Identification and estimation of treatment and interference effects in observational studies on networks

Laura Forastiere[*], Edoardo M. Airoldi[†], and Fabrizia Mealli[‡]

[*]*Yale University,* [†]*Harvard University and* [‡]*University of Florence*

Abstract

Causal inference on a population of units connected through a network often presents technical challenges, including how to account for interference. In the presence of local interference, for instance, potential outcomes of a unit depend on its treatment as well as on the treatments of other local units, such as its neighbors according to the network. In observational studies, a further complication is that the typical unconfoundedness assumption must be extended—say, to include the treatment of neighbors, and individual and neighborhood covariates—to guarantee identification and valid inference. Here, we propose new estimands that define treatment and interference effects. We then derive analytical expressions for the bias of a naive estimator that wrongly assumes away interference. The bias depends on the level of interference but also on the degree of association between individual and neighborhood treatments. We propose an extended unconfoundedness assumption that accounts for interference, and we develop new covariate-adjustment methods that lead to valid estimates of treatment and interference effects in observational studies on networks. Estimation is based on a generalized propensity score that balances individual and neighborhood covariates across units under different levels of individual treatment and of exposure to neighbors' treatment. We carry out simulations, calibrated using friendship networks and covariates in a nationally representative longitudinal study of adolescents in grades 7-12, in the United States, to explore finite-sample performance in different realistic settings.

1   Introduction

1.1   Motivation

Many experimental and observational studies are affected by the presence of interference among units. *Interference* is said to be present when a treatment, exposure or intervention, on one unit has an effect on the response of another unit (Cox, 1958). This phenomenon can be due to social or physical interaction among units. For instance, widespread use of preventive measures for infectious diseases, such as bed nets for malaria control or vaccines for other diseases, may benefit unprotected individuals by reducing the reservoir of infection and by affecting the vector of transmission (Binka et al., 1998; Howard et al., 2000; Hawley et al., 2003). In education, students assigned to attend a tutoring program might interact with other students who were not assigned to the program and spillover effects may arise thanks to the transmission of knowledge among peers. In social media, advertisements are shared by users of the same network or the behavior of a single user might affect other users. The effects of interference are typically referred to as *spillover effects*, in economics, or as *peer influence effects*, in social sciences. A unit's treatment might affect other units' outcomes through a variety of mechanisms: the spread of the treatment; through a change in its own outcome which, in turn, affect the outcomes of other units; or through other pathways involving intermediate variables. Regardless of the mechanism through which it takes place, this dependence between units' treatments and outcomes poses statistical challenges, because potential outcomes for each unit must be indexed also by the treatment received by other units.

When the target estimand is the average (individual) treatment effect, interference is seen as a nuisance and has to be taken into account, or dealt with in some way, to prevent bias (e.g., see Sävje, et al., 2017). However, scientists and people making decisions about policies, interventions, products, and campaigns are also interested in the extent to which individuals are influenced by others, in order to leverage or prevent mechanisms of interference. In settings with limited resources, beneficial spillover effects could be leveraged to increase the cost-effectiveness of an intervention. Similarly, marketers are interested in exploiting how specific individuals can have a large influence on their friends' decisions to promote

the purchase of some products. On the contrary, when spillover effects are detrimental, researchers are interested in investigating the extent to which and how interference play a role, in order to preserve the effectiveness of the program.

In this paper, we consider the problem of estimating both the causal effect of the individual treatment and the spillover effect, in situations where interference occurs between units that are related on a known network and the assignment mechanism of the treatment is not known. Our first aim is to formulate the problem under the potential outcome framework (Rubin, 1974, 1986) and to define causal estimands that are best suited for network data. We then provide identifying assumptions and discuss their plausibility in dependent data. Our second aim is to quantify the bias of a naive estimator that neglects the presence of interference. Understanding the sources of bias is crucial in both the study design phase when information on connections is not easy to collect and in the final stage of drawing conclusions from the results. Our third aim is to tackle the confounding problem of observational studies in networks by extending the propensity score approach to settings with interference from first order neighbors. By revealing the balancing properties of this generalized propensity score we lay the basis for developing covariate adjustment methods for the estimation of treatment and spillover effects in observational network data. Finally, we rely on these results to propose a propensity score-based method that builds upon the literature on generalized propensity scores for continuous treatments (Hirano & Imbens, 2004).

1.2   Related work

When assessing causal effects, the standard approach is to rule out the presence of interference. The assumption of no interference is also called Individualistic Treatment Response (ITR) (Manski, 2013), or, combined with the assumption of no hidden versions of treatment, it is referred to as the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980). In many situations, however, no interference may not be plausible. The risk of this assumption is demonstrated by Sobel (2006). Examining randomized housing mobility studies in which households in poor areas are financed to relocate to better neighborhoods, Sobel shows that ignoring interference can lead to entirely wrong conclusions about the effectiveness of the program. In fact, the observed mean difference between treated and control units can be decomposed into a total effect,i.e., the sum of the direct effect of receiving the treatment and the spillover effect from other units' treatment, for treated units and a spillover effect for control units.

In both cases, when spillover effects are seen as an inconvenience for the estimation

of the treatment effect or when they are the target estimands, different strategies to get unbiased estimates of either the treatment or the spillover effects have been proposed, in both the design and the estimation phases. These strategies depend on the extension of the interference mechanism. When individuals can be partitioned into groups, it is often plausible to assume that interference occurs within groups but not across groups. This assumption is referred to as *partial interference* (Sobel, 2006). Under this assumption, cluster randomized trials are sometimes designed to rule out the presence of interference between treated and untreated clusters. Furthermore, to estimate both the treatment and spillover effects, one possible design is a sequential two stage randomization, where in the first stage groups are randomized to different treatment allocation strategies and in the second stage individuals are randomized to treatment or control conditional on the strategy assigned to their group in the first stage (Hudgens & Halloran, 2008). Some more recent work has examined interference when units are connected along more complicated network structures. Design strategies involve rearranging assignment of treatment to subjects in a manner that, incorporating information on network connectivity, is able to make a trade-off between bias and variance (Toulis & Kao, 2013; Ugander et al., 2013; Eckles et al., 2014).

With regard to inference strategies, there are various suggestions on how to conduct causal inference under interference for randomized experiments. Rosenbaum (2007) uses randomization inference to both test and estimate the total treatment effect, that is, the sum of treatment and spillover effect, even in the presence of interference. Bowers et al. (2013) and Aronow (2012) propose the use of randomization inference to separately test for interference. Building on these works, Athey et al. (2015) provide a general framework that applies to a much larger class of non-sharp null hypotheses. Aronow & Samii (2017) perform inference using the Horvitz-Thompson estimator, which requires that inclusion probabilities for complex network sampling designs can be computed.

The literature on spillover effects is also rapidly evolving for observational studies. However, most of the recently proposed methods rely on the assumption of partial interference. Hong & Raudenbush (2006), who evaluate the effect of grade retention of low-achieving children on test scores, use a parametric multilevel approach to mimic a two-stage experiment and base their analysis on the assumption that the extent to which each child is affected by the retention of other children only depends on whether they are exposed to a low versus high proportion of kindergartners in the school. Another similar example is found in Verbitsky-Savitz & Raudenbush (2012), evaluating the effects of community policing program on neighborhoods crime rates using a three-level generalized hierarchical linear model. Building on the work by Hudgens & Halloran (2008), Tchetgen Tchetgen & VanderWeele

(2012) proposed inverse-probability-weighted estimators (IPW) for treatment effects also in the presence of partial interference. Perez-Heydrich et al. (2014) showed the performance of these IPW estimators using simulation studies and applied them to analyze an individually-randomized, placebo-controlled trial of cholera vaccination.

On the other hand, work on causal inference for observational network data is still in its infancy. Liu et al. (2016) extend these IPW estimators in the presence of general forms of interference on networks. They focus on direct and indirect effects based on Bernoulli allocation strategies similar to those defined by Hudgens & Halloran (2008). Although their estimators allows for general forms of interference, their estimator and their asymptotic results are proved under partial interference and clustered data. Van der Laan (2014) and then Sofrygin & van der Laan (2017) propose a TMLE estimator for similar treatment and spillover effects and prove asymptotic results under IID assumptions. Finally Ogburn et al. (2017) extend this TMLE estimator to allow for dependence due to both contagion and homophily and derive asymptotic results that allow the number of ties per node to increase as the network grows.

1.3   Contributions

We start by providing a general formalization of the problem of interference in networks under the potential outcome framework. As in previous works (Van der Laan, 2014; Aronow & Samii, 2017), we replace SUTVA with an assumption that limits the propagation of treatment to immediate neighbors, ruling out the influence by neighbors' of neighbors, and simplifies the mechanism of interference, in that only a summary of the neighbors' treatment vector matters. This assumption will reduce the number of possible potential outcomes, which are indexed only by the treatment received by the unit, the individual treatment, and a summary of the treatment received by his neighbors, the neighborhood treatment. Under this framework, we provide new causal estimands for treatment and spillover effects. We first define causal estimands as average comparisons of potential outcomes under different values of the treatment of both the unit and his neighbors. We then marginalize these estimands over the observed distribution of the neighbors' treatment. When compared to the estimands in Van der Laan (2014) and in Liu et al. (2016), where the treatment vector is drawn from a hypothetical intervention, our estimands have a more descriptive nature.

We provide an uncounfoundedness assumption that is similar to the conditional exchangeability in Liu et al. (2016) and the randomization assumption in Van der Laan (2014). However, our assumption is weaker because it only implies the independence of the treatments

with the potential outcomes of each single unit. In addition we prove identification results and discuss the implications and the plausibility of this assumption in network settings.

Key contributions of this paper is the derivation of bias formulas for the treatment effect when SUTVA is wrongly assumed. Our results differ from the one provided by Sobel (2006) in that we deal with observational studies and we explicitly show the two key factors driving the bias: the extent to which a unit's potential outcome is affected by the treatment received by his neighbors and the residual association between the individual and the neighborhood treatment after conditioning for covariates. These results are fundamental for applied researchers as they demystify the misconception that bias resulting from ignoring interference solely depends on the level of interference itself. A simulation study is used to validate the analytical results in different scenarios, with different level of interference and different correlation between the individual and the neighborhood treatment.

In the second part of the paper, we provide a clear definition of the propensity score under interference and elucidate its balancing properties. Under interference on the immediate neighborhood, balance must be achieved across arms defined not only by the individual treatment but also by the neighborhood treatment. In this spirit, under the neighborhood interference, we first define the joint propensity score as the probability of assignment to a particular individual and neighborhood treatment given the observed covariates. We discuss the inclusion of different kinds of covariates representing individual characteristics, neighbors' characteristics but also network properties. We then prove the balancing properties of this joint propensity score and we show that unconfoundedness holds not only conditional on the joint propensity score but also conditional on the individual and neighborhood propensity scores. Relying on this unconfoudedness assumption, we propose a joint propensity score-based estimator, based on subclassification on the individual binary-treatment propensity score and parametric adjustment for the neighborhood multivalued-treatment propensity score.

The remainder of this article is organized as follows. In Section 2, we introduce notation and formalize the problem by providing general definitions of estimands and identifying assumptions. The bias for the treatment effect when we wrongly assume no interference is derived in Section 3. In Section 4 we give a formal definition and show the balancing properties of the generalized propensity score under neighborhood interference. In Section 5 we present the joint propensity score-based estimator for main effects and spillover effects. The simulation study, which validates the bias formulas and assesses the performance of our estimator, is described in Section 6. Finally, in Section 7 we discuss our findings and highlight potential future work. In Appendix A we develop propensity score-based estimators

for conditional effects, that is, main and spillover effects among units that are observed under specific values of individual and neighborhood treatments. In Appendix B we detail the data generating model for the simulation study, leveraging Add-Health network and covariate data. Details of the proposed estimator, the specific models used for the simulation study, as well as a proposed approach on how to conduct statistical inference, are presented in Appendix C. Proofs of the theorems and corollaries are reported in Appendix D.

2 Interference based on the exposure to neighborhood treatment

2.1 Notation

Let us denote an undirected network G as a pair $(\mathcal{N}, \mathbb{E})$, where $\mathcal{N}$ is a set of nodes (units) with cardinality N, and $\mathbb{E}$ is a set of edges with generic element $(i, j) = (j, i)$, which represents the presence of a link between unit i and unit j. This network G is our population of interest.

Let us define a partition of the set $\mathcal{N}$ around node i as $(i, \mathcal{N}_i, \mathcal{N}_{-i})$, where the set $\mathcal{N}_i$ has cardinality N_i and contains all nodes j connected to i by an edge in $\mathbb{E}$, i.e., the *neighboring* nodes of i in G (e.g. friends or neighbors), and the set $\mathcal{N}_{-i}$ contains all nodes other than i that are not in $\mathcal{N}_i$. $\mathcal{N}_i$ is referred to as the *neighborhood* of unit i, regardless of whether the presence of edges is defined by physical proximity, friendship or any other type of relationship. For consistency with the literature of social networks, the number of neighbors N_i is referred to as *degree* of unit i. It is worth noting that our proposed formulation and methods can be easily extended to directed networks. Moreover, here we assume connections to be fixed and known. This assumption is more plausible when the type of relationship underlying the mechanism of interference is a concrete and objective concept, and information on connections between units can be objectively gathered from available databases. An example of objective relationship is the one defined by geographic proximity, membership to the same group, collaboration or networking between firms or units, friendship on social media, etc. However, real world networks are often uncertain, as social interactions between individuals may be either unobservable, or measured with error. Here we assume that information on social ties is correctly measured. Extensions to include uncertainty in the social network are possible, but are beyond the scope of this paper.

Let now $Z_i \in \{0, 1\}$ be a binary variable representing the treatment assignment to unit i

and $Y_i \in \mathcal{Y}$ the observed outcome of unit i. $\mathbf{Z}$ and $\mathbf{Y}$ are the corresponding vectors in the whole population $\mathcal{N}$. For each unit i, the partition $(i, \mathcal{N}_i, \mathcal{N}_{-i})$ defines the following partitions of the treatment and outcome vectors: $(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}})$ and $(Y_i, \mathbf{Y}_{N_i}, \mathbf{Y}_{N_{-i}})$. Finally, let $\mathbf{X}_i \in \mathcal{X}$ be a vector of K covariates for unit i. In social network data, $\mathbf{X}_i$ can be decomposed into two subvectors: individual covariates, denoted with $\mathbf{X}_i^{ind} \in \mathcal{X}^{ind}$, and neighborhood covariates, denoted with $\mathbf{X}_i^{neigh} \in \mathcal{X}^{neigh}$. $\mathbf{X}_i^{ind}$ denotes the set of individual-level characteristics (e.g., demographic factors, socio-economic factors, health status, ...) or contextual covariates (e.g., environment socio-economic factors, geographical factors, ...). Conversely, $\mathbf{X}_i^{neigh}$ may include three types of neighborhood-level covariates: variables representing the structure of the neighborhood $\mathcal{N}_i$ (the degree N_i, the topology, etc.), network properties at node-level representing the position of unit's neighborhood in the graph (e.g., centrality, betweenness,the number of shared neighbors, ...), and aggregational covariates as functions of individual-level covariates in the neighborhood, i.e., $X_{ik}^{neigh} = h_{ik}(\mathbf{X}_{k\mathcal{N}_i}^{ind}).$, where $\mathbf{X}_{k\mathcal{N}_i}^{ind}$ is the vector of the k individual characteristic in the neighborhood of unit i and $h_{ik}(\cdot)$ is a summarizing function. For instance, if the covariate X_{ik}^{ind} is sex of unit i then X_{ik}^{neigh} can be the proportion of males among the neighbors of unit i. Similarly, if the covariate X_{ik}^{ind} is income of unit i then X_{ik}^{neigh} can be the average income among the neighbors of unit i. $h_{ik}(\cdot)$ could also be a function of comparison between the covariate If the covariate X_{ik}^{ind} of unit i and the same covariate for his neighbors, e.g. if X_{ik}^{ind} is town of residence of unit i then X_{ik}^{neigh} can be the number of friends living in the same town. With the distinction between $\mathbf{X}_i^{ind}$ and $\mathbf{X}_i^{neigh}$ we want to emphasize that in network settings we need to include covariates representing the structure of the neighborhood or the type of neighbors, i.e., $\mathbf{X}_i^{neigh}$, that are not usually taken into account in settings without interference or in settings with partial interference.

2.2   Potential outcomes and neighborhood interference

Here, we extend the potential outcomes notation to include the presence of (network) interference. In general, the outcome observable at node i is a function of the entire treatment assignment vector $\mathbf{Z}$ and can be written as $Y_i(\mathbf{Z})$. As pointed out by Rubin (1986), this potential outcome is well defined only if the following assumption holds:

Assumption 1 (No Multiple Versions of Treatment (Consistency))**.**

$$Y_i = Y_i(\boldsymbol{Z})$$

This says that the mechanism used to assign the treatments does not matter and assigning the treatments in a different way does not constitute a different treatment.

This assumption is the first component of a fundamental assumption usually made in the potential outcomes approach to causal inference: the stable unit treatment value assumption (SUTVA) (Rubin, 1980, 1986). The second component of SUTVA is the more critical assumption of no interference between individuals (Cox, 1958). If we write the assignment vector $\mathbf{Z}$ as $(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}})$, this assumption states that $Y_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = Y_i(Z_i, \mathbf{Z}'_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_{-i}})$, $\forall\ \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}'_{\mathcal{N}_{-i}}$. In contrast, in this paper we are interested in relaxing the no interference assumption and allow the existence of network interference. Formally, we replace the no interference assumption by a new assumption of interference within the neighborhood.

