies the condition $\bar{\mathcal{N}}_M^2/n^{1/2} = o(1)$. This follows under Assumption A.1, since the maximum degree is uniformly bounded. This completes the proof. □

Theorem E.3. *Let Assumptions 2.1, 4.2, 4.3 hold. Then for all $w_N \in \mathcal{W}_N$,*

$$\frac{V_N(w_N)}{\hat{V}_n(w_N)} - 1 \to_p 0. \qquad (67)$$

Proof of Theorem E.3. First, notice that under Assumption 2.1, 2.2, Lemma 3.3 holds, and therefore, the conditional variance can be written as a function of $\sigma(.), \eta(.)$. With an abuse of notation, we will refer to $n = 1^\top \mathbf{R}$.

Next, we prove consistency pointwise for each element in $\mathcal{W}_n$. Throughout the proof we denote $\eta(i,j) = \eta\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k, T_j, D_j, \sum_{k\in N_j} D_k\Big)$ and $\sigma^2(i) = \sigma^2\Big(T_i, D_i, \sum_{k\in\mathcal{N}_i} D_k\Big)$. For notational convenience, we denote $w_N(i,.)$ omitting the last arguments when clear from the context.

We have

$$|nV_N(w_N) - n\hat{V}_n(w_N)| \leq \underbrace{\left|\frac{1}{n}\sum_{i:R_i=1} w_N^2(i)(\hat{\sigma}^2(i) - \sigma^2(i))\right|}_{(a)} + \underbrace{|\frac{1}{n}\sum_{i:R_i=1}\sum_{j\in\mathcal{N}_i} w_N(i,\mathbf{D}^R,\mathbf{R},\mathbf{T})w_N(j)(\hat{\eta}(i,j) - \eta(i,j))|}_{(b)}. \tag{68}$$

Consider first term (a). Then we can write

$$(a) \leq \max_{o\in\{1,\cdots,n\}} w_N(o)^2 \frac{1}{n}\sum_{i:R_i=1}\left|(\hat{\sigma}^2(i) - \sigma^2(i))\right| = o_p(1). \tag{69}$$

Consider now the covariance component. We have

$$\begin{aligned}(b) &\leq \max_{o\in\{1,\cdots,n\}} |w_N(o)|\frac{1}{n}\sum_{i:R_i=1}\left|\sum_{j\in\mathcal{N}_i} w_N(j)(\hat{\eta}(i,j) - \eta(i,j))\right| \\ &\leq \underbrace{\max_{o\in\{1,\cdots,n\}} |w_N(o)|\frac{1}{n}\sum_{i:R_i=1}\left|\sum_{j\in\mathcal{N}_i} w_N(j)(\hat{\eta}(i,j) - \eta(i,j))\right|}_{(J)}.\end{aligned} \tag{70}$$

We have

$$\begin{aligned}(J) \leq &\max_{o\in\{1,\cdots,n\}} |w_N(o)|\frac{1}{n}\sum_{i:|\mathcal{N}_i|\leq L}\left|\sum_{j\in\mathcal{N}_i} w_N(j)(\hat{\eta}(i,j) - \eta(i,j))\right| \\ &+ \max_{o\in\{1,\cdots,n\}} |w_N(o)|\frac{1}{n}\sum_{i:|\mathcal{N}_i|\geq L}\left|\sum_{j\in\mathcal{N}_i} w_N(j)(\hat{\eta}(i,j) - \eta(i,j))\right|.\end{aligned} \tag{71}$$