Assumption 2 (Neighborhood Interference). *Given a function $g_i : \{0,1\}^{N_i} \to \mathcal{G}_i$, $\forall i \in \mathcal{N}$, $\forall\ \mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}'_{\mathcal{N}_{-i}}$ and $\forall\ \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_i} : g_i(\mathbf{Z}_{\mathcal{N}_i}) = g_i(\mathbf{Z}'_{\mathcal{N}_i})$, the following equality holds:*

$$Y_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = Y_i(Z_i, \mathbf{Z}'_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_{-i}})$$

Assumption 1 and Assumption 2 together can be referred to as Stable Unit Treatment on Neighborhood Value Assumption (SUTNVA, pronounced *Sut-Ton-Va*). Assumption 2 rules out the dependence of the outcome of unit i, Y_i, from the treatment received by units outside his neighborhood, i.e., $\mathbf{Z}_{\mathcal{N}_{-i}}$, but allows Y_i to depend on the treatment received by his neighbors, i.e., $\mathbf{Z}_{\mathcal{N}_i}$. Moreover, this dependence is assumed to be through a specific function $g_i(\cdot)$. Let us denote with G_i the variable resulting from applying this function to the neighborhood treatment vector, i.e., $G_i = g_i(\mathbf{Z}_{\mathcal{N}_i})$. G_i can be, for instance, the simple number or the proportion of treated neighbors, i.e., $G_i = \sum_{j \in \mathcal{N}_i} Z_j$ or $G_i = \frac{\sum_{j \in \mathcal{N}_i} Z_j}{N_i}$, respectively. If we think that different neighbors can affect the outcome of unit i in a different way, then G_i can also be a weighted sum of the treatment vector, $G_i = \sum_{j \in \mathcal{N}_i} w_{ij} Z_j$, where the weights can be covariates or the elements of the adjacency matrix in a weighted network. For example, the spillover effect of the treatment received by the best friends can be higher than the one of other friends. We represent this situation by making the potential outcome of unit i depend on a variable G_i, which results from giving more weight to the treatment of the best friends. The domain of G_i will depend on how the function $g_i(\cdot)$ is defined. For example, if we consider the simple number of treated neighbors, then $\mathcal{G} = \{0, 1, \ldots, N_i\}$. Note that G_i can be any function of treatment vector in the neighborhood $\mathbf{Z}_{\mathcal{N}_i}$, even the identity function. This formulation is similar to the 'exposure mapping' introduced by Aronow & Samii (2017) and the one in Van der Laan (2014). Here we assume the function $g_i(\cdot)$ to be known and well-specified. A discussion on consequences of misspecification of this function is beyond

the scope of this paper and can be found in Aronow & Samii (2017).

2.3 Individual and neighborhood treatments

Under Assumption 2, each unit is subject to two treatments: the *individual treatment*, Z_i, and the *neighborhood treatment*, G_i. We define as *joint treatment* the bivariate treatment to which each unit is exposed. To fix ideas, think of a dichotomous treatment Z_i, such as whether or not an individual received a vaccine, in epidemiology applications, or a coupon, in marketing applications. We say that node i or, equivalently, experimental unit i, is *assigned to treatment* if $Z_i = 1$, and that it is *assigned to control* (or to an alternative treatment) if $Z_i = 0$. We say that a node i is *exposed to neighborhood treatment* $G_i = g$ if $g_i(\mathbf{Z}_{\mathcal{N}_i}) = g$. The distinction between the two expressions referring to either assignment or exposure to a treatment is due to the different nature of the two treatments. In fact, while the individual treatment Z_i can be assigned or self-selected depending on individual and neighborhood characteristics, the neighborhood treatment G_i is the result of a mapping function that operates on the treatments received by the neighbors.

The assignment mechanism is then the probability distribution of the joint treatment in the whole sample, given all covariates and potential outcomes. Formally, the assignment mechanism can be written as follows: $P(\mathbf{Z}, \mathbf{G} | \mathbf{X}, \{\mathbf{Y}(z, g), z = 0, 1; g \in \mathcal{G}\})$, where $\mathbf{Z}$ is the vector of the individual treatments received by all the units in the sample, $\mathbf{G}$ is the the vector of the neighborhood treatments to which units are exposed, $\mathbf{X}$ is the covariate matrix collecting all the vectors $\mathbf{X}_i$ in the sample, and $\mathbf{Y}(z, g)$ is the collection of the potential outcomes $Y_i(z, g)$, under treatments z and g, for all units. Nevertheless, the vector of neighborhood treatments $\mathbf{G}$ is a function of $\mathbf{Z}$, where the deterministic function that links the two treatment vectors depends on the mechanism of interference, which determines the function $g_i(\cdot)$, and on the structure of the social network, which determines the neighborhood of each unit and thus the subvector $\mathbf{Z}_{\mathcal{N}_i}$ to which the function $g_i(\cdot)$ is applied. As a consequence, given the social network and the function $g_i(\cdot)$ for all units, the domain of the joint treatment $(\mathbf{Z}, \mathbf{G})$ has cardinality 2^N and it is clearly a subset of $\{0, 1\}^N \times \mathcal{G}$. The assignment mechanism then reduces to

$$P(\mathbf{Z}, \mathbf{G} | \mathbf{X}, \{\mathbf{Y}(z, g), z = 0, 1; g \in \mathcal{G}\}) = \begin{cases} P(\mathbf{Z} | \mathbf{X}, \{\mathbf{Y}(z, g), z = 0, 1; g \in \mathcal{G}\}) & \text{if } \mathbf{G} = \mathbf{g}(\mathbf{Z}) \\ 0 & \text{otherwise} \end{cases}$$

(1)

where $\mathbf{g}(\mathbf{Z})$ is the N-vector $[g_1(\mathbf{Z}_{\mathcal{N}_1}), \ldots, g_N(\mathbf{Z}_{\mathcal{N}_N})]$. Expression (1) for the assignment mech-

anism reflects the fact that the only intervention we can conceive is on the individual treatments $\mathbf{Z}$, whereas the neighborhood treatments $\mathbf{G}$ directly follow from that same intervention.

2.4 Causal estimands: Main effects and spillover effects

Under SUTNVA (Assumption 1 and Assumption 2), potential outcomes can be indexed only by the the individual treatment and the neighborhood treatment, reducing to $Y_i(Z_i, G_i)$. The potential outcome $Y_i(z, g)$ – which is the simplified expression for $Y_i(Z_i = z, G_i = g)$ – represents the potential outcome of unit i under individual treatment $Z_i = z$ and if a summary of the treatment vector of his neighborhood, $\mathbf{Z}_{\mathcal{N}_i}$, through the function $g_i(\cdot)$, had value g. A potential outcome $Y_i(z, g)$ is defined only for a subset of nodes where G_i can take on value g. We denote this subset by $V_g = \{i : g \in \mathcal{G}_i\}$, with cardinality v_g. For instance, in the case where G_i is the number of treated neighbors, V_g is the set of nodes with degree $N_i \geq g$, that is, with at least g neighbors. It is worth noting that each unit can belong to different subsets V_g, depending on the cardinality of $\mathcal{G}_i$.

Additional assumptions are needed for the definition of potential outcomes for units with degree N_i equal to zero In fact, for these kind of units the neighborhood treatment is in principle not defined and therefore they do not belong to any subset V_g. We denote by $V_\emptyset$ the subsets of units without neighbors, i.e., $V_\emptyset = \{i : N_i = 0\}$. For these units we could define potential outcomes of the form $Y_i(z)$ and analyze them separetely. Otherwise, we could make the assumption that units with no neighbors would exhibit outcomes as if they did have neighbors and their neighborhood treatment were zero, i.e., $Y_i = Y_i(Z_i, 0)$ if $N_i = 0$. In this case $V_\emptyset \subset V_0$.

We take a perspective where the potential outcomes of the population G of cardinality N are fixed quantities and expectations are simple averages of these outcomes. This perspective is sometimes referred to as 'super-population' perspective (Imbens & Rubin, 2015; Hernán & Robins, 2018).

With regard to the individual treatment, we first define the causal effect for a fixed level of the neighborhood treatment. Formally, we define the (individual) *treatment effect*, also called *main effect*, by

$$\tau(g) = E\left[Y_i(Z_i = 1, G_i = g) - Y_i(Z_i = 0, G_i = g) | i \in V_g \right] \tag{2}$$

that is, $\tau(g)$, with $g \in \mathcal{G}$, denotes the causal effect of the individual treatment when the

neighborhood treatment is set to level g. Next, define the *overall main effect τ* by the average effect of the individual treatment over the probability distribution of the neighborhood treatment, that is

$$\tau = \sum_{g \in \mathcal{G}} \tau(g) P(G_i = g) \tag{3}$$

where $\mathcal{G} = \bigcup_i \mathcal{G}_i$ [1]. We now define the causal effects of the neighborhood treatment, often referred to as spillover effects or peer effects. We define the *spillover effect* of having the neighborhood treatment set to level g versus 0, when the unit is under the individual treatment z, by

$$\delta(g;z) = E\left[\, Y_i(Z_i = z, G_i = g) - Y_i(Z_i = z, G_i = 0) | \, i \in V_g \,\right] \tag{4}$$

Finally, define the *overall spillover effect $\Delta(z)$* by the average of the spillover effects $\delta(g;z)$ over the distribution of the neighborhood treatment, that is

$$\Delta(z) = \sum_{g \in \mathcal{G}} \delta(g;z) P(G_i = g) \tag{5}$$

The main effects $\tau(g)$ in (2) and spillover effects $\delta(z;g)$ in (4) are average comparisons of potential outcomes under fixed values of the individual and neighborhood treatment. Differently, in the overall main and spillover effects in (3) and (5), the individual treatment is kept fixed while the neighborhood treatment is drawn from its observed distribution. The latter estimands are similar to the ones introduced in the few papers on causal inference in observational network data. However, previous work takes the average over hypothetical interventions, (e.g. general stochastic interventions in Van der Laan (2014) and its extensions or Bernoulli trial in Liu et al. (2016)) On the contrary, our estimands in (3) and (5) replace the hypothetical intervention on the whole treatment vector with a hypothetical intervention that fixes the treatment of a unit i and leaves the treatment of its neighbors to the observed value. Thanks to this definition, we can disentangle the total effect that we could observe on units that receive the treatment and are also exposed to the treatment. If we define *total effect* as the following average comparison

$$TE = \sum_{g \in \mathcal{G}} E\left[\, Y_i(Z_i = 1, G_i = g) - Y_i(Z_i = 0, G_i = 0) | \, i \in V_g \,\right] P(G_i = g) \tag{6}$$

---

[1]Under the 'super-population perspective' the expected value operator $E[\cdot | i \in U]$, where U is a subset of the super-population, must be understood as $\frac{1}{|U|} \sum_{i \in U} (\cdot)$. Similarly the $P(\cdot)$ operator equals $\frac{1}{N} \sum_{i \in \mathcal{N}} I(\cdot)$ (Imbens & Rubin, 2015; Hernán & Robins, 2018).

it is straightforward to show that this is equal to the sum of the overall main and spillover effects:

$$
\begin{aligned}
&= \sum_{g \in \mathcal{G}} E\left[\, Y_i(Z_i = 1, G_i = g) - Y_i(Z_i = 0, G_i = g) \middle| i \in V_g \,\right] P(G_i = g) \\
&+ \sum_{g \in \mathcal{G}} E\left[\, Y_i(Z_i = 0, G_i = g) - Y_i(Z_i = 0, G_i = 0) \middle| i \in V_g \,\right] P(G_i = g) \\
&= \tau + \Delta(0)
\end{aligned}
$$

Both the main and spillover effects $\tau(g)$ and $\delta(z; g)$ are based on the comparison between the marginal mean of two different potential outcomes. Formally, we denote with $\mu(z, g)$ the following quantity:

$$
\mu(z, g) = E\big[Y_i(z, g) \big| i \in V_g\big] \qquad \forall z \in \{0, 1\}, g \in \mathcal{G}
$$

that is, the marginal mean of the potential outcome $Y_i(z, g)$ in the subset V_g of units where this potential outcome is well-defined. We can view $\mu(z, g)$ as an average dose-response function (ADRF) depending on the dose of two different treatments, i.e., the individual treatment, which is binary, and the neighborhood treatment, which is a discrete variable.

2.5 Unconfoundedness of the joint treatment

Because the causal effects of interest depend on the comparison between two quantities $\mu(z, g)$ with different values of the joint treatment, identification results can focus on the identification of the ADRF $\mu(z, g)$. In the presence of interference, the typical unconfoundedness assumption for identification of causal effects under SUTVA must be restated. In particular, under SUTVNA the unconfoundedness assumption must be defined for both the individual and neighborhood treatment.

Assumption 3 (Unconfoundedness of Individual and Neighborhood Treatment).

$$
Y_i(z, g) \perp\!\!\!\perp Z_i, G_i \mid \boldsymbol{X}_i \qquad \forall z \in \{0, 1\}, g \in \mathcal{G}_i, \forall i.
$$

This assumption states that the individual and neighborhood treatments are independent of the potential outcomes of unit i, conditional on the vector of covariates $\boldsymbol{X}_i$.

Theorem 1 (Identification of ADRF). *Under Assumption 1 (No Multiple Versions of Treatment), Assumption 2 (Neighborhood Interference) and Assumption 3 (Unconfoundedness),*

we have

$$E\big[Y_i(z,g)|\, i \in V_g\big] = \sum_{\boldsymbol{x}\in\mathcal{X}} E\,\big[\,Y_i|Z_i = z, G_i = g, \boldsymbol{X}_i = \boldsymbol{x}, i \in V_g\big] P(\boldsymbol{X}_i = \boldsymbol{x}\,|\, i \in V_g) \quad (7)$$

Proof. See Appendix D.

For convenience we denote the conditional outcome mean in the second term of Equation (7) by $\overline{Y}_{z,g}^{obs}$. Theorem 1 implies that the average potential outcome $Y_i(z,g)$ among the subset of units V_g is equal to the weighted average of the observed outcomes of units with $Z_i = z$ and $G_i = g$ and with the same values of covariates, i.e. $\overline{Y}_{z,g}^{obs}$. If the population at hand $\mathcal{N}$ is actually the population of interest we can easily compute the latter average. Therefore, we can obtain an unbiased estimate of the dose-response function $\mu(z,g)$, for all values of z and g, and consequently all the causal effects of interest, namely the main effects and the spillover effects. In case we only have a sample, of size $M \ll N$, of the population of interest $\mathcal{N}$, $\overline{Y}_{z,g}^{obs}$ must be estimated form the sample. In this case, Theorem 1 implies that an unbiased estimator of $\mu(z,g)$ can be obtained by an unbiased estimator of the conditional outcome mean $\overline{Y}_{z,g}^{obs}$.

Assumption 3 does not specify the entire assignment mechanism $p(\mathbf{Z}, \mathbf{G}|\{\mathbf{Y}(z,g), z \in 0, 1, g \in \mathcal{G}\}, \mathbf{X})$, and thus has no implications on the independence between the treatment assignments or between the potential outcomes of different units. Therefore, Assumption 3 may hold irrespective of the independence between the treatment assignments or between the potential outcomes of different units.

In this regard, it is worth discussing the plausibility of Assumption 3 in network settings, because certain dependences may cause the assumption to fail. Assuming the unconfoundedness of the joint treatment means assuming that the vector $\mathbf{X}_i$ contains all the potential confounders of the relationship between the joint treatment and the potential outcomes for a specific unit i. The plausibility of this assumption depends on how the vector $\mathbf{X}_i$ is defined in relation to the assignment mechanism.

Assumption 3 rules out the presence of latent variables (not included in $\mathbf{X}_i$) that affect both the individual treatment and/or the neighborhood treatment and the potential outcome of a specific unit i. This has several implications. For one, in principle the assumption does not rule out the presence of homophily, that is tendency of individuals who share similar characteristics to form ties. In fact, homophily does not violate the unconfoudedness assumption in the cases where characteristics driving the homophily mechanism i) are included in $\mathbf{X}_i$, ii) even if unobserved they do not affect the outcomes iii) they correspond to

13

treament, that is people who share the same treatment/exposure variable tend to form ties. The only situation where homophily is a threat to identification is when variables underlying the network formation process are not included in $\mathbf{X}_i$ and affect the outcome.

In addition Assumption 3 has implications on the correlation of units' potential outcomes. Consider the case where $Y_i(z,g)$ are correlated with the vector of potential outcomes in his neighborhood, i.e., $\mathbf{Y}_{\mathcal{N}_i}(z,g)$, and depend on the vector of covariates $\mathbf{X}_i$: this does not pose in general identification threats, unless $\mathbf{Y}_{\mathcal{N}_i}(z,g)$ is associated with G_i after conditioning on $\mathbf{X}_i$. For example, second wave covariates, may be associated with both G_i and $Y_{\mathcal{N}_i}(z,g)$ and therefore their omission may induce association between G_i and $Y_i(z,g)$ trough the dependence between $Y_i(z,g)$ and $Y_{\mathcal{N}_i}(z,g)$. In this case, Assumption 3 would be violated. The dependence of G_i from second wave covariates may however be diluted in the application of two summarizing function, i.e. $h_i(\cdot)$ and $g_i(\cdot)$, as well as in the link function that relates the probability of the individual treatment and the neighborhood covariates. Therefore, in many circumstances, we believe that, depending on the choice of $\mathbf{X}_i$, Assumption 3 may plausibly hold, at least approximately, with dependent outcomes. If we instead consider the case where potential outcomes of neighboring units are independent, the dependence of $\mathbf{Y}_{\mathcal{N}_i}(z,g)$ with second wave covariates do not obviously pose identification issues.

2.6   Conditional main and spillover effects

In Section 2.4 and 2.5 we have defined marginal causal effects and provided identification results for the marginal mean of potential outcomes, the ADRF. Given the bivariate and multi-valued nature of the joint treatment, the estimation of these marginal quantities poses some challenges. In fact, for most of the units in the sample we do not observe neither of the two potential outcomes involved in the effect of interest. We will see in Section 5 one possible estimation approach. Nevertheless, we could focus on a specific set of units who exhibits specific values of either the individual or the neighborhood treatment and for whom the observed outcome corresponds to one of the two potential outcomes of interest. For instance, for the estimation of the effect of the individual treatment we might focus on units for whom the neighborhood treatment is zero, i.e., $G_i = 0$. Causal effects defined on such a subpopulation of units require weaker identifying conditions and can be estimated using methods for binary treatments. Definitions of these conditional causal effects, together with identification results and possible estimation methods, are reported in Appendix A.

14

3 Bias when SUTVA is wrongly assumed

3.1 Naive Estimator

Under SUTVA, potential outcomes can be indexed only by the individual treatment, i.e., $Y_i(z)$, regardless of the treatment received by other units. Therefore, in this case, the average (individual) treatment effect can be defined as

$$\tau_{sutva} = E\left[\,Y_i(Z_i = 1) - Y_i(Z_i = 0)\,\right] \tag{8}$$

In observational studies, several covariate-adjusted estimators for this quantity have been proposed. All these estimators are designed to consistently estimate the quantity

$$\tau_{X^\star}^{obs} = \sum_{\mathbf{x} \in \mathcal{X}^\star} E\left[\,Y_i | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}\,\right] - E[Y_i | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}\,]\, P(\mathbf{X}_i^\star = \mathbf{x}) \tag{9}$$

where $\mathbf{X}_i^\star \in \mathcal{X}^\star$ is the subset of covariates used for the adjustment methods. If interference is ruled out, we can imagine that we would not take neighborhood covariates into account and we would only adjust for individual covariates, eventually including contextual covariates, i.e., $\mathbf{X}_i^\star = \mathbf{X}_i^{ind}$. Thus, under SUTVA, if individual covariates form a sufficient set for unconfoundedness, that is, $Y_i(z) \perp\!\!\!\perp Z_i | \mathbf{X}^\star \; \forall z = 0, 1$, then $\tau_{sutva} = \tau_{X^\star}^{obs}$, and hence an unbiased estimator of $\tau_{X^\star}^{obs}$ would provide an unbiased estimate of the average treatment effect τ_{sutva} (e.g., Rosenbaum, 2010; Imbens & Rubin, 2015).

However, in the presence of interference, potential outcomes of the form $Y_i(z)$ are not well-defined and, thus, these estimators would clearly not estimate the quantity τ_{sutva}. Moreover, in general they would not even estimate main effects $\tau(g)$ or τ, given that estimators of τ_x^{obs} compare units belonging to the two treatment arms defined by the individual treatment Z_i, regardless of the neighborhood treatment G_i.

3.2 Outline of Bias Results

Here, we derive results for the bias for the overall main effect τ of a naïve approach that neglects interference. Bias is expressed as the difference between τ and the quantity τ_x^{obs}. When τ_x^{obs} is estimated from a sample, then the difference between τ and the quantity τ_x^{obs} represents the bias for τ of an unbiased estimator of τ_x^{obs}, that we would naïvely use if SUTVA is assumed,

We must distinguish between two cases depending on whether or not Assumption 3 holds conditioning only on the subset of covariates $\mathbf{X}^\star$, that is, $Y_i(z, g) \perp\!\!\!\perp Z_i, G_i \mid \mathbf{X}_i^\star, \forall z \in \{0, 1\}, g \in \mathcal{G}_i$.

Section 3.3 is concerned with bias results under the unconfoudedness assumption:

- Theorem 2.A provides an expression for the the quantity τ_x^{obs} in terms of potential outcomes of the form $Y_i(z, g)$

- Corollary 2.A.1 shows that if Z_i and G_i are conditionally independent an unbiased estimator of $\tau_{X^\star}^{obs}$ is unbiased for the overall main effect τ

- Corollary 2.A.2 provides expressions for the different between τ_x^{obs} and τ when Z_i and G_i are conditionally dependent and highlights the two main sources of bias

Section 3.4 is concerned with bias results when the unconfoudedness assumption does not hold conditional on a set of covariates $\mathbf{X}^\star$:

- Theorem 2.B provides an expression for the the difference between τ_x^{obs} and τ, combining the bias due to interference and the bias due to unmeasured confounders

- Corollary 2.B.1 simplifies the expression of the bias when Z_i and G_i are conditionally independent

- Corollary 2.B.2 shows that when SUTVA does hold the difference between τ_x^{obs} and τ is only due to unmeasured confounders and is the same as in the previous case where SUTVA does not hold but Z_i and G_i are conditionally independent

3.3   Bias of Naive Estimator When Unconfoudedness Holds

Theorem 2.A. *Let $G = (\mathcal{N}, \mathbb{E})$ be a known social network and let $\mathcal{N}_i$ be the neighborhood of unit i as defined by the presence of edges. Let $Z_i \in \{0, 1\}$ be a binary treatment assigned to unit i and let G_i be a deterministic function of the subset of the treatment vector $\mathbf{Z}$ in the neighborhood $\mathcal{N}_i$, that is, $G_i = g_i(\mathbf{Z}_{\mathcal{N}_i})$, with $g_i : \{0, 1\}^{N_i} \to \mathcal{G}_i$. If*

1. *Assumption 1 holds*

2. *Assumption 2 holds, given function $g_i(\cdot)$ for each unit $i \in \mathcal{N}$*

3. *Assumption 3 holds conditional on $\boldsymbol{X}_i^\star$, i.e., $Y_i(z,g) \perp\!\!\!\perp Z_i, G_i | \boldsymbol{X}_i^\star, \forall z \in \{0,1\}, g \in \mathcal{G}_i$*

then the following equality holds

$$
\begin{aligned}
\tau_{X^\star}^{obs} = \sum_{\boldsymbol{x} \in \mathcal{X}^\star} \bigg(& \sum_{g \in \mathcal{G}} E[Y_i(1,g)|\boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] P(G_i = g | Z_i = 1, \boldsymbol{X}_i^\star = \boldsymbol{x}) \\
& - E[Y_i(0,g)|\boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] P(G_i = g | Z_i = 0, \boldsymbol{X}_i^\star = \boldsymbol{x}) \bigg) P(\boldsymbol{X}_i^\star = \boldsymbol{x})
\end{aligned}
\tag{10}
$$

Proof. See Appendix D.

Corollary 2.A.1. *Under the three conditions of Theorem 2.A and the additional condition*

4. *Z_i and G_i are independent conditional on $\boldsymbol{X}_i^\star$, i.e., $Z_i \perp\!\!\!\perp G_i | \boldsymbol{X}_i^\star$*

the following equality holds:

$$
\tau_{X^\star}^{obs} = \tau
$$

Therefore, an unbiased estimator of $\tau_{X^\star}^{obs}$ is unbiased for the overall main effect τ, even in the presence of interference, if Z_i and G_i are conditionally independent.

Proof. See Appendix D.

Corollary 2.A.2. *Under the three conditions of Theorem 2.A and if $Z_i \not\perp\!\!\!\perp G_i | \boldsymbol{X}_i^\star$, an unbiased estimator of $\tau_{X^\star}^{obs}$ would be biased for the overall main effect τ, with bias given by*

$$
\begin{aligned}
\tau_{X^\star}^{obs} - \tau = \sum_{\boldsymbol{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \bigg(& E[Y_i|Z_i = 1, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] - E[Y_i|Z_i = 1, G_i = g', \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \bigg) \\
& \bigg(P(G_i = g|Z_i = 1, \boldsymbol{X}_i^\star = \boldsymbol{x}) - P(G_i = g|\boldsymbol{X}_i^\star = \boldsymbol{x}) \bigg) P(\boldsymbol{X}_i^\star = \boldsymbol{x}) \\
- \sum_{\boldsymbol{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \bigg(& E[Y_i|Z_i = 0, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] - E[Y_i|Z_i = 0, G_i = g', \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \bigg) \\
& \bigg(P(G_i = g|Z_i = 0, \boldsymbol{X}_i^\star = \boldsymbol{x}) - P(G_i = g|\boldsymbol{X}_i^\star = \boldsymbol{x}) \bigg) P(\boldsymbol{X}_i^\star = \boldsymbol{x})
\end{aligned}
\tag{11}
$$

If the spillover effect of the neighborhood treatment G_i at level g vs level g' does not depend

17

on the individual treatment Z_i, the bias formula is reduced to

$$= \sum_{\boldsymbol{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \left(E[Y_i | Z_i = z, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] - E[Y_i | Z_i = z, G_i = g', \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \right)$$

$$\left(P(G_i = g | Z_i = 1, \boldsymbol{X}_i^\star = \boldsymbol{x}) - P(G_i = g | Z_i = 0, \boldsymbol{X}_i^\star = \boldsymbol{x}) \right) P(\boldsymbol{X}_i^\star = \boldsymbol{x})$$

$$(12)$$

irrespective of the value of $z \in \{0, 1\}$.

Proof. See Appendix D.

Theorem 2.A concerns the bias of the estimation approach when we wrongly assume away the presence of interference, but we are able to adjust for a set of covariates $\mathbf{X}^\star$ that satisfy the uncondoudedness assumption (Assumption 3). The core result of the theorem is the expression for $\tau_{X^\star}^{obs}$ in (10), which shows that $\tau_{X^\star}^{obs}$ is actually comparing different types of potential outcomes with different values of the neighborhood treatment. Then, Corollary 2.A.1 states that if the individual and neighborhood treatments are independent conditional on $\mathbf{X}_i^\star$, then using covariates-adjusted estimation methods that assume SUTVA would yield unbiased estimates for the overall main effect τ, even if SUTVA does not hold. On the contrary, as stated in Corollary 2.A.2, a residual correlation between Z_i and G_i, after conditioning on $\mathbf{X}_i^\star$, would result in bias. The formula presented in (12) shows that the bias depends on two factors: the level of interference and the residual association between the individual treatment Z_i and the neighborhood treatment G_i after conditioning for $\mathbf{X}_i^\star$. There can be several reasons for such an association. For instance, the individual treatment Z_i and the neighborhood treatment G_i can be linked through neighborhood covariates that are not included in $\mathbf{X}_i^\star$. Moreover, there can be peer influence in the treatment uptake. Such situation is plausible in most realistic applications where a unit's choice to take the treatment might depend also on other units's choices. Another cause of an association between Z_i and G_i can be the presence of homophily, that is, similar characteristics underlying the neighborhood structure and driving the assignment mechanism.

3.4 Bias of Naive Estimator When Unconfoudedness Does Not hold

Theorem 2.B. *Under Assumptions 1 and 2 , if Assumption 3 does not hold conditional on $\boldsymbol{X}_i^\star$, i.e., $Y_i(z, g) \not\perp\!\!\!\perp Z_i, G_i | \boldsymbol{X}_i^\star$, but holds conditional on $\boldsymbol{X}_i^\star$ and an additional vector of covariates $\mathbf{U}_i \in \mathcal{U}$, i.e., $Y_i(z, g) \perp\!\!\!\perp Z_i, G_i | \boldsymbol{X}_i^\star, \mathbf{U}_i$, an unbiased estimator of $\tau_{X^\star}^{obs}$ would be*

biased for the overall main effect τ, with bias $\tau_{X^\star}^{obs} - \tau$ given by

$$
= \sum_{x \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \Bigg(E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = 1, G_i = g', \mathbf{U}_i = \mathbf{u}', \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \Bigg)
$$

$$
\Bigg(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}) P(G_i = g | Z_i = 1, \boldsymbol{X}_i^\star = \boldsymbol{x})
$$

$$
- P(\mathbf{U}_i = \mathbf{u} | G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}) P(G_i = g | \boldsymbol{X}_i^\star = \boldsymbol{x}) \Bigg) P(\boldsymbol{X}_i^\star = \boldsymbol{x})
$$

$$
- \sum_{x \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \Bigg(E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \tag{13}
$$

$$
- E[Y_i | Z_i = 0, G_i = g', \mathbf{U}_i = \mathbf{u}', \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \Bigg)
$$

$$
\Bigg(P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}) P(G_i = g | Z_i = 0, \boldsymbol{X}_i^\star = \boldsymbol{x})
$$

$$
- P(\mathbf{U}_i = \mathbf{u} | G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}) P(G_i = g | \boldsymbol{X}_i^\star = \boldsymbol{x}) \Bigg) P(\boldsymbol{X}_i^\star = \boldsymbol{x})
$$

If both the spillover effect of the neighborhood treatment G_i at level g vs level g' and the effect of the unmeasured confounder $\mathbf{U}_i$ at level $\mathbf{u}$ vs level $\mathbf{u}'$ do not depend on the individual treatment Z_i, the bias formula reduces to

$$
= \sum_{x \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \Bigg(E[Y_i | Z_i = z, G_i = g, \mathbf{U}_i = \mathbf{u}, \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = z, G_i = g', \mathbf{U}_i = \mathbf{u}', \boldsymbol{X}_i^\star = \boldsymbol{x}, i \in V_g] \Bigg)
$$

$$
\Bigg(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}) P(G_i = g | Z_i = 1, \boldsymbol{X}_i^\star = \boldsymbol{x})
$$

$$
- P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \boldsymbol{X}_i^\star = \boldsymbol{x}) P(G_i = g | Z_i = 0, \boldsymbol{X}_i^\star = \boldsymbol{x}) \Bigg) P(\boldsymbol{X}_i = \boldsymbol{x})
$$

$$
\tag{14}
$$

irrespective of the value of $z \in \{0, 1\}$.

Proof. See Appendix D.

Corollary 2.B.1. *Under the following conditions*

1. *Assumption 1 holds*

2. *Assumption 2 holds given the function $g_i(\cdot)$ for each unit $i \in \mathcal{N}$*

3. *Assumption 3 holds conditional on $\boldsymbol{X}_i^\star$ and $\mathbf{U}_i$, i.e., $Y_i(z, g) \perp\!\!\!\perp Z_i, G_i | \boldsymbol{X}_i^\star, \mathbf{U}_i, \forall z \in \{0, 1\}, g \in \mathcal{G}_i$*

4. *Z_i and G_i are independent conditional on $\boldsymbol{X}_i^\star$, i.e., $Z_i \perp\!\!\!\perp G_i | \boldsymbol{X}_i^\star$*

an unbiased estimator of $\tau_{X^\star}^{obs}$ would be biased for the overall main effect τ, with bias $\tau_{X^\star}^{obs} - \tau$ given only by the unmeasured confounder $\mathbf{U}_i$:

$$
= \sum_{\boldsymbol{x} \in \mathcal{X}^\star} \sum_{u \in \mathcal{U}} \left(E[Y_i | Z_i = z, \mathbf{U}_i = \mathbf{u}, \boldsymbol{X}_i^\star = \boldsymbol{x}] - E[Y_i | Z_i = z, \mathbf{U}_i = \mathbf{u}', \boldsymbol{X}_i^\star = \boldsymbol{x}] \right)
$$
$$
\left(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, \boldsymbol{X}_i^\star = \boldsymbol{x}) - P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, \boldsymbol{X}_i^\star = \boldsymbol{x}) \right) P(\boldsymbol{X}_i = \boldsymbol{x})
$$
(15)

where $z \in \{0, 1\}$

Proof. See Appendix D.

Corollary 2.B.2. *Under the following conditions*

1. *SUTVA holds*

2. *Assumption 3 holds conditional on $\boldsymbol{X}_i^\star$ and $\mathbf{U}_i$, i.e., $Y_i(z, g) \perp\!\!\!\perp Z_i, G_i | \boldsymbol{X}_i^\star, \mathbf{U}_i, \forall z \in \{0, 1\}, g \in \mathcal{G}_i$*

an unbiased estimator of $\tau_{X^\star}^{obs}$ would be biased for the overall main effect τ, an unbiased estimator of $\tau_{X^\star}^{obs}$ would be biased for the overall main effect τ, with bias $\tau_{X^\star}^{obs} - \tau$ given only by the unmeasured confounder $\mathbf{U}_i$ as in Equation (15).

Proof. See Appendix D.

Theorem 2.B states that, if in the estimation of the treatment effect we wrongly assume SUTVA and adjust for a set of covariates $\mathbf{X}_i^\star$ that does not suffice for unconfoundedness to hold, the bias due to interference is combined with the bias due to unmeasured confounders $\mathbf{U}_i = \mathbf{X}_i \backslash \mathbf{X}_i^\star$. However, if either the individual treatment Z_i and the neighborhood treatment G_i are independent given $\mathbf{X}_i^\star$ or there is no interference between units, then covariate-adjusted estimators would be biased only because of unmeasured confounders. The vector of unmeasured confounders $\mathbf{U} : i$ can include neighborhood covariates $\mathbf{X}_i^{neigh}$ that might affect the individual treatment Z_i directly or through the neighborhood treatment G_i. Typically, when SUTVA is assumed these kind of covariates are not taken into account in the estimation procedure. Theorem 2.B shows that we should pay careful attention to the problem of dependence between units, even if proper interference can be ruled out.

4 Definition and Properties of Generalized Propensity Score Under Neighborhood Interference

In this section we contribute to the literature by extending the definition of propensity score under neighborhood interference. Under Assumption 2, in fact, each unit is exposed to a bivariate treatment; therefore the definition of the propensity score must be generalized to be the joint probability of an individual with certain observed characteristics of being assigned to an individual treatment and being exposed to a neighborhood treatment. We show that the new propensity score has balancing properties that are similar to the propensity score under SUTVA.

4.1 Joint Propensity Scores

We define as the *joint propensity score*, denoted by $\psi(z; g; x)$, as the joint probability distribution of the individual treatment and the neighborhood treatment given the observed covariates:

$$\psi(z; g; x) = P(Z_i = z, G_i = g | \mathbf{X}_i = \mathbf{x}) \tag{16}$$

$\psi(z; g; x)$ is the probability for unit i of being exposed to treatment z and neighborhood treatment g, given his observed individual and neighborhood characteristics $\mathbf{x}$.

The joint propensity score does not necessarily correspond to the unit-level assignment probability, which is in general expressed as $Pr(Z_i = z, G_i = g | \mathbf{X}, \{\mathbf{Y}(z, g), z = 0, 1; g \in \mathcal{G}\})$ (unit-level version of assignment mechanism in (1)). However, if the unit-level assignment probability of being exposed to treatment z and neighborhood treatment g only depends unit-level variables, i.e., $Pr(Z_i = z, G_i = g | \mathbf{X}, \{\mathbf{Y}(z, g), z = 0, 1; g \in \mathcal{G}\}) = Pr(Z_i = z, G_i = g | \mathbf{X}_i, \{Y_i(z, g), z = 0, 1; g \in \mathcal{G}_i\})$, [2] and unconfoundedness holds given $\mathbf{X}_i$ (Assumption 3), then the unit-level assignment probability coincides with the joint propensity score.

Given the definition of the joint propensity score in (16), we can prove the following two properties.