We have by Holder's inequality and Assumption 4.2,

$$\max_{o\in\{1,\cdots,n\}} |w_N(o)|\frac{1}{n}\sum_{i:|\mathcal{N}_i|\leq L}\left|\sum_{j\in\mathcal{N}_i} w_N(j)(\hat{\eta}(i,j) - \eta(i,j))\right| \leq L\bar{C}\frac{1}{n}\sum_{i:|\mathcal{N}_i|\leq L}\max_j |\hat{\eta}(i,j) - \eta(i,j)| = o_p(1) \tag{72}$$

where the last equality follows by Assumption 4.2, for a constant $\bar{C}$. The second component reads as follows:

$$\max_{o\in\{1,\cdots,n\}} |w_N(o)| \frac{1}{n} \sum_{i:|\mathcal{N}_i|\geq L} \Big| \sum_{j\in\mathcal{N}_i} w_N(j)(\hat{\eta}(i,j)-\eta(i,j)) \Big| \leq \bar{C}\mathcal{N}_{\max} \frac{1}{n} \sum_{i:|\mathcal{N}_i|\geq L} \max_j |\hat{\eta}(i,j)-\eta(i,j)|. \tag{73}$$

By Assumption 4.3, we have that

$$\bar{C}\mathcal{N}_{\max} \frac{1}{n} \sum_{i:|\mathcal{N}_i|\geq L} \max_j |\hat{\eta}(i,j) - \eta(i,j)| \leq O_p(1)\mathcal{N}_{\max} n^{3/4}/n = o_p(1). \tag{74}$$

Here $\max_j |\hat{\eta}(i,j) - \eta(i,j)| = O_p(1)$ since $\hat{\eta}$ converges uniformly to η . Uniform consistency follows from the union bound, since $|\mathcal{W}_n|$ is finite dimensional. The proof is complete by the fact that $nV_N(w_N) > 0$ and the continuous mapping theorem. □

Corollary. *Theorem 4.2 holds.*

Proof. The proof follows from Theorem E.2 and Theorem E.3 by Slutsky theorem. □

F Optimization: MILP for Difference in Means Estimators

In this sub-section we discuss the optimization algorithm for difference in means estimators.

To show that the optimization problem admits a mixed-integer linear program formulation, we first introduce the following proposition, which follows similarly to what discussed in Viviano (2019).

Lemma F.1. *(Viviano, 2019) Any function g_i that depends on D_i and $\sum_{k\in\mathcal{N}_i} D_k$ can be written as*

$$g_i(D_i, \sum_{k\in\mathcal{N}_i} D_k) = \sum_{h=0}^{|\mathcal{N}_i|} (g_i(1,h) - g_i(0,h))u_{i,h} + (t_{i,h,1} + t_{i,h,2} - 1)g_i(0,h), \tag{75}$$

where $u_{i,h}, t_{i,h,1}, t_{i,h,2}$ are defined by the following linear inequalities.

$$
\begin{aligned}
&(A) \quad \frac{D_i + t_{i,h,1} + t_{i,h,2}}{3} - 1 < u_{i,h} \leq \frac{D_i + t_{i,h,1} + t_{i,h,2}}{3}, \quad u_{i,h} \in \{0,1\} \quad \forall h \in \{0, ..., |\mathcal{N}_i|\}, \\
&(B) \quad \frac{(\sum_k A_{i,k} D_k - h)}{|\mathcal{N}_i| + 1} < t_{i,h,1} \leq \frac{(\sum_k A_{i,k} D_k - h)}{|\mathcal{N}_i| + 1} + 1, \quad t_{i,h,1} \in \{0,1\}, \quad \forall h \in \{0, ..., |\mathcal{N}_i|\} \\
&(C) \quad \frac{(h - \sum_k A_{i,k} D_k)}{|\mathcal{N}_i| + 1} < t_{i,h,2} \leq \frac{(h - \sum_k A_{i,k} D_k)}{|\mathcal{N}_i| + 1} + 1, \quad t_{i,h,2} \in \{0,1\}, \quad \forall h \in \{0, ..., |\mathcal{N}_i|\}.
\end{aligned} \tag{76}
$$

Proof. We define the following variables:

$$
t_{i,h,1} = 1\{\sum_{k \in \mathcal{N}_i} D_k \geq h\}, \quad t_{i,h,2} = 1\{\sum_{k \in \mathcal{N}_i} D_k \leq h\}, \quad h \in \{0, ..., |\mathcal{N}_i|\}.
$$

The first variable is one if at least h neighbors are treated, and the second variable is one if at most h neighbors are treated.