Proposition 3 (Balancing Property). *The joint propensity score is a balancing score, that is,*

$$P(Z_i = z, G_i = g | \boldsymbol{X}_i, \psi(z; g; \boldsymbol{X}_i)) = P(Z_i = z, G_i = g | \psi(z; g; \boldsymbol{X}_i))$$

---

[2]This property is similar to the individualistic property of the assignment mechanism in Imbens & Rubin (2015). However, here it is defined on the extended assignment mechanism defined on both the individual and the neighborhood treatment and the vector of covariates $\mathbf{X}_i$ include neighbors' characteristics.

.

Proof. See Appendix D.

Proposition 3 implies that if a group of units have the same $\psi(z; g; \mathbf{x})$, then the distribution of covariates $\mathbf{X}$ for the group is the same between the arm with $Z_i = z$ and $G_i = g$ and all the other arms.

Proposition 4 (Conditional Unconfoundedness of Z_i and G_i given the joint propensity score). *If Assumption 3 holds given $\mathbf{X}_i$, then*

$$Y_i(z, g) \perp\!\!\!\perp Z_i, G_i | \psi(z; g; \mathbf{X}_i) \qquad \forall z \in \{0, 1\}, g \in \mathcal{G}_i$$

Proof. See Appendix D.

Proposition 4 states that if unconfoundedness holds conditional on covariates $\mathbf{X}_i$ than the distribution of the potential outcome $Y_i(z, g)$ is independent of the individual and the neighborhood treatment of units with the same value of the joint propensity score $\psi(z; g; \mathbf{X}_i)$ for the corresponding values z and g. This is a crucial result in that it allows imputing missing potential outcomes $Y_i(z, g)$ across arms for units with the same propensity of being exposed to treatment z and neighborhood treatment g. Therefore, it is sufficient to adjust the joint propensity score to account for confounding bias in the estimation of both main and spillover effects.

4.2 Individual Propensity Score and Neighborhood Propensity Score

We can consider the following factorization of the joint propensity score:

$$\begin{aligned}
\psi(z; g; x) &= P(Z_i = z, G_i = g | \mathbf{X}_i = \mathbf{x}) \\
&= P(G_i = g | Z_i = z, \mathbf{X}_i^g = \mathbf{x}^g) P(Z_i = z | \mathbf{X}_i^z = \mathbf{x}^z)
\end{aligned} \tag{17}$$

where $\mathbf{X}_i^g \in \mathcal{X}^g \subset \mathcal{X}$ is the subset of covariates affecting the neighborhood treatment, and $\mathbf{X}_i^z \in \mathcal{X}^z \subset \mathcal{X}$ is the subset of covariates affecting the individual treatment. In principle vectors $\mathbf{X}_i^z$ and $\mathbf{X}_i^g$ could be different. In fact, individual characteristics collected in $\mathbf{X}_i^{ind}$ should be included in $\mathbf{X}_i^z$, but the type of neighboring units and the neighborhood structure, i.e., $\mathbf{X}_i^{neigh}$, could also affect the probability of individual treatment for unit i. Also, the probability of a unit's neighbors receiving treatments $\mathbf{Z}_{\mathcal{N}_i}$ summarized in g will depend on

neighborhood characteristics in $\mathbf{X}_i^{neigh}$, on neighborhood structure that is likely to affect the mapping of $\mathbf{Z}_{\mathcal{N}_i}$ into g, but might as well depend on individual characteristics of unit i in $\mathbf{X}_i^{ind}$. We denote the probability of having the neighborhood treatment at level g conditional on a specific value z of the individual treatment and on the vector of covariates $\mathbf{X}_i^g$, i.e., $P(G_i = g | Z_i = z, \mathbf{X}_i^g = \mathbf{x}^g)$, with $\lambda(g; z; x^g)$ and refer to it as *neighborhood propensity score*. Similarly, we denote the probability of having the individual treatment at level z conditional on covariates $\mathbf{X}_i^z$, i.e., $P(Z_i = z | \mathbf{X}_i^z = \mathbf{x}^z)$, with $\phi(z; x^z)$ and we refer to it as the *individual propensity score*. Given this factorization (Eq. 17), the unconfoundedness assumption holds conditioning on the two types of propensity scores separately.

Proposition 5 (Conditional Unconfoundedness of Z_i and G_i given the individual propensity score and the neighborhood propensity score)**.** *If Assumption 3 holds given $\boldsymbol{X}_i$, then*

$$ Y_i(z, g) \perp\!\!\!\perp Z_i, G_i | \lambda(g; z; \boldsymbol{X}_i^g), \phi(1; \boldsymbol{X}_i^z) \qquad \forall z \in \{0, 1\}, g \in \mathcal{G}_i $$

Proof. See Appendix D.

As Proposition 4, Proposition 5 is a key result in that it allows deriving adjustment methods that separately adjust for the individual and the neighborhood treatment.

5 Propensity Score-Based Estimator for Main Effects and Spillover Effects

Here we propose an estimator that relies on results of the previous section and follows directly from the formalization of the individual and neighborhood treatment. Consider a sample of the population G. The unbiasedness of the proposed estimator relies on a random sampling mechanism that preserves the connections among units. Different sampling schemes are possible. Here we consider two sampling schemes. A possible method is cluster sampling where disjoint clusters are randomly sampled and but networks in each cluster are not necessarily fully connected. School sampling with friendship networks within school is a good example. Another option is an egocentric sampling method where randomly selected units, called 'ego', are the units of analysis, but, via interview, they are asked to nominate a list of persons ('alters') with whom they have a specific type of relationship(Kolaczyk , 2009; Perri et al., 2018)

5.1 Individual and Neighborhood Propensity Score Estimator

Based on the identification result of Theorem 1, we could obtain an unbiased estimator of $\mu(z,g)$ using an unbiased estimator of the conditional mean $\overline{Y}_{z,g}^{obs}$. For example, this quantity could be estimated by taking the mean of the observed outcomes within cells defined by covariates (stratification) (Imbens & Rubin, 2015). Nevertheless, the presence of continuous covariates or a large number of covariates poses some challenges in the estimation of $\overline{Y}_{z,g}^{obs}$. Relying on results of the previous section, we propose a propensity score-based estimator of the average dose-response function $\mu(z,g)$, that allows estimating marginal main and spillover effects $\tau(g)$ and $\delta(z;g)$.

If Assumption 3 of unconfoundedness holds given $\mathbf{X}_i$, then we can define the joint propensity score as the probability distribution of the individual and neighborhood treatments given covariates $\mathbf{X}_i$. [3] As a consequence of Proposition 4, we can get an unbiased estimator of $\mu(z,g)$ by adjusting for the joint propensity score $\psi(z;g;\mathbf{X}_i)$, that is, using an unbiased estimator of the quantity

$$E\big[E[Y_i|Z_i = z, G_i = g, \psi(z;g;\mathbf{X}_i)]|Z_i = z, G_i = g\big]$$

where the outer expectation is taken over the empirical distribution of the joint propensity score in the population. However, because the joint treatment is bivariate and the neighborhood treatment G_i can potentially take on many different values, depending on the function $g_i(\cdot)$, it is not easy to stratify or match on this joint propensity score (Imbens , 2000). To solve this issue, we propose a semi-parametric approach, which exploits the factorization of the joint propensity score in (17).

According to Proposition 5, we can adjust for the joint propensity score by adjusting separately for both the individual propensity score $\phi(1;x^z)$ and for the neighborhood propensity score $\lambda(g;z;x^g)$,that is, using an unbiased estimator of the quantity

$$E\big[E[Y_i|Z_i = z, G_i = g, \phi(1;\mathbf{X}_i^z), \lambda(g;z;\mathbf{X}_i^g)]|Z_i = z, G_i = g\big]$$

The need for propensity score-adjustment has a different meaning for the two types of

---

[3]Note that we included in the joint propensity score the set of covariates that is sufficient to satisfy the unconfoundedness assumption. Oftentimes, the probability of unit i being exposed to a certain neighborhood treatment does depend on non-neighboring characteristics, given that the probability of his neighbors being assigned to treatment might depend on their neighbors' covariates. In this case the joint propensity score will not coincide with the unit-level assignment mechanism.

propensity scores. When estimating spillover effects of the form $\delta(z, g)$, adjustment for the neighborhood propensity score $\lambda(g; z, \mathbf{X}_i^g)$ is required in order to correct for the covariate imbalance across the two arms with the values of the neighborhood treatment that we are comparing, i.e., $G_i = g$ and $G_i = 0$. Conversely, adjustment for the individual propensity score $\phi(1; \mathbf{X}_i^z)$ is required when we are estimating marginal spillover effects. In fact, the ADRF $\mu(z, g)$ for each value of $z \in \{0, 1\}$ is estimated using data of the corresponding individual treatment arm. Conditioning on $\phi(1; \mathbf{X}_i^z)$ allows generalizing the result to the opposite arm. When estimating main effects this argument is reversed.

Moreover, the two propensity scores are of different nature. The individual treatment is binary and thus we can adjust for the individual propensity score using the standard propensity score-based adjustment methods for binary treatment. On the contrary, the neighborhood treatment has a discrete domain $\mathcal{G}_i$ with cardinality depending on the function g_i. $\lambda(g; z; x^g)$ can be seen as the *generalized propensity score* (GPS) defined by Hirano & Imbens (2004) for continuous treatments. To adjust for the neighborhood propensity score, we can then use a similar model-based approach. To reduce the bias due to a possible model-misspecification, we propose the use of a subclassification on the individual propensity score $\phi(1; x^z)$ and, within subclasses that are approximately homogenous in $\phi(1; x^z)$, a model-based approach for the neighborhood propensity score .

5.2 Estimating Procedure: Subclassification and GPS

Here we describe the details of the estimating procedure (*Subclassification and GPS*). Note that the proposed estimation strategy is particularly appropriate when the domain of the neighborhood treatment $\mathcal{G}_i \subset \mathcal{R}$.

1) We derive a subclassification on the individual propensity score $\phi(1; \mathbf{X}_i^z)$ as follows:

    a) We estimate $\phi(1; \mathbf{X}_i^z)$ with a logistic regression for Z_i conditional on covariates $\mathbf{X}_i^z$;

    b) We predict $\phi(1; \mathbf{X}_i^z)$ for each unit;

    c) We identify J subclasses B_j, with $j = 1, \ldots, J$, defined by similar values of $\phi(1; \mathbf{X}_i^z)$ and where there is sufficient balance between individual treatment groups, i.e., $\mathbf{X}_i^z \perp\!\!\!\perp Z_i | i \in B_j$.

2) Within each subclass B_j, we repeat the following steps to estimate $\mu_j(z, g) = E[Y_i(z, g) | i \in B_j^g]$, where $B_j^g = V_g \cap B_j$:

a) We estimate the parameters of a model for the neighborhood propensity score $\lambda(g; z; x^g)$: $\lambda(z; g; \mathbf{X}_i^g) = Pr(G_i = g | Z_i = z, \mathbf{X}_i^g) = f^G(g, z, \mathbf{X}_i^g)$

b) We use the observed data $(Y_i, Z_i, G_i, \mathbf{X}_i^g)$ and $\widehat{\Lambda} = \lambda(G_i; Z_i; \mathbf{X}_i^g)$ to estimate the parameters of a model $Y_i(z, g) \mid \lambda(z; g; \mathbf{X}_i^g) \sim f^Y(z, g, \lambda(g; z; \mathbf{X}_i^g))$;

c) For a particular level of the joint treatment $(Z_i = z, G_i = g)$, for each unit $i \in B_j^g$ we predict the neighborhood propensity score evaluated at that level of the treatment, i.e., $\lambda(g; z; \mathbf{X}_i^g)$, and use it to predict the potential outcome $Y_i(z, g)$.

d) To estimate the dose-response function $\mu_j(z, g)$ we average the potential outcomes over $\lambda(z; g; \mathbf{X}_i^g)$

$$\widehat{\mu}_j(z, g) = \frac{\sum_{i \in B_j^g} \widehat{Y}_i(z, g)}{|B_j^g|}$$

3) We derive the average dose-response function as follows:

$$\widehat{\mu}(z, g) = \sum_{j=1}^J \widehat{\mu}_j(z, g) \pi_j^g$$

where $\pi_j^g = \frac{|B_j^g|}{v_g}$.

It is worth noting that there are three key differences between our GPS approach and the one in Hirano & Imbens (2004): i) it is performed within each subclass defined by $\phi(1; x^z)$, ii) both the GPS and the outcome model will include the individual treatment $Z_i = z$, iii) G_i is a discrete treatment whose domain $\mathcal{G}_i$ depends on the function $g_i(\cdot)$, and iv) the dose-response function $\mu(z, g)$ is not always defined for all units.

The problem of statistical inference with units connected in network is not straightforward, given the correlation structure of the data. In Appendix C we propose a bootstrapping method with resampling at different levels depending on the sampling scheme.

6 Realistic Simulation Study leveraging Add Health data

We use a simulation study to illustrate how the proposed methods may be applied. Our aim is twofold: i) to validate the analytical derivation of the bias for the main effect when interference is wrongly ruled out; ii) to show the performance of the proposed estimators in a

realistic sample. We use friendship network data collected through the National Longitudinal Study of Adolescent Health (Add Health). We limited our analysis to 29 schools for a total of 16410 students. Let us assume that we are interested in estimating the effect of health coverage on flu infection. Such effect could, in principle, be estimated from the real data. However, to show the performance of the estimation procedures in different scenarios, both the treatment and the outcome are generated using different generating processes. Let the individual treatment variable Z_i denote whether student i was covered or not by some health insurance, and let Y_i denote the number of days student i missed school because of illness in one given year. The assumption of no-interference might be questionable here, because we can think that health insurance leads to better health for the covered student, which in turn can reduce the chance of spreading infectious diseases, such as flu, to the student's friends. Therefore, even if we are only interested in estimating the main effect of the individual health insurance, not taking interference into account might result in biased estimates. For simplicity, let us consider two individual covariates: $\texttt{race}_i$, indicating student i's race (1 if white and 0 if other), and $\texttt{grade}_i$, a discrete variable indicating student i's grade. Let $\mathbf{X}_i^{ind} = (\texttt{race}_i, \texttt{grade}_i)$ and $\mathbf{X}_i^{neig} = (\frac{\sum_{k \in \mathcal{N}_i} \texttt{race}_k}{N_i}, \frac{\sum_{k \in \mathcal{N}_i} \texttt{grade}_k}{N_i}, N_i)$.

The simulation study considers four scenarios of dependence between Z_i and G_i. In all scenarios but the third, G_i is the proportion of friends with health insurance among the first five best friends. In the third scenario G_i is the number of 'treated' friends among all friends.

Scenario 1: Z_i is generated depending on individual race and grade. Hence, Z_i and G_i are independent conditional on $\mathbf{X}_i^{ind}$.

Scenario 2: Z_i is generated depending on individual race and grade, and on friends' race and grade. Hence, Z_i and G_i are dependent if we condition only on $\mathbf{X}_i^{ind}$, but independent conditional on $\mathbf{X}_i^{neig}$.

Scenario 3: Z_i is generated depending on individual race and grade, and on the student's degree. Hence, Z_i and G_i are dependent if we condition only on $\mathbf{X}_i^{ind}$, but independent conditional on $\mathbf{X}_i^{neig}$.

Scenario 4: Z_i is generated depending on individual race and grade and on G_i. Hence Z_i and G_i are directly correlated and are not independent even if we condition on $\mathbf{X}_i^{ind}$ or $\mathbf{X}_i^{neig}$. (Here data are generated using an iterative procedure).

Details of the association between Z_i, G_i and the covariates in the four scenarios are in Appendix B.1. The outcome models are described in Appendix B.2. It is worth noting that

27

Table 1: Computed Bias for τ.

Scenario	interference	bias($\emptyset$)	bias($\mathbf{X}_i^{ind}$)	bias($\mathbf{X}_i^z$)
1 $(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind})$	low	-5.977	-0.045	-0.045
	medium	-6.200	-0.072	-0.072
	high	-6.323	-0.090	-0.090
2 $(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	-5.749	-1.636	-0.034
	medium	-6.498	-2.618	-0.054
	high	-6.998	-3.273	-0.068
3 $(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind}, \texttt{degree}_i)$	low	-4.158	-1.247	-0.047
	medium	-4.792	-2.079	-0.075
	high	-5.744	-3.327	-0.095
4 $(Z_i \not\perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	-9.504	-1.414	-1.415
	medium	-11.681	-2.263	-2.263
	high	-13.132	-2.829	-2.825

the generating models are slightly different in the two sets of simulations where the focus is either on main effects or on spillover effects. In the first set of simulations, the outcome distribution, reported in Equation (28) in Appendix B.2.1, only depends on individual covariates and does not depend on neighborhood characteristics. Therefore, unconfoundedness (Assumption 3) holds conditional on $\mathbf{X}_i^{ind}$. This is to show that, even if the outcome does not depend on neighborhood covariates that do affect Z_i, a bias can still occur due to the induced correlation between G_i and Z_i and interference. In the second set of simulations with a focus on spillover effects, the outcome model, reported in Equation (31) in Appendix B.2.2, depends on both individual and neighborhood covariates, and has a more complicated structure with additional interaction terms. For each scenario, we consider different levels of interference by changing the value of the parameter δ in Equations (28) and (31).

6.1 Main Effect: Bias of naïve estimators and GPS-based estimator

The focus of this first simulation study is on the estimation of the overall main effect τ (Eq. 3). Our aim here is to show in different scenarios the bias resulting from neglecting the presence of interference and using typical estimators of the treatment effect and compare it to the analytical results presented in Section 3.

Table 1 shows the bias computed using formulas in Theorem 2. In particular, for each

scenario, we derived the bias resulting from neglecting interference and from adjusting for different sets of covariates: $\mathbf{X}_i^\star = \{\emptyset, \mathbf{X}_i^{ind}, \mathbf{X}_i^z\}$. Given that unconfoundedness holds conditional on $\mathbf{X}_i^{ind}$, for the approach that does not adjust for any covariates, i.e., $\mathbf{X}_i^\star = \emptyset$, the bias is due to both unmeasured confounders and interference. Equation (13) has been used for the computation, since in the outcome model (Equation (28) of Appendix B.2) there is an interaction between the individual treatment Z_i and $\texttt{race}_i$ (an unmeasured confounder in this case). When we do adjust for either the individual covariates $\mathbf{X}_i^{ind}$ or all covariates $\mathbf{X}_i^z$, the bias is derived using Equation (12). In fact, in these approaches the potential confounders $\mathbf{X}_i^{ind}$ are included in the adjustment set and, thus, we do not have the bias due to unmeasured confounders. However, when the the adjustment set is not sufficient to rule out the dependence between Z_i and G_i, the presence of interference does produce bias.