Since each unit has $|\mathcal{N}_i|$ neighbors and zero to $|\mathcal{N}_i|$ neighbors can be treated, there are in total $\sum_{i=1}^{n}(2|\mathcal{N}_i| + 2)$ of such variables.

The variable $t_{i,h,1}$ can be equivalently be defined as

$$
\frac{(\sum_k A_{i,k} D_k - h)}{|\mathcal{N}_i| + 1} < t_{i,h,1} \leq \frac{(\sum_k A_{i,k} D_k - h)}{|\mathcal{N}_i| + 1} + 1, \quad t_{i,h,1} \in \{0,1\}. \tag{77}
$$

The above equation holds for the following reason. Suppose that $h < \sum_k A_{i,k} D_k$. Since $\frac{(\sum_k A_{i,k} D_k - h)}{|\mathcal{N}_i|+1} < 0$, the left-hand side of the inequality is negative and the right hand side is positive and strictly smaller than one. Since $t_{i,h,1}$ is constrained to be either zero or one, in such case, it is set to be zero. Suppose now that $h \geq \sum_k A_{i,k} D_k$. Then the left-hand side is bounded from below by zero, and the right-hand side is bounded from below by one. Therefore $t_{i,h,1}$ is set to be one. Similarly, we can write

$$
\frac{(h - \sum_k A_{i,k} D_k)}{|\mathcal{N}_i| + 1} < t_{i,h,2} \leq \frac{(h - \sum_k A_{i,k} D_k)}{|\mathcal{N}_i| + 1} + 1, \quad t_{i,h,2} \in \{0,1\}. \tag{78}
$$

By definition,

$$t_{i,h,1} + t_{i,h,2} = \begin{cases} 1 \text{ if and only if } \sum_{k \in \mathcal{N}_i} D_k \neq h \\ 2 \text{ otherwise .} \end{cases} \tag{79}$$

Therefore, we can write

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{h=0}^{|\mathcal{N}_i|} (g_i(1,h) - g_i(0,h)) D_i (t_{i,h,1} + t_{i,h,2} - 1) + (t_{i,h,1} + t_{i,h,2} - 1) g_i(0,h). \tag{80}$$

Finally, we introduce the variable $u_{i,h} = D_i(t_{i,h,1} + t_{i,h,2} - 1)$. Since $D_i, t_{i,h,1}, t_{i,h,2} \in \{0,1\}$ it is easy to show that such variable is completely determined by the above constraint. This completes the proof. □

We first start from the case where $|\mathcal{W}_N| = 1$ and then we extend to the case of multiple estimators. By Lemma F.1, we showcase that each function of the individual and neighbors' treatment assignment can be written as a linear function of the decision variables under linear constraints.

We define

$$\begin{aligned} &\tilde{\sigma}_i^2(D_i, \sum_{k \in \mathcal{N}_i} D_k) = \sigma(T_i, D_i, \sum_{k \in \mathcal{N}_i} D_k) \\ &\tilde{\eta}_{i,j}(D_i, \sum_{k \in \mathcal{N}_i} D_k, D_j, \sum_{k \in N_j} D_k) = \eta(T_i, D_i, \sum_{k \in \mathcal{N}_i} D_k, T_j, D_j, \sum_{k \in N_j} D_k) \end{aligned}$$

the variance function and $\tilde{\eta}_{i,j}(.)$ the covariance for unit i and j, given their number of neighbors and the observed treatment assignments.