We then run 500 replications of each scenario, applying the following estimators (Imbens & Rubin, 2015) of the overall main effect:

- A simple difference in means estimator (Unadjusted Estimator) comparing treated and untreated units;

- An ordinary least squares estimator that regresses the outcome on individual treatment Z_i, adjusting only for individual covariates $\mathbf{X}_i^{ind}$;

- An estimator based on a subclassification on the individual propensity score $\hat{\phi}(1, \mathbf{X}_i^{ind})$, which is estimated using only individual covariates $\mathbf{X}_i^{ind}$;

- An ordinary least squares estimator that regresses the outcome on individual treatment Z_i, adjusting for individual and neighborhood covariates, i.e, $\mathbf{X}_i^{ind}$;

- An estimator based on a subclassification on the individual propensity score $\hat{\phi}(1, \mathbf{X}_i^z)$, which is estimated using individual and neighborhood covariates, i.e, $\mathbf{X}_i^z$;

- The estimator proposed in Section 5 (Sublcassification & GPS), which is based on subclassification on the individual propensity score and model-based adjustment for the neighborhood propensity score.

Table 2 reports the mean bias and root mean squared error (RMSE) of all these estimators in all scenarios. The first five estimators ignore the presence of interference, resulting in a bias that is proportional to the level of interference and the level of association between individual treatment and neighborhood treatment. In Scenario 1, where Z_i and G_i are independent conditional on individual covariates, all subclassification-based estimators

Table 2: Estimation of main effect τ.

Scenario	δ	Unadjusted		Regression $\sim Z_i, \mathbf{X}_i^{ind}$		Subclass $\hat{\phi}(1, \mathbf{X}_i^{ind})$		Regression $\sim Z_i, \mathbf{X}_i^z$		Subclass $\hat{\phi}(1; \mathbf{X}_i^z)$		Subclass $\hat{\phi}(1, \mathbf{X}_i^z)$ & GPS	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1 $(Z_i \perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind})$	low	-5.104	5.104	-3.429	3.431	-0.016	0.130	-3.254	3.256	0.006	0.074	-0.002	0.064
	medium	-5.278	5.278	-3.529	3.532	-0.021	0.149	-3.256	3.258	0.007	0.090	0.002	0.067
	high	-5.396	5.396	-3.598	3.601	-0.021	0.167	-3.258	3.261	0.015	0.108	0.003	0.067
2 $(Z_i \perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	-5.136	5.136	-3.477	3.477	-1.652	1.668	-2.156	2.157	-0.066	0.331	-0.000	0.036
	medium	-5.884	5.884	-4.521	4.521	-2.620	2.631	-2.318	2.319	-0.097	0.511	0.003	0.039
	high	-6.384	6.384	-5.222	5.223	-3.272	3.281	-2.430	2.432	-0.123	0.635	-0.002	0.036
3 $(Z_i \perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind}, \texttt{degree}_i)$	low	-4.106	4.107	-2.079	2.089	-1.146	1.149	-1.033	1.049	0.078	0.553	-0.002	0.230
	medium	-4.725	4.726	-2.886	2.893	-1.911	1.913	-1.041	1.056	0.082	0.556	0.019	0.233
	high	-5.655	5.656	-4.098	4.103	-3.058	3.059	-1.054	1.068	0.084	0.559	0.012	0.234
4 $(Z_i \not\perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	-8.670	8.670	-7.104	7.104	-1.443	1.445	-7.076	7.076	-1.738	1.740	0.000	0.105
	medium	-10.761	10.761	-8.861	8.862	-2.309	2.309	-8.824	8.823	-2.784	2.785	-0.003	0.095
	high	-12.154	12.154	-10.030	10.030	-2.883	2.884	-9.985	9.985	-3.477	3.478	0.010	0.090

are essentially unbiased, for all levels of interference. Regression-based estimators are affected by a bias due to model-misspecification, in the range of -3.50, when adjusting for individual covariates, and -3.25, when adjusting for individual and neighborhood covariates. Finally, in the unadjusted difference in means estimators we have the usual bias due to unmeasured confounders. In scenarios where the association between the Z_i and G_i is due to neighborhood covariates (Scenario 2 and Scenario 3), for the unadjusted estimator and the regression estimator that only adjust for individual covariates, an additional bias due to interference is combined with the aforementioned bias due to unmeasured confounders and model-misspecification, respectively. The unadjusted estimator and the two subclassification-based estimators, adjusting for individual covariates or all covariates, show a bias that is comparable to the corresponding analytical bias reported in Table 1. The two estimators –regression and subclassification-based – that adjust for both individual and neighborhood covariates are able to reduce the bias due to interference. The model-misspecification bias for the regression estimator is still present, whereas the semiparametric sublcassifcation-based estimator is able to cancel out the bias. In Scenario 4, where there is a direct correlation between Z_i and G_i, an adjustment for neighborhood covariates cannot remove this correlation. Therefore, all regression-based and sublcassifcation-based estimators are affected by the bias due to interference. Only an estimator that explicitly takes interference into account is able to eliminate this kind of bias. The estimator proposed in this paper (Sublcassification & GPS), which adjusts for both individual and neighborhood

Table 3: Estimation of $\Delta(0)$

Scenario	interference	Unadjusted Regression $\sim Z_i, G_i$		Subclass $\hat{\phi}(1, \mathbf{X}_i^z)$		GPS		Subclass $\hat{\phi}(1, \mathbf{X}_i^z)$ & GPS	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1 $(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind})$	low	1.576	1.581	2.871	2.901	-1.457	1.465	0.003	0.091
	medium	1.568	1.574	2.838	2.872	-1.462	1.472	0.005	0.114
	high	1.571	1.580	2.878	2.910	-1.467	1.479	0.000	0.148
2 $(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	3.503	3.506	4.132	5.259	-0.574	0.590	0.033	0.098
	medium	3.506	3.510	4.037	4.933	-0.579	0.599	0.041	0.118
	high	3.485	3.489	4.204	5.232	-0.592	0.613	0.034	0.129
3 $(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind}, \texttt{degree}_i)$	low	5.445	5.446	7.005	7.048	0.380	0.399	-0.035	0.091
	medium	5.455	5.456	7.009	7.050	0.381	0.401	-0.033	0.091
	high	5.441	5.442	7.054	7.096	0.381	0.400	-0.033	0.099
3 $(Z_i \not\perp\!\!\!\perp G_i \mid \mathbf{X}_i^z, \mathbf{X}_i^{neigh})$	low	3.002	3.002	2.577	2.584	-1.201	1.202	0.071	0.111
	medium	3.002	3.002	2.556	2.563	-1.202	1.203	0.068	0.111
	high	3.005	3.005	2.548	2.555	-1.200	1.201	0.066	0.111

propensity scores, shows no bias in all scenarios and for all levels of interference, thanks to Assumption 3 holding. In fact, relying on Theorem 1 and Proposition 5, an unbiased estimator of the quantity $E_{\Lambda,\Phi}\left[\,E[Y_i|Z_i = z, G_i = g, \phi(1; \mathbf{X}_i^z), \lambda(g; z, \mathbf{X}_i^g)]|Z_i = z, G_i = g\,\right]$ is unbiased for the ADRF $\mu(z, g)$.

6.2 Spillover Effects

The second simulation study focuses on the estimation of the marginal spillover effect $\Delta(0)$ and $\Delta(1)$. Here we consider estimators that do take interference into account and are explicitly designed for the estimation of spillover effects. In observational studies, the presence of confounders $\mathbf{X}_i$ requires the use of covariate-adjusted methods. Given unconfoundedness (Assumption 3), according to Proposition 4 and Proposition 5, we might adjust for either the joint treatment propensity score or separately for the individual propensity score and the neighborhood propensity score, as in our proposed estimator. Failing to adjust for either of the two propensity score would result in biased estimates.

We compare the performance of the following estimators, which differ for the way this covariate-adjustment is handled:

Table 4: Estimation of $\Delta(1)$

Scenario	interference	Unadjusted Regression $\sim Z_i, G_i$ Bias	RMSE	Subclass $\hat{\phi}(1, \mathbf{X}_i^z)$ Bias	RMSE	GPS Bias	RMSE	Subclass $\hat{\phi}(1, \mathbf{X}_i^z)$ & GPS Bias	RMSE
1	low	3.195	3.196	2.870	2.878	0.432	0.441	0.003	0.055
$(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind})$	medium	3.197	3.198	2.863	2.871	0.437	0.449	0.002	0.083
	high	3.198	3.200	2.869	2.877	0.434	0.452	0.004	0.103
2	low	2.483	2.485	2.842	2.862	0.386	0.393	0.001	0.059
$(Z_i \perp\!\!\!\perp$	medium	2.481	2.485	2.885	2.907	0.385	0.394	-0.000	0.072
$G_i \mid \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	high	2.473	2.477	2.882	2.919	0.379	0.391	-0.005	0.088
3	low	3.668	3.669	5.921	5.951	0.559	0.562	-0.038	0.045
$(Z_i \perp\!\!\!\perp G_i \mid \mathbf{X}_i^{ind}, \texttt{degree}_i)$	medium	3.668	3.668	5.944	5.975	0.558	0.561	-0.040	0.046
	high	3.669	3.669	5.943	5.982	0.558	0.561	-0.038	0.048
3	low	-3.280	3.280	-5.182	5.182	0.367	0.370	-0.019	0.056
$(Z_i \not\!\perp\!\!\!\perp G_i \mid \mathbf{X}_i^z, \mathbf{X}_i^{neigh})$	medium	-3.277	-3.277	-5.182	5.182	0.368	0.371	-0.021	0.057
	high	-3.281	3.281	-5.178	5.178	0.364	0.367	-0.026	0.059

- An unadjusted regression-based estimator, that regresses the outcome on both the individual treatment Z_i and the neighborhood treatment G_i – with an interaction term –, without adjusting for covariates;

- An estimator based on a subclassification on the individual propensity score $\phi(1, \mathbf{X}_i^z)$, which is estimated using both individual and neighborhood covariates;

- A model-based estimator that regresses the outcome on both the individual treatment Z_i and the neighborhood treatment G_i – with an interaction term –, adjusting for the neighborhood propensity score $\lambda(g; z, \mathbf{X}_i^g)$ (GPS approach, Hirano & Imbens, 2004);

- The estimator proposed in Section 5 (Sublcassification & GPS), which is based on subclassification on the individual propensity score and model-based adjustment for the neighborhood propensity score.

In Tables 3 and 4 we report the bias and root mean squared error (RMSE) of all estimators for spillover effects $\Delta(0)$ and $\Delta(1)$, respectively (Eq. 5). In observational studies affected by interference, when SUTNVA assumption holds, estimators of spillover effects, relying on the correct function $g_i(\cdot)$, are unbiased as long as they properly adjust for confounding covariates. The unadjusted estimator and the two estimators that only adjust for
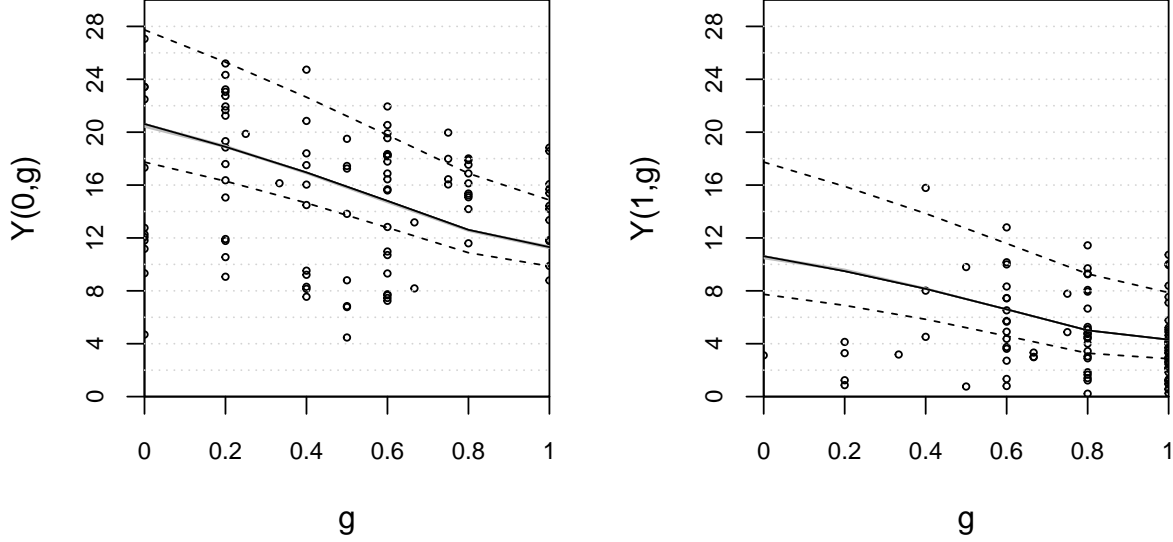
Figure 1: Estimated Dose-Response Functions $\mu(0, g)$ and $\mu(1, g)$.

one of the two propensity scores show a significant bias. The estimator that adjusts for the neighborhood propensity score $\lambda(g; z, \mathbf{X}_i^g)$ performs better than the one based only on the individual propensity score $\phi(1, \mathbf{X}_i^z)$, given the different nature of adjustment. Our proposed estimator based on both propensity scores performs well in all scenarios.

Figure 1 depicts the scatterplot of the observed outcomes and the estimated ADRFs $\mu(0, g)$ and $\mu(1, g)$ against all possible values of the neighborhood treatment $g \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. These plots are derived for Scenario 2, although all scenarios show similar results (Scenario 3 has a different domain for g in the x-axis). Solid lines represent the estimated marginal ADRFs, whereas dashed lines represent $\mu(0, g)$ (left) and $\mu(1, g)$ (right) estimated in two subclasses defined by the individual propensity score: the top and bottom lines correspond to a low and high value of $\phi(1, \mathbf{X}_i^z)$, respectively. The nonlinear trend is a result of the dependence of the outcome from the neighborhood propensity score in both the generation and the estimation model. 95% confidence intervals are obtained using bootstrap methods and are illustrated by the gray bands. Here the small standard deviation is due to the large sample size. As mentioned, our dose-response estimator allows estimating both main and spillover effects, which can be easily computed from the functions $\mu(0, g)$ and $\mu(1, g)$. In particular, spillover effects $\delta(0, g)$, shown in Figure 2 (left), can be computed by subtracting $\mu(0, 0)$ from $\mu(0, g)$ for each value of g.

Similarly, the difference between $\mu(1, g)$ and $\mu(1, 0)$ leads to the estimation of spillover effects $\delta(1, g)$, shown in Figure 2 (right). We can see that health insurance coverage among friends has an effect on a student's health. In particular, a greater proportion of insured
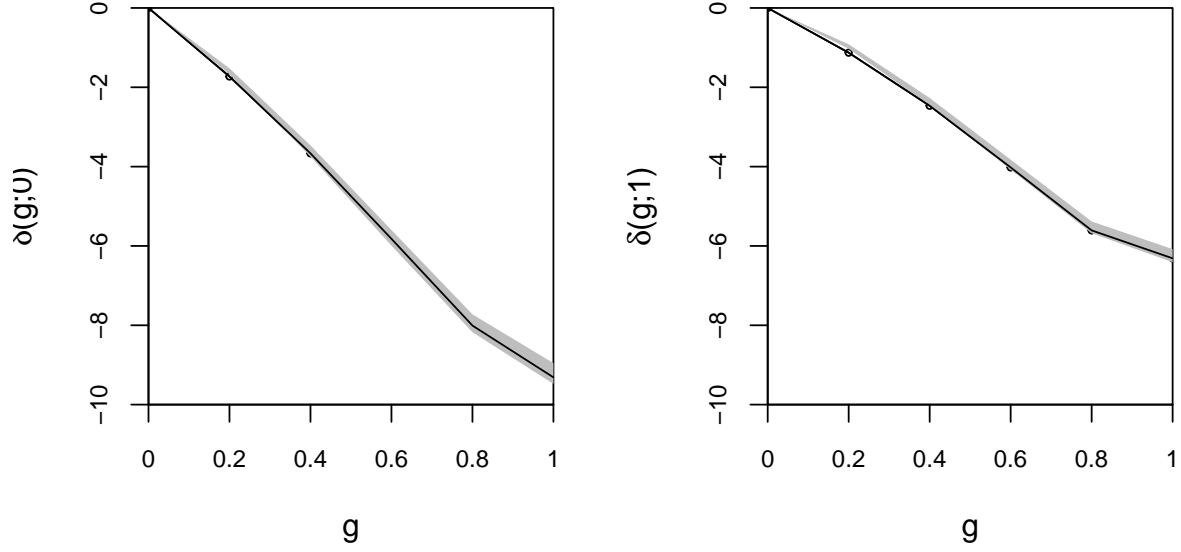
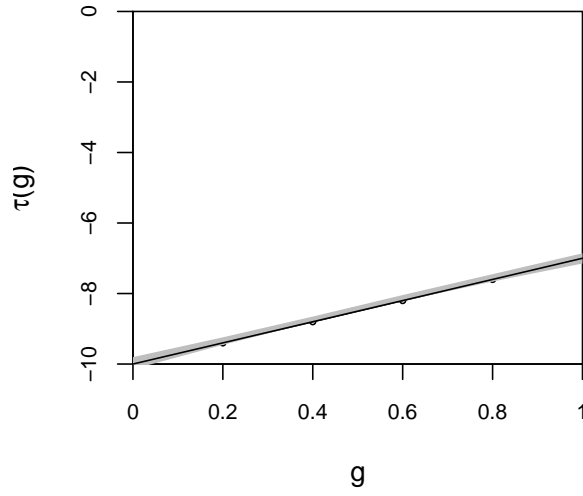Figure 2: Estimated Spillover Effects $\delta(g; z)$.