We define

$$\begin{aligned} v_i^1(D_i, \sum_{k \in \mathcal{N}_i} D_k) &= 1\{D_i = d_1, \sum_{k \in \mathcal{N}_i} D_k = s_1, T_i = l\}, \\ v_i^0(D_i, \sum_{k \in \mathcal{N}_i} D_k) &= 1\{D_i = d_0, \sum_{k \in \mathcal{N}_i} D_k = s_0, T_i = l\}. \end{aligned} \tag{81}$$

The objective function reads as follows.

$$\sum_{i:R_i=1} R_i\Big(\frac{v_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)\tilde{\sigma}_i(D_i,\sum_{k\in\mathcal{N}_i}D_k)}{\sum_{i:R_i=1}R_iv_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}\Big)^2+R_i\Big(\frac{v_i^0(D_i,\sum_{k\in\mathcal{N}_i}D_k)\tilde{\sigma}_i(D_i,\sum_{k\in\mathcal{N}_i}D_k)}{\sum_{i:R_i=1}R_iv_i^0(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}\Big)^2$$
$$+\frac{R_iv_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)}{\sum_{i:R_i=1}R_iv_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}\times$$
$$\times\sum_{j\in\mathcal{N}_i}R_j\Big(\frac{v_j^1(D_j,\sum_{k\in N_j}D_k)}{\sum_{i:R_i=1}R_iv_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}-\frac{v_j^0(D_j,\sum_{k\in N_j}D_k)}{\sum_{i:R_i=1}R_iv_i^0(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}\Big)\tilde{\eta}_{i,j}\Big(D_i,\sum_{k\in\mathcal{N}_i}D_k,D_j,\sum_{k\in N_j}D_k\Big)$$
$$-\frac{R_iv_i^0(D_i,\sum_{k\in\mathcal{N}_i}D_k)}{\sum_{i:R_i=1}R_iv_i^0(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}\times$$
$$\times\sum_{j\in\mathcal{N}_i}R_j\Big(\frac{v_j^1(D_j,\sum_{k\in N_j}D_k)}{\sum_{i:R_i=1}R_iv_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}-\frac{v_j^0(D_j,\sum_{k\in N_j}D_k)}{\sum_{i:R_i=1}R_iv_i^0(D_i,\sum_{k\in\mathcal{N}_i}D_k)/n}\Big)\tilde{\eta}_{i,j}\Big(D_i,\sum_{k\in\mathcal{N}_i}D_k,D_j,\sum_{k\in N_j}D_k\Big). \tag{82}$$

We now introduce the following auxiliary variables: $n\times\sum_{i:R_i=1}|\mathcal{N}_i|$ variables $t_{i,h,1}=1\{\sum_{k\in\mathcal{N}_i}D_k\geq h\}$ and $n\times\sum_{i:R_i=1}|\mathcal{N}_i|$ variables $t_{i,h,2}=1\{\sum_{k\in\mathcal{N}_i}D_k\leq h\}$. We define $\tilde{t}_{i,h}=t_{i,h,1}+t_{i,h,2}-1$ and we define $u_{i,h}=D_i\times\tilde{t}_{i,h}$. Such variables are fully characterize by the two linear constraints for each variable as discussed in Lemma F.1 and the 0-1 constraint for each variable. By Lemma F.1, each function or product of functions of the variables $(D_i,\sum_{k\in\mathcal{N}_i}D_k)$ can now be described as a linear function of these new decision variables. Consider for example, $(v_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)\tilde{\sigma}_i(D_i,\sum_{k\in\mathcal{N}_i}D_k))^2$ first. Then such function is rewritten as

$$(v_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)\tilde{\sigma}_i(D_i,\sum_{k\in\mathcal{N}_i}D_k))^2=\sum_{h=1}^{|\mathcal{N}_i|}(v_i^1(1,h)^2\tilde{\sigma}_i(1,h)^2-v_i^1(0,h)^2\tilde{\sigma}_i(0,h)^2)u_{i,h}+v_i^1(0,h)^2\tilde{\sigma}_i(0,h)^2\tilde{t}_{i,h}. \tag{83}$$

Similarly, consider the following function

$$K(D_i,D_j,\sum_{k\in\mathcal{N}_i}D_k,\sum_{k\in N_j}D_k):=v_i^1(D_i,\sum_{k\in\mathcal{N}_i}D_k)v_j^1(D_j,\sum_{k\in N_j}D_k)\tilde{\eta}_{i,j}\Big(D_i,\sum_{k\in\mathcal{N}_i}D_k,D_j,\sum_{k\in N_j}D_k\Big). \tag{84}$$

By Lemma F.1, the function can be written as

$$\sum_{h=0}^{|\mathcal{N}_i|} \Big(K(1, D_j, h, \sum_{k \in N_j} D_k) - K(0, D_j, h, \sum_{k \in N_j} D_k) \Big) u_{i,h} + \tilde{t}_{i,h} K(0, D_j, h, \sum_{k \in N_j} D_k). \quad (85)$$

We can now linearize the function and obtain the following equivalent formulation

$$\begin{aligned} \sum_{h'=0}^{|N_j|} \Big(\sum_{h=0}^{|\mathcal{N}_i|} & \Big(K(1,1,h,h') - K(0,1,h,h) \Big) u_{i,h} + \tilde{t}_{i,h} K(0,1,h,h') \\ & - \Big(K(1,0,h,h') - K(0,0,h,h) \Big) u_{i,h} + \tilde{t}_{i,h} K(0,0,h,h') \Big) u_{j,h'} \\ & + \Big(K(1,0,h,h') - K(0,0,h,h) \Big) u_{i,h} \tilde{t}_{j,h'} + \tilde{t}_{i,h} K(0,0,h,h') \tilde{t}_{j,h'} \Big). \end{aligned} \quad (86)$$

which is quadratic in the decision variables, as defined in Lemma F.1. Therefore, each function in the numerators and denominators of Equation (82) can be written as a linear or quadratic function in the decision variables $D_i, u_{i,h}, \tilde{t}_{i,h}$. We now linearize the quadratic expressions in the numerator and denominators, to show that also quadratic expression have a linear formulation. To do so we introduce a new set of variables that we denote as

$$A_{i,j,h',h'} = u_{i,h} u_{j,h'}, \quad B_{i,j,h',h'} = u_{i,h} \tilde{t}_{j,h'}, \quad C_{i,h,h',h} = \tilde{t}_{i,h} \tilde{t}_{j,h'}. \quad (87)$$

Since each of the above variable takes values in $\{0, 1\}$, such variables can be expressed with linear constraints. For instance, $A_{i,j,h',h'}$ is defined as follows.

$$\frac{u_{i,h} + u_{j,h'}}{2} - 1 < A_{i,j,h',h'} \leq \frac{u_{i,h} + u_{j,h'}}{2}, \quad A_{i,j,h',h'} \in \{0, 1\}. \quad (88)$$

In fact, if both $u_{i,h}, u_{j,h'}$ are both equal to one, the left hand size is zero, and under the 0-1 constraint, the resulting variable is equal to one. This follows similarly also for the other variables. Finally, notice that since also $R_i \in \{0, 1\}$, the product of R_i for any other 0-1 variable can be similarly linearized. Therefore, the above problem reads as a mixed-integer *fractional* linear program. By the linear representation of

fractional linear programming discussed in Charnes and Cooper (1962), the proof completes for the case where $|\mathcal{W}_N| = 1$.

To solve the optimization problem over multiple weights $\mathcal{W}_N$, we can add an auxiliary variables λ, and solve the following program

$$\min \lambda, \quad \lambda \geq f_{w_N} \forall w_N \in \mathcal{W}_N \tag{89}$$

where f_{w_n} denotes the linearized objective function for each $w_N \in \mathcal{W}_N$.