Figure 3: Estimated Main Effects $\tau(g)$.

friends result in a reduction of the average number of school days that a student would miss because of illness. The effect is larger for those who are not insured themselves ($\delta(0, g) > \delta(1, g)$, $\forall g$). Figure 3 depicts estimated main effects $\tau(g)$, which can be obtained from the difference between $\mu(1, g)$ and $\mu(0, g)$ for each value of g. The individual effect of being covered by some health insurance is reduced for those with a higher coverage among friends. More estimation and inference results are reported in Appendix C.

7   Concluding Remarks

The paper contains several contributions. We have introduced a relaxed version of SUTVA, named SUTNVA, that simplifies in a meaningful way the mechanism of interference in networks. Under SUTNVA, potential outcomes are defined as a function of a bivariate treatment variable, (Z_i, G_i), which consists of the individual treatment and the neighborhood treatment, which is a function of the vector of treatments received by the unit's neighbors. The framework presented in this paper is based on a compound assignment mechanism for the bivariate individual and neighborhood treatment, which is seen as a joint treatment to which each unit is assigned. Within this framework, we have introduced new causal estimands for the treatment and spillover effects , which disentangle the total effect of the individual treatment receipt and the exposure to the neighbors' treatments. We have then laid out an unconfoundedness assumption for the joint treatment, discussing its implications and plausibility in networks.

We have derived explicit bias formulas under several scenarios for the treatment effects when SUTVA is wrongly assumed. In this case, there can be a combination of a bias due to not adjusting for neighborhood confounders and a bias due to the presence of interference. The latter is proportional to the level of interference and to the association between the individual and the neighborhood treatments. Note that this result does not depend on the choice of the function defining the neighborhood treatment. An important result is that, when the set of covariates in the adjustment set suffices for the independence of the two treatments, the bias due to interference is canceled out; it demystifies the misconception that bias resulting from neglecting interference solely depends on the level of interference itself. Regardless of the level of interference, the treatment effect estimated using a naive approach would not be biased if the treatments received by units of the same neighborhood are conditionally independent. Oftentimes, this is likely to be true. This means that we would only need to worry when there are reasons to believe that there is a direct dependence among treatments in the form of peer influence or that we have not included important neighborhood covariates that are part of the assignment mechanism. The implications of our derivations are significant for researchers in both experimental and observational studies. In experiments, our results help to understand and predict the risk of different design strategies. In observational studies, researchers will have a better sense of whether they should trust their conclusions and will make them think more carefully of ways to reduce the probability of bias due to interference. For example, they may collect covariates that are not necessarily confounders but are predictive of both the individual treatment and the

neighborhood treatment. These results may be used to develop a formal sensitivity analysis which is part of our future research agenda.

If we do take interference into account, proper covariate-adjustment methods are needed under the new version of unconfoundedness. Formal theory for designing and analyzing observational studies with units organized in a network is still in its infancy. We have extendend the theory on propensity score as a balancing score. The proven balancing and unconfoudedness properties of both the joint propensity score and the individual and neighborhood propensity scores can open a whole new field of research. It could give rise to different propensity score-based estimators that remove bias by adjusting for either the joint propensity score or for the individual and neighborhood propensity scores separately. The easier way to adjust for the propensity score of a multivariate treatment is through a Horvitz-Thompson inverse probability-weighted estimator. This idea has already been used in the interference literature, however only in experimental settings, or in observational studies but under partial interference (e.g. Tchetgen Tchetgen & VanderWeele (2012); Aronow & Samii (2017); Liu et al. (2016)). A clear use of Horvitz-Thompson inverse probability-weighted estimator in observational network data is still lacking. We have instead focused on another adjustment method based on Hirano and Imbens (2004) model-based generalized propensity score approach.

We have proposed a semi-parametric propensity-score based estimator of both treatment and spillover effects. We define a joint propensity score that balances individual and neighborhood covariates across units under different levels of such a bivariate treatment. If we assume interference to operate only through a function of the vector of friends' treatments, then spillover effects can be seen as a dose-response function of a multivalued treatment. We factorize the bivariate-treatment propensity score and estimate the dose response function by combining subclassification on the individual binary-treatment propensity score and parametric adjustment for the neighborhood multivalued-treatment propensity score. The reason for the focus on this type of estimator stems from its flexibility. In fact, although it heavily relies on the correctness of the outcome model, extensions to more complex non-linear models is straightforward. In addition, thanks to its imputation approach that uses the estimated model to impute the specific potential outcomes of interest, it can be used to estimate different types of causal estimands.

The first simulation study, focused on the main effect, was used to i) assess the performance of the proposed estimator in the estimation of the main effect, ii) show the bias resulting from neglecting interference and using typical parametric and semi-parametric estimators of the treatment effect, ii) shed light on the different sources of bias through different scenar-

ios iii) validate the nonparametric bias formulas. The performance of our semi-parametric propensity-score based estimator was also assessed in the second simulation study for the estimation of the spillover effects. These simulations were also useful to emphasize the need for the adjustment for the joint propensity score in different scenarios.

As with all papers, this also has some limitations. The causal estimands are defined conditional on the observed social network that we assume fully known and fixed. The estimator we propose is consistent with the way we formulated the problem of interference and, for this reason, it is unbiased for the estimation of treatment and spillover effects. This is confirmed by our simulation study, where the estimator is consistent with the generating model. However, because the estimating procedure is based on the subclassification on the individual propensity score and a model-based approach for the neighborhood propensity score, in real studies it might be affected by a bias due to a residual imbalance in the subclasses or to model-misspecification. Due to the complexity of the problem, we expect the true model to be non linear, and, thus, we prefer our semi-parametric estimator over linear regression on individual and neighborhood treatments. We are working on making our semi-parametric estimator less model-dependent.

The problem of statistical inference with units connected in network is very tricky given the correlation structure of the data. Aronow & Samii (2017) derive finite-sample variance for an IPW estimator in experiments. Liu et al. (2016) rely on the assumption of partial interference, whereas Van der Laan (2014) and its extensions rely on independence assumptions. To the best of our knowledge, this problem is still an open research area. In Appendix C we propose a bootstrapping procedure with resampling at the unit-level or at the cluster-level. These resampling technique quantify uncertainty due to the two sampling schemes, egocentric or cluster sampling, respectively. Under alternative sampling schemes, different resampling methods to derive standard errors should be further investigated.

Directions for future research include using Bayesian semiparametric approaches to inference, which could potentially overcome the problems of quantifying uncertainty, developing novel sensitivity analysis techniques to evaluate how departures from the unconfoundedness and SUTNVA affect treatment effect estimates, developing new methodological tools to account for network uncertainty when true interactions between individuals are either not observed or measured with error.

References

Aronow, P. M. (2012). A General Method for Detecting Interference Between Units in Randomized Experiments. *Sociological Methods & Research*, 41(3), 3–16.

Aronow, P. M. & Samii, C.(2017). Estimating Average Causal Effects Under General Interference. *Forthcoming. Annals of Applied Statistics. Preprint:arxiv:1305.6156.*

Athey, S., Eckles, D., & Imbens, G.W. (2015). Exact p-values for network interference. *NBER Working Paper 21313.*

Binka, F. N., Indome, F., & Smith, T. (1998). Impact of spatial distribution of permethrin-impregnated bednets on child mortality in rural northern Ghana. *American Journal of Tropical Medicine and Hygiene*, 59, 80–5.

Bowers, J., Fredrickson, M.M., & Panagopoulos, C. (2013). Reasoning about interference between units: a general framework. *Political Analysis*, 21(1), 97–124.

Cox, D. R. (1958). Planning of Experiments. *Wiley; New York.*

Eckles, D., Karrer, B., & Ugander, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. *arxiv:1404.7530.*

Halloran, M. E. & Hudgens, M. G. (2016). Dependent Happenings: A Recent Methodological Review. *Current Epidemiology Reports*, 3(4), 297–305.

Halloran, M. E., & Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology*, 6(2),142–151.

Hawley, W. A., Phillips-Howard, P. A., ter Kuile, F. O., Terlouw, D. J., Vulule, J. M., Ombok, M., Nahlen B. L., Gimnig, J. E., Kariuki, S. K., Kolczak, M. S., & Hightower, A. W. (2003). Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. *American Journal of Tropical Medicine Hygiene*, 68, 121–27.

Hernán, M. A., & Robins, J. M. (2018). Causal Inference. *Boca Raton: Chapman & Hall/CRC, forthcoming.*

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. *In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, ed. A. Gelman and X.-L. Meng, 73–84. West Sussex, England: Wiley InterScience.*

Hong, G. & Raudenbush, S. W. (2006). Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data. *Journal of the American Statistical Association*, 101, 901–910.

Howard, W. A., Omumbo, J., Nevill, C., Some, E. S., Donnelly C. A., &Snow, R. W. (2000). Evidence for a mass community effect of insecticide-treated bednets on the

incidence of malaria on the Kenyan coast. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 94, 357–60.

HUDGENS, M. G. & HALLORAN, M. E. (2008). Towards causal inference with interference. *Journal of the American Statistical Association*, 103, 832–842.

IMBENS, G. W. (2000). The Role of the Propensity Score in Estimating Dose-Response Functions. *Biometrika*, 87, 706–710.

IMBENS, G. W., & RUBIN, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. *Cambridge University Press New York, NY, USA, 2015.*

KOLACZYK, E. D. (2009). Statistical Analysis of Network Data. *New York: Springer–Verlag.*

LIU, L. & HUDGENS, M.G. (2013). Large Sample Randomization Inference of Causal Effects in the Presence of Interference . *Journal of the American Statistical Association*, 109(505), 288–301.

LIU, L., HUDGENS, M.G., & BECKER-DREPS, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika*, 103, 829–842.

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

OGBURN, E.L., SOFRYGIN, O., DIAZ, I. & VAN DER LAAN, M.J. (2017). Causal inference for social network data. https://arxiv.org/abs/1705.08527.

PEREZ-HEYDRICH, C., HUDGENS, M.G., HALLORAN, M.E., CLEMENS, J.D., ALI, M. & EMCH, M.E. (2014). Assessing effects of cholera vaccination in the presence of interference. *Biometrics*, 70(3), 731–44.

PERRI, B., PESCOSOLIDO, B. & BORGATTI, S. (2018). Egocentric Network Analysis: Foundations, Methods, and Models (Structural Analysis in the Social Sciences). *Cambridge: Cambridge University Press.*

ROSENBAUM, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, 102(477), 191–200.

ROSENBAUM, P. R. (2010). Design of observational studies. *New York, NY: Springer-Verlag.*

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

RUBIN, B. D. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology* 66, 688–701.

RUBIN, D. B. (1980). Comment on "Randomization Analysis of Experimental Data in the

Fisher Randomization Tes" by D. Basu. *Journal of the American Statistical Association*, 75, 591–593.

RUBIN, D. B. (1986). Which Ifs have Causal Answers? Comment on "Statistics and Causal Inference" by P. Holland. *Journal of the American Statistical Association*, 81, 961–962.

SÄVJE, F., ARONOW, P. M. & HUDGENS, M. G. (2017). Average treatment effects in the presence of unknown interference. *arXiv.org*, no. 1711.06399.

SOBEL, M. E. (2006). What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference. *Journal of the American Statistical Association*, 101, 1398–1407.

SOFRYGIN, O. & VAN DER LAAN, M. (2017). Semi-Parametric Estimation and Inference for the Mean Outcome of the Single Time-Point Intervention in a Causally Connected Population. *Journal of Causal Inference*, 5(1).

SINCLAIR, B. (2011). Design and Analysis of Experiments in Multilevel Populations. *In Cambridge Handbook of Experimental Political Science. Cambridge University Press p. 906.*.

UGANDER, J., KARRER, B., BACKSTROM, L. & KLEINBERG, J. (2013). Graph Cluster Randomization: Network Exposure to Multiple Universes. *Proc. 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*.

TCHETGEN TCHETGEN, E. J. & VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21, 55–75.

TOULIS, P. & KAO, E.K. (2013). Estimation of Causal Peer Influence Effects. *ICML*, (3) 2013: 1489–1497.

VAN DER LAAN, M.J. (2014). Causal Inference for a Population of Causally Connected Units. *Journal of Causal Inference*, 0, 2193–3677.

VANDERWEELE, T. J., HONG, G., JONES, S. M., & BROWN, J. L. (2013). Mediation and Spillover Effects in Group-Randomized Trials: A Case Study of the 4Rs Educational Intervention. *Journal of the American Statistical Association*, 108:502, 469–482.

VANDERWEELE, T. J., TCHETGEN TCHETGEN, E. J., & HALLORAN, M. E. (2012). Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. *Epidemiology*, 23(5), 751–761.

VANDERWEELE, T. J., TCHETGEN TCHETGEN, E. J., & HALLORAN, M. E. (2014). Interference and sensitivity analysis. *Statistical Science*, 29(4), 687–706.

VERBITSKY-SAVITZ, N. & RAUDENBUSH, S. W. (2012). Causal inference under interference in spatial settings: a case study evaluating community policing program in Chicago. *Epidemiological Methods*, 1(1), 106–130.

Online Appendix

A Propensity Score-Based Estimator for Conditional Main and Spillover Effects

A.1 Conditional Main Effects

In one experiment, we can observe one of the two potential outcomes $Y_i(1, g)$ and $Y_i(0, g)$ defining the main effect $\tau(g)$ only for units who exhibit a neighborhood treatment $G_i = g$. For all the other units in V_g these potential outcomes are potentially observable but none of them is actually observed. For this reason let us define a *conditional main effect*, which is the average effect of the individual treatment only for those with $G_i = g$:

$$\tau_g(g) = E \left[Y_i(Z_i = 1, G_i = g) - Y_i(Z_i = 0, G_i = g)|G_i = g \right] \tag{18}$$

If we condition on a particular value of G_i, then we only have one treatment, which is the individual treatment Z_i. In observational studies, this treatment depends on a set o covariates. To be able to get an unbiased estimate of $\tau_g(g)$, we need to make the following assumption:

Assumption 4 (Conditional Unconfoundedness of the Individual Treatment)**.**

$$Y_i(z, g) \perp\!\!\!\perp Z_i \mid G_i = g, \boldsymbol{X}_i^{z|g} \qquad \forall z \in \{0, 1\}, g \in \mathcal{G}_i$$

Assumption 4 states that units with neighborhood treatment $G_i = g$ and with the same value of the vector $\mathbf{X}_i^{z|g}$ have the same probability of being treated, which does not depend on the potential outcomes $Y_i(z, g)$. $\mathbf{X}_i^{z|g}$ is a sub-vector of $\mathbf{X}_i$ containing all those covariates that would affect the decision of taking the treatment for unit i, given that the neighborhood treatment is g. It is worth noting that Assumption 3 given $\mathbf{X}_i$ implies Assumption 4. The proof just follows from the fact that joint independence implies conditional independence and that by definition of $\mathbf{X}_i^{z|g}$ we have that $P(Z_i = 1|G_i, \mathbf{X}_i) = P(Z_i = 1|G_i, \mathbf{X}_i^{z|g})$. Under Assumption 4 an unbiased estimator of (18) can be obtained by any unbiased estimator of

the quantity

$$\tau_{g,x}^{obs} = \sum_{\mathbf{x} \in \mathcal{X}} E\left[\, Y_i | Z_i = 1, G_i = g, \mathbf{X}_i = \mathbf{x}\,\right] - E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i = \mathbf{x}\,]\; P(\mathbf{X}_i = \mathbf{x})$$

A.1.1 Conditional Propensity Scores of the Individual Treatment

If the vector $\mathbf{X}_i^{z|g}$ has large dimension or includes continuous covariates, then adjustment for these covariates cannot be performed by stratification. As Rosenbaum & Rubin (1983) proposed for the no-interference case, adjustment for covariates can be achieved through the use of propensity scores. Let us define the *conditional individual propensity score* as the probability of being treated, given a value of the neighborhood treatment and the vector of covariates:

$$e(x^{z|g}; g) = P(Z_i = 1 | G_i = g, \mathbf{X}_i^{z|g} = \mathbf{x}) \tag{19}$$

The balancing property and the property of unconfoundedness given the conditional individual propensity score can be formalized as follows.

Proposition 6 (Balancing Property of the Conditional Individual Propensity Score)**.** *The propensity score of the individual treatment is a balancing score:*

$$Z_i \perp\!\!\!\perp \boldsymbol{X}_i^{z|g} \mid G_i = g, e(\boldsymbol{X}_i^{z|g}; g)$$

Proof. Refer to Rosenbaum & Rubin (1983).

Proposition 7 (Conditional Unconfoundedness of Z_i given the Conditional Individual Propensity Score)**.** *If assumption 4 holds*

$$Y_i(z, g) \perp\!\!\!\perp Z_i \mid G_i = g, e(\boldsymbol{X}_i^{z|g}; g)$$

Proof. Refer to Rosenbaum & Rubin (1983).

A.1.2 Estimator based on Propensity Scores

One possible estimator based on the propensity score $e(\mathbf{X}_i^{z|g}; g)$ is a subclassification estimator (Imbens & Rubin, 2015), which stratifies the subpopulation where $G_i = g$ into J strata $B_j, j \in \{1, \ldots, J\}$, defined by similar values of the propensity score $e(\mathbf{X}_i^{z|g}; g)$, i.e.,

$B_j = \{i \in V_g : G_i = g, b_{j-1} < e(\mathbf{X}_i^{z|g}; g) \leq b_j\}$. Within each stratum the conditional main effect can be estimated non-parametrically by

$$\widehat{\tau}_g^j(g) = \frac{\sum_{i \in B_j} Y_i \mathbb{1}(Z_i = 1)}{\sum_{i \in B_j} \mathbb{1}(Z_i = 1)} - \frac{\sum_{i \in B_j} Y_i \mathbb{1}(Z_i = 0)}{\sum_{i \in B_j} \mathbb{1}(Z_i = 0)} \tag{20}$$

Finally the conditional main effect $\tau_g(g)$ can be estimated by the following average:

$$\widehat{\tau}_g(g) = \sum_{j=1}^{J} \widehat{\tau}_g^j(g) \pi_j \tag{21}$$

where $\pi_j = \frac{|B_j|}{\sum_{i \in V_g} \mathbb{1}(G_i = g)}$

A.2 Conditional Spillover Effects

Let us now focus on spillover effects for units that are in the control group. We define as *conditional spillover effect*, denoted with $\delta_0(g; 0)$, the following quanity:

$$\delta_0(g; 0) = E\left[\, Y_i(Z_i = 0, G_i = g) - Y_i(Z_i = 0, G_i = 0) \,\middle|\, i \in V_g, Z_i = 0 \,\right] \tag{22}$$

In this case, even units with $Z_i = 0$ can exhibit an outcome that is not one of the two potential outcomes defining the effect. This is because G_i is not a binary treatment. Nevertheless, if we are only interested in estimating $\delta_0(g; 0)$ for a specific value g, we can ease the problem by defining a new variable

$$T_i(g) = \begin{cases} 1 & \text{if} \quad G_i = g \\ 0 & \text{if} \quad G_i = 0 \\ \text{NA} & \text{if} \quad G_i \neq g, 0 \end{cases}$$

We can then focus on the subpopulation with $Z_i = 0$ and $T_i \in \{0, 1\}$ and proceed with a standard estimation procedure for the effect of a binary treatment binary T_i. The identifying unconfoudedness assumption can be expressed as follows.

Assumption 5 (Conditional Unconfoundedness of the Neighborhood Treatment)**.**

$$Y_i(0, g) \perp\!\!\!\perp T_i(g) \mid Z_i = 0, \boldsymbol{X}_i^{g|0}, G_i = \{g, 0\}$$

3

In principle, $\mathbf{X}_i^{g|0}$ can be a subvector of $\mathbf{X}_i$. This is the set of covariates that makes the probability of having a specific value of the neighborhood treatment for units under control independent of the potential outcome $Y_i(0, g)$. As with Assumption 4 in Section A.1, Assumption 3 does imply Assumption 5.

A.2.1 Conditional Propensity Scores of the Neighborhood Treatment

To use a propensity score-based estimator, we can define the *conditional neighborhood propensity score* as follows:

$$r_{g,0}(x) = P(T_i = 1 | Z_i = 0, \mathbf{X}_i^{g|0} = \mathbf{x}, G_i = \{g, 0\}) \tag{23}$$

satisfying the following properties.

Proposition 8 (Balancing Property of the Conditional Neighborhood Propensity Score)**.** *The conditional neighborhood propensity score is a balancing score:*

$$T_i \perp\!\!\!\perp \boldsymbol{X}_i^g \mid Z_i = 0, G_i = \{0, 1\}, r_{g,0}(\boldsymbol{X}_i^g)$$

Proof. Refer to Rosenbaum & Rubin (1983). □

Proposition 9 (Conditional Unconfoundedness of Z_i given the Conditional Neighborhood Propensity Score)**.** *If assumption 5 holds*

$$Y_i(z, g) \perp\!\!\!\perp T_i(g) \mid Z_i = 0, r_{g,0}(\boldsymbol{X}_i^{g|0}), G_i = \{g, 0\}$$

Proof. Refer to Rosenbaum & Rubin (1983). □

One possible estimator exploiting this result can be a subclassification-based estimator, as in Section A.1.1.

As shown in Sections A.1 and A.2, according to Propositions 7 and 9, conditional effects can be estimated using common propensity score-based estimators for binary treatments on a subset of units. We do not illustrate the performance of these estimators, which has been extensively assessed in the literature (e.g., Imbens & Rubin, 2015).

B Data Generating Model for Simulations

As introduced in Section 6, the simulation study exploits the friendship network collected in the Add Health Data. This dataset also provided information on two students' characteristics, namely, race and grade. In our simulation we have: $\texttt{race}_i \in \{0,1\}$ and $\texttt{race}_i \in \{6,7,...,12\}$. On the contrary, the treatment (Z_i, G_i) and the outcome Y_i variables were generated using different generating processes. Here we describe the data generating model for the four scenarios.

B.1 Correlation Structure of the Joint Treatment and Covariate Balance

Scenario 1. In the fist scenario, the individual treatment Z_i is generated depending on individual race and grade, with the following propensity score

$$\texttt{logit}(P(Z_i = 1)) = -18 + 2\texttt{grade}_i + 3\texttt{race}_i \tag{24}$$

and the neighborhood treatment G_i is defined as the proportion of treated friends among the first five best friends. This generating procedure leads to a specific correlation structure between neighborhood and individual treatments and covariates. Here we describe the distribution of these variables in one simulated dataset. The following table shows the number of treated and untreated students and the average proportion of treated friends in the whole population, for treated and for untreated units. The proportion of treated students of 0.743 is similar to the actual proportion of students with health insurance in the real dataset.

	All	$Z_i = 1$	$Z_i = 0$
N	16410	12188	4222
$\overline{Z}$	0.743	1	0
$\overline{G}$	0.662	0.677	0.617

Table 5 describes the balance of covariate distributions across individual treatment arms. The imbalance of individual covariates follows from the propensity score, whereas for neighborhood covariates the small difference is presumably due to the presence of homophily, that is, students are more likely to have friends with similar characteristics. For the same reason, the proportion of treated friends depends on individual covariates, other than neighborhood

5

Table 5: Covariate balance across individual treatment arms

Variables	$\bar{X}_T$	$\bar{X}_C$	Stand. Diff.
Grade	10.097	7.531	1.873
Race	0.754	0.364	0.905
Friends' Grade	9.186	8.949	0.227
Friends' Race	0.588	0.541	0.144
Degree	7.892	8.053	-0.037
G_i	0.694	0.633	0.230

Table 6: Covariate balance across dichotomized neighborhood treatment arms

Variables	$\bar{X}_{G_i \geq 0.5}$	$\bar{X}_{G_i < 0.5}$	Stand. Diff.
Grade	9.546	9.070	0.270
Race	0.651	0.661	-0.020
Friends' Grade	9.451	8.027	1.584
Friends' Race	0.636	0.373	0.847
Degree	7.783	8.442	-0.153
Z_i	0.764	0.671	0.219

covariates as follows from the individual propensity score depending on individual covariates. The covariate balance across units with different neighborhood treatment is described by Table 6, which shows the distribution of covariates in two arms defined by a dichotomized neighborhood treatment, and by Table 6, which reports the coefficients of a weighted logistic regression. In this scenario, the assignment mechanism only based on individual covariates, should result in a zero correlation between the neighborhood treatment and the individual treatment after conditioning for $\mathbf{X}_i^{ind}$. The small association seen in Table 5 and 6, is perhaps due to the homophily mechanism, as we can see in Table 7, where the association is zero after conditioning for covariates.

Scenario 2. The individual treatment Z_i is generated depending on individual race and grade, and on friends' race and grade with the following propensity score

$$\texttt{logit}(P(Z_i = 1)) = -47 + 2\texttt{grade}_i + 4\texttt{race}_i + 3\texttt{friends.grade}_i + 5\texttt{friends.race}_i \quad (25)$$

Table 7: Coefficients of logistic regression of neighborhood treatment

Variables	Estimate	SE
Grade	0.011	0.007
Race	-0.021	0.020
Friends' Grade	0.92	0.010
Friends' Race	1.401	0.028
Degree	-0.002	0.002
Z_i	0.017	0.028

Table 8: Covariate balance across individual treatment arms

Variables	$\bar{X}_T$	$\bar{X}_C$	Stand. Diff.
Grade	10.024	7.650	1.660
Race	0.752	0.355	0.919
Friends' Grade	9.286	8.637	0.656
Friends' Race	0.632	0.405	0.730
Degree	7.906	8.018	-0.026
G_i	0.687	0.505	0.687

The neighborhood treatment G_i is the proportion of treated friends among the first five best friends. This assignment mechanism leads to the following distribution of Z_i and G_i

	All	$Z_i = 1$	$Z_i = 0$
N	16410	12064	4346
$\overline{Z}$	0.735	1	0
$\overline{G}$	0.626	0.686	0.459

and covariate balance across individual treatment arms, shown in Table 8. As in Scenario 1, the neighborhood treatment G_i will depends on neighborhood covariates, as seen in Table 9 and Table 10. Hence, Z_i and G_i are dependent if we condition only on $\mathbf{X}_i^{ind}$, but independent conditional on $\mathbf{X}_i^{neig}$. In Table 10 we report the coefficients of a weighted logistic regression of the neighborhood treatment on covariates. We used this estimated model to predict the neighborhood propensity score $\lambda(g; 0, \mathbf{X}_i^g)$ for each unit. As an illustration, here we report the difference in the predicted $\lambda(g; 0, \mathbf{X}_i^g)$ across units with different race (Figure 4), grade (Figure 5), friend's race and grade (Figure 6). The dependence of the the neighborhood treatment from friends' grade is presumably due to homophily, since students are likely to have friends in the same class.

Table 9: Covariate balance across dichotomized neighborhood treatment arms

Variables	$\bar{X}_{G_i \geq 0.5}$	$\bar{X}_{G_i < 0.5}$	Stand. Diff.
Grade	9.561	9.120	0.300
Race	0.656	0.649	0.013
Friends' Grade	9.512	8.128	1.585
Friends' Race	0.632	0.432	0.630
Degree	7.803	8.272	-0.109
Z_i	0.820	0.393	0.627

Table 10: Coefficients of logistic regression of neighborhood treatment

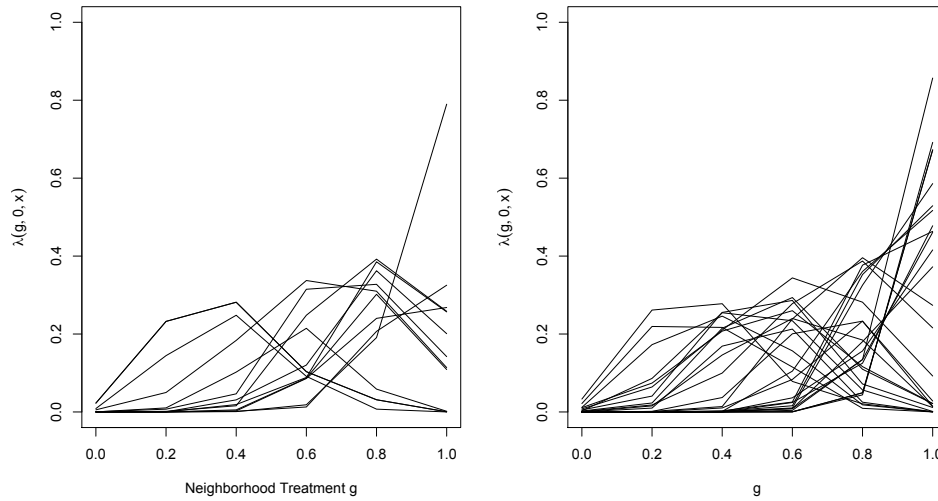| Variables	Estimate	SE
Grade	0.016	0.006
Race	0.025	0.020
Friends' Grade	0.935	0.010
Friends' Race	0.978	0.029
Degree	0.000	0.002
Z_i	0.044	0.028

Figure 4: Neighborhood propensity score across race: 0 (left), 1 (right).
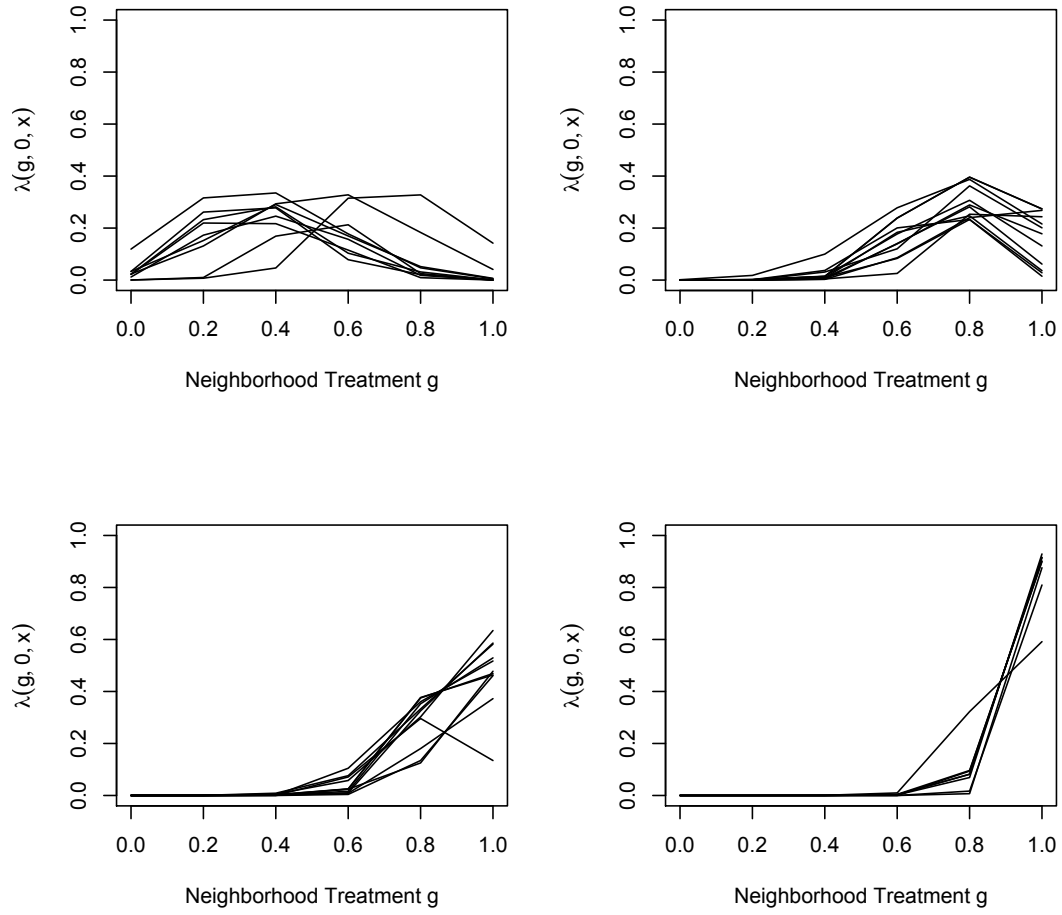
Figure 5: Neighborhood propensity score across grade: 7 (top left), 9 (top right), 10 (bottom left), 12 (bottom right).
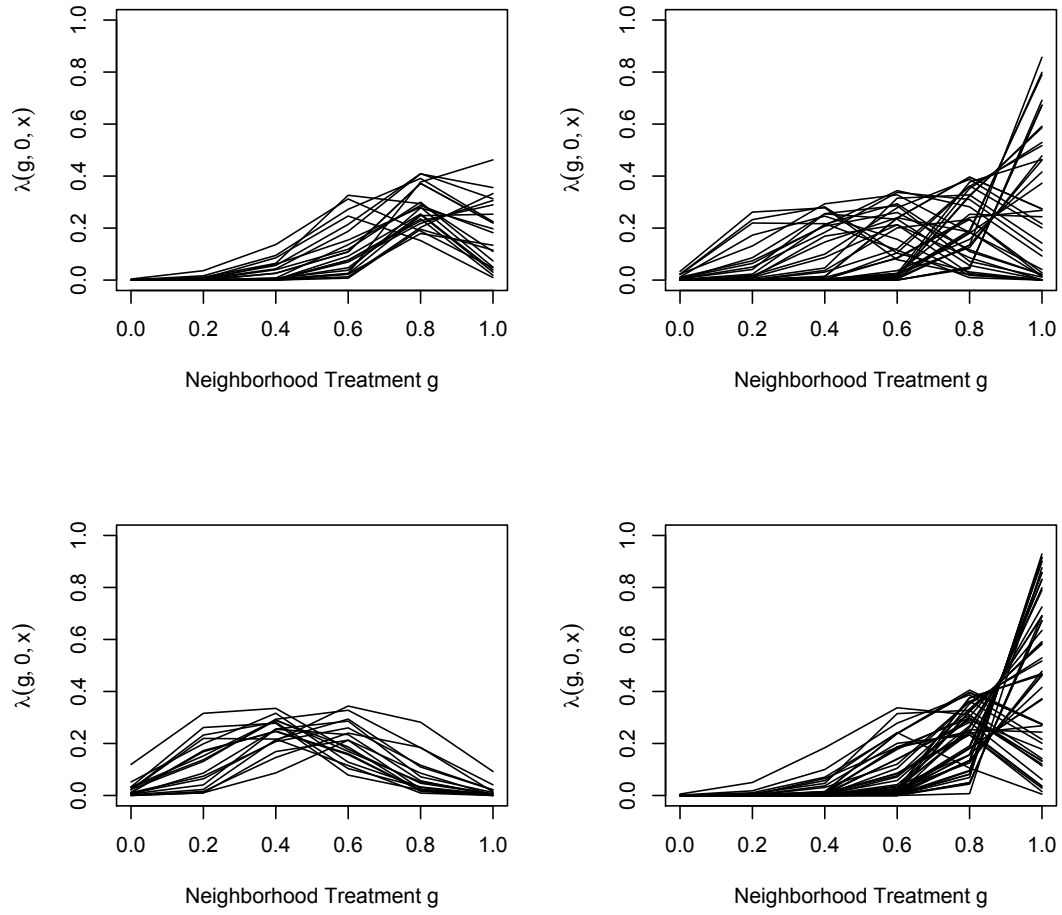
Figure 6: Neighborhood propensity score across friends' race (top left: prevalence of non-white friends, top right: prevalence of white friends) and grade (bottom left: prevalence of friends in lower grades, bottom right: prevalence of friends in higher grades).

Table 11: Covariate balance across individual treatment arms

Variables	$\bar{X}_T$	$\bar{X}_C$	Stand. Diff.
Grade	9.694	8.628	0.653
Race	0.692	0.532	0.346
Friends' Grade	9.631	8.806	0.566
Friends' Race	0.678	0.586	0.268
Degree	9.417	3.264	1.597
G_i	6.786	2.323	1.440

Table 12: Covariate balance across dichotomized neighborhood treatment arms

Variables	$\bar{X}_{G_i \geq 0.5}$	$\bar{X}_{G_i < 0.5}$	Stand. Diff.
Grade	9.408	9.464	-0.033
Race	0.666	0.642	0.050
Friends' Grade	9.512	8.128	1.585
Friends' Race	0.632	0.432	0.630
Degree	11.284	4.87	1.845
Z_i	0.989	0.548	4.422

Scenario 3. This scenario was generated with the purpose of making Z_i and G_i associated trough a unit's degree. In fact, Z_i is generated depending on individual covariates, and on the student's degree, using

$$\texttt{logit}(P(Z_i = 1)) = -49 + 3\texttt{grade}_i + 4\texttt{race}_i + 4N_i \tag{26}$$

G_i is the number of treated friends among all friends and, thus, it depends on the degree. From the following distribution we can see that the average number of treated friends differs in the two individual treatment arms.

	All	$Z_i = 1$	$Z_i = 0$
N	16410	11865	4544
$\overline{Z}$	0.723	1	0
$\overline{G}$	5.209	6.119	2.833

Covariate balance across treatments are reported in Tables 11, 12 and 13.

Scenario 4. In this scenario the generating procedure is such that the individual and neighborhood treatments are directly correlated. We used an iterative procedure that, starting form initial values, it assigns the individual treatment depending on individual covariates

Table 13: Coefficients of logistic regression of neighborhood treatment

Variables	Estimate	SE
Grade	-0.004	0.007
Race	0.012	0.018
Friends' Grade	0.014	0.008
Friends' Race	0.084	0.024
Degree	-0.002	0.002
Z_i	0.033	0.024

Table 14: Covariate balance across individual treatment arms

Variables	$\bar{X}_T$	$\bar{X}_C$	Stand. Diff.
Grade	10.204	7.431	2.197
Race	0.744	0.417	0.750
Friends' Grade	9.200	8.928	0.261
Friends' Race	0.591	0.536	0.170
Degree	7.939	7.919	0.004
G_i	0.930	0.233	4.125

and on the neighborhood treatment, using the following model

$$\texttt{logit}(P(Z_i = 1)) = -20 + 2\texttt{grade}_i + 3\texttt{race}_i + 4G_i \qquad (27)$$

At each iteration, the neighborhood treatment G_i is simply computed as the proportion of treated friends among the first 5 best friends.

	All	$Z_i = 1$	$Z_i = 0$
N	16410	11773	4637
$\overline{Z}$	0.717	1	0
$\overline{G}$	0.719	0.915	0.223

Covariate balance across treatments are reported in Tables 14, 15 and 16. Here we can see that there is still a residual dependence between Z_i and G_i even after conditioning for all types of covariates.

Table 15: Covariate balance across dichotomized neighborhood treatment arms

Variables	$\bar{X}_{G_i \geq 0.5}$	$\bar{X}_{G_i < 0.5}$	Stand. Diff.
Grade	10.085	7.455	1.778
Race	0.716	0.463	0.560
Friends' Grade	9.205	8.880	0.313
Friends' Race	0.589	0.534	0.169
Degree	7.948	7.887	0.014
Z_i	0.925	0.105	3.122

Table 16: Coefficients of logistic regression of neighborhood treatment

Variables	Estimate	SE
Grade	0.802	0.013
Race	0.534	0.027
Friends' Grade	0.176	0.039
Friends' Race	0.106	0.012
Degree	0.026	0.003
Z_i	1.934	0.034

B.2   Outcome Models

B.2.1   Outcome Models for Main Effects Simulations

In the first simulation study for the estimation of main effects, the potential outcomes have the following distribution:

$$
\begin{aligned}
Y_i(z, g) | \mathbf{X}_i^{ind} &\sim \mathcal{N}\big(\mu(z, g, \mathbf{X}_i^{ind}), 1\big) \\
\mu(z, g, \mathbf{X}_i^{ind}) &= 15 - 7\mathbf{I}(\phi(1; \mathbf{X}_i^{ind}) \geq 0.7) - 15z + 3z\mathbf{I}(\phi(1; \mathbf{X}_i^{ind}) \geq 0.7) + \delta g
\end{aligned}
\tag{28}
$$

with $\mathbf{X}_i^{ind} = [\texttt{race}_i + \texttt{grade}_i]$ and $\mathbf{X}_i^g = [\texttt{race}_i, \texttt{grade}_i, \texttt{friends.race}_i, \texttt{friends.grade}_i, \texttt{degree}_i]$ and $\delta \in (-5, -8, -10)$ (or $\delta \in (-0.3, -0.5, -0.8)$ for Scenario 3), corresponding to a low, medium and high level of interference. According to this model, the main effects are

$$
\tau(g) = -15 + 3\mathbf{I}(\phi(1; \mathbf{X}_i^{ind}) \geq 0.7) \quad \forall g \in \mathcal{G} \quad \Longrightarrow \quad \tau = -15 + 3\mathbf{I}(\phi(1; \mathbf{X}_i^{ind}) \geq 0.7) \tag{29}
$$

and spillover effects are

$$
\delta(g; z) = \delta g \quad \Longrightarrow \quad \Delta(z) = \delta E[G_i] \quad \forall z = 0, 1 \tag{30}
$$

13

B.2.2 Outcome Models for Spillover Effects Simulations

In the second simulation study for the estimation of spillover effects, the potential outcomes have the following distribution:

$$Y_i(z,g)|\mathbf{X}_i^{ind}, \mathbf{X}_i^g \sim \mathcal{N}\big(\mu(z,g,\mathbf{X}_i^{ind}, \mathbf{X}_i^g), 1\big)$$

$$\mu(z,g,\mathbf{X}_i^{ind}, \mathbf{X}_i^g) = 15 + \texttt{friends.grade}_i + 7\texttt{friends.race}_i - 10\mathbf{I}(\phi(1;\mathbf{X}_i^{ind}) \geq 0.7) - 10z$$
$$+ \delta g - 10\lambda(g;z,\mathbf{X}_i^g) + 5g\mathbf{I}(\phi(1;\mathbf{X}_i^{ind}) \geq 0.7) + 3zg \tag{31}$$

with $\mathbf{X}_i^{ind} = [\texttt{race}_i + \texttt{grade}_i]$ and $\mathbf{X}_i^g = [\texttt{race}_i, \texttt{grade}_i, \texttt{friends.race}_i, \texttt{friends.grade}_i, \texttt{degree}_i]$ and $\delta \in (-5, -8, -10)$ (or $\delta \in (-0.3, -0.5, -0.8)$ for Scenario 3), corresponding to a low, medium and high level of interference. In Scenario 3, where G_i is the number of treated friends, the coefficients of the last two terms of Equation (31) are scaled by $1/10$. According to this model, the main effects are

$$\tau(g) = -10 + 3g \quad \Longrightarrow \quad \tau = -10 + 3E[G_i] \tag{32}$$

and spillover effects are

$$\delta(g;z) = \delta g 5g E[\mathbf{I}(\phi(1;\mathbf{X}_i^{ind}) \geq 0.7)] - 10\lambda(g;z,\mathbf{X}_i^g) + 3zg$$
$$\Longrightarrow \Delta(z) = \delta E[G_i] + 5E[G_i]E[\mathbf{I}(\phi(1;\mathbf{X}_i^{ind}) \geq 0.7)] - 10E[\lambda(G_i;z,\mathbf{X}_i^g)] + 3zE[G_i] \tag{33}$$

B.3 True Main and Spillover Effects

In Table 17 we present a description of all scenarios in terms of correlation between Z_i and G_i, the true values of the overall main effect τ (according to the outcome model in Equation (28)), and the overall spillover effects $\Delta(0)$ and $\Delta(1)$ (according to the outcome model in Equation (31)). True values are computed as the mean of the values estimated in each simulation data according to equations (29) and (33), respectively. We also report, for each scenario, a measure of partial correlation between Z_i and G_i, conditional on the individual covariates:

$$\rho_{ZG|X^{ind}} = \frac{R^2_{G|Z,X^{ind}} - R^2_{G|X^{ind}}}{1 - R^2_{G|X^{ind}}}$$

It is worth noting that the bias for the overall main effect τ, when $\mathbf{X}_i^\star = \mathbf{X}_i^{ind}$, is not directly proportional to this measure of association, but to the quantity $P(G_i = g|Z_i = 1, \mathbf{X}_i^{ind} = \mathbf{x}) - P(G_i = g|Z_i = 0, \mathbf{X}_i^{ind} = \mathbf{x})$, for each value $g \in \mathcal{G}$, as seen in Equation (12).

14

Table 17: Simulation Scenarios and True Effects

Scenario	interference	$\rho_{ZG|X^{ind}}$	τ	$\Delta(0)$	$\Delta(1)$
1 $(Z_i \perp\!\!\!\perp G_i|\mathbf{X}_i^{ind})$	low	0.002	-13.107	-1.961	-0.513
	medium	0.002	-13.107	-3.421	-1.974
	high	0.002	-13.107	-4.384	-2.937
2 $(Z_i \perp\!\!\!\perp G_i|\mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	0.166	-13.109	-2.542	-0.905
	medium	0.166	-13.109	-4.223	-2.585
	high	0.166	-13.109	-5.336	-3.699
3 $(Z_i \perp\!\!\!\perp G_i|\mathbf{X}_i^{ind}, \texttt{degree}_i)$	low	0.236	-13.177	-0.997	0.605
	medium	0.236	-13.177	-2.065	-0.462
	high	0.236	-13.177	-3.667	-2.064
4 $(Z_i \not\perp\!\!\!\perp G_i|\mathbf{X}_i^{z}, \mathbf{X}_i^{neigh})$	low	0.403	-13.109	-1.789	-1.239
	medium	0.403	-13.109	-3.861	-3.311
	high	0.403	-13.109	-5.243	-4.693

C Subclassification and Generalized Propensity Score Estimator

1) We derive a subclassification on the individual propensity score $\phi(1; \mathbf{X}_i^z)$ as follows:

   a) We estimate $\phi(1; \mathbf{X}_i^z)$ with a logistic regression for Z_i conditional on covariates $\mathbf{X}_i^z$;

   b) We predict $\phi(1; \mathbf{X}_i^z)$ for each unit;

   c) We identify J subclasses B_j, with $j = 1, \ldots, J$, defined by similar values of $\phi(1; \mathbf{X}_i^z)$ and where there is sufficient balance between individual treatment groups, i.e., $\mathbf{X}_i^z \perp\!\!\!\perp Z_i | i \in B_j$.

2) Within each subclass B_j, we repeat the following steps to estimate $\mu_j(z, g) = E[Y_i(z, g)| i \in B_j^g]$, where $B_j^g = V_g \cap B_j$:

   a) We estimate the parameters of a model for the neighborhood propensity score $\lambda(g; z; x^g)$: $\lambda(z; g; \mathbf{X}_i^g) = Pr(G_i = g | Z_i = z, \mathbf{X}_i^g) = f^G(g, z, \mathbf{X}_i^g)$

   b) We use the observed data $(Y_i, Z_i, G_i, \mathbf{X}_i^g)$ and $\widehat{\Lambda} = \lambda(G_i; Z_i; \mathbf{X}_i^g)$ to estimate the parameters of a model $Y_i(z, g) | \lambda(z; g; \mathbf{X}_i^g) \sim f^Y(z, g, \lambda(g; z; \mathbf{X}_i^g))$;

   c) For a particular level of the joint treatment $(Z_i = z, G_i = g)$, for each unit

$i \in B_j^g$ we predict the neighborhood propensity score evaluated at that level of the treatment, i.e., $\lambda(g; z; \mathbf{X}_i^g)$, and use it to predict the potential outcome $Y_i(z, g)$.

d) To estimate the dose-response function $\mu_j(z, g)$ we average the potential outcomes over $\lambda(z; g; \mathbf{X}_i^g)$

$$\widehat{\mu}_j(z, g) = \frac{\sum_{i \in B_j^g} \widehat{Y}_i(z, g)}{|B_j^g|}$$

3) We derive the average dose-response function as follows:

$$\widehat{\mu}(z, g) = \sum_{j=1}^{J} \widehat{\mu}_j(z, g) \pi_j^g$$

where $\pi_j^g = \frac{|B_j^g|}{v_g}$.

C.1 Individual Treatment, Neighborhood Treatment and Outcome Models

Here we detail the estimation procedure used in the simulation study. We provide details on the models used for the individual propensity score, the neighborhood propensity score and the outcome.

1) We derive a subclassification on the individual propensity score $\phi(1; \mathbf{X}_i^z)$ as follows:

a) We first assume a Bernoulli distribution for the individual treatment,

$$Z_i \sim Ber(\phi(1; \mathbf{X}_i^z)) \tag{34}$$

with

$$\text{logit}(\phi(1; \mathbf{X}_i^z)) = \alpha_0 + \boldsymbol{\alpha_x^T} \mathbf{X}_i^z \tag{35}$$

Parameters $\boldsymbol{\alpha} = [\alpha_0, \boldsymbol{\alpha_x^T}]$ are estimated using a logistic regression.

b) Given estimated parameters $\widehat{\boldsymbol{\alpha}}$, for each unit we compute the individual propensity score $\phi(1; \mathbf{X}_i^z)$.

c) We identify J subclasses B_j, with $j = 1, \ldots, J$, defined by similar values of $\phi(1; \mathbf{X}_i^z)$ and where there is sufficient balance between individual treatment groups, i.e., $\mathbf{X}_i^z \perp\!\!\!\perp Z_i | i \in B_j$. In the simulation study we choose 5 subclasses based on

16

quintiles. For recommendations on the choice of the number of subclasses and the boundaries see Imbens & Rubin (2015)

2) Within each subclass B_j, we repeat the following steps to estimate $\mu_j(z,g) = E\big[Y_i(z,g) | i \in B_j^g\big]$, where $B_j^g = V_g \cap B_j$:

a) We posit a model for the neighborhood treatment assuming a distribution that depends on its specific definition. In scenarios 1,2, and 4, where G_i is defined as the proportion of treated neighbors, we posit the following model:

$$G_i N_i \sim Bin(N_i, \pi(z, \mathbf{X}_i^g)) \tag{36}$$

with

$$\text{logit}(\pi(z, \mathbf{X}_i^g)) = \gamma_0 + \gamma_z z + \boldsymbol{\gamma_x^T} \mathbf{X}_i^g \tag{37}$$

Parameters $\boldsymbol{\gamma} = [\gamma_0, \gamma_z, \boldsymbol{\gamma_x^T}]$ are estimated using a binomial logistic regression. In scenarios 3, where G_i is defined as the number of treated neighbors, we simply have:

$$G_i \sim Bin(N_i, \pi(z, \mathbf{X}_i^g)) \tag{38}$$

b) Given estimated parameters $\widehat{\boldsymbol{\gamma}}$, for each unit we compute $\widehat{\Lambda}_i = \lambda(G_i; Z_i; \mathbf{X}_i^g)$

c) We use the observed data $(Y_i, Z_i, G_i, \mathbf{X}_i^g)$ and $\widehat{\Lambda}_i = \lambda(G_i; Z_i; \mathbf{X}_i^g)$ to estimate the parameters of a model $Y_i(z,g) \mid \lambda(z; g; \mathbf{X}_i^g) \sim f^Y(z, g, \lambda(g; z; \mathbf{X}_i^g))$. In particular, in the simulation study we fit the following linear regression:

$$Y_i(z,g) = \beta_0 + \beta_z z + \beta_g g + \beta_{zg} zg + \beta_\lambda \widehat{\lambda}_i(g; z; \mathbf{X}_i^g) + \beta_{g\lambda} g\widehat{\lambda}_i(g; z; \mathbf{X}_i^g) \tag{39}$$

Alternative models could be used. In a more realistic scenario we recommend using the cubic polynomial function proposed by Hirano & Imbens (2004).

d) For a particular level of the joint treatment $(Z_i = z, G_i = g)$, for each unit $i \in B_j^g$ we predict the neighborhood propensity score evaluated at that level of the treatment, i.e., $\lambda(g; z; \mathbf{X}_i^g)$, and use it to predict the potential outcome $Y_i(z,g)$.

e) To estimate the dose-response function $\mu_j(z,g)$ we average the potential outcomes over $\lambda(z; g; \mathbf{X}_i^g)$

$$\widehat{\mu}_j(z,g) = \frac{\sum_{i \in B_j^g} \widehat{Y}_i(z,g)}{|B_j^g|}$$

3) We derive the average dose-response function as follows:

$$\widehat{\mu}(z, g) = \sum_{j=1}^{J} \widehat{\mu}_j(z, g)\pi_j^g$$

where $\pi_j^g = \frac{|B_j^g|}{v_g}$.

C.2   Statistical Inference and Interval Estimation

In this section, we focus on the derivation of standard errors and confidence intervals for our proposed propensity score estimator. We propose the use of a bootstrap method with an independent resampling scheme with replacement. After computing the neighborhood treatment G_i and the neighborhood covariates for each unit, these are considered as node attributes. the bootstrap procedure resamples units independently.

In Table 18 we present the results of using this estimator to perform inference about main and spillover effects. For every scenario, we simulated one dataset and applying the procedure described in Section 5 to derive point estimates, standard errors, and confidence intervals. Standard errors are the result of 1000 bootstrap replications. Resampling is performed at unit-level. Also shown are bootstrapped 95% confidence intervals based on normal approximation.

It is worth noting that standard errors are consistent with root mean square errors obtained through Monte Carlo simulations, reported in Tables 2, 3 and 4.

The variance of our estimator is assessed with respect to the distribution induced by random sampling using bootstrap methods. The properties of this variance estimator do depend on the sampling mechanism. In particular, the bootstrap resampling mechanism should match the one that was actually used to derive the sample. We have proposed a resampling scheme that matches an egocentric sampling mechanism (Kolaczyk , 2009; Perri et al., 2018), in that we take independent samples with replacement where sampled units carry with them both the neighborhood treatment G_i and neighborhood covariates, regardless of whether their neighbors have been resampled. This approach is valid when the observed sample was actually obtained by an egocentric sampling mechanism. In such a case, the properties of our variance estimator should be similar to the ones discussed in Imbens & Rubin (2015). In the simulation study, the performance of our estimator has been assessed under outcomes independence. The results of our simulations are promising. However, the

Table 18: Estimated Effects, Standard Errors and Confidence Intervals

Scenario	interference	$\widehat{\tau}$	SE	CI	$\widehat{\Delta(0)}$	SE	CI	$\widehat{\Delta(1)}$	SE	CI
1 $(Z_i \perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind})$	low	-13.174	0.095	[-13.356,-12.985]	-2.072	0.075	[-2.219, -1.923]	-0.576	0.045	[-0.663,-0.487]
	medium	-13.133	0.127	[-13.379,-12.878]	-3.572	0.078	[-3.729,-3.423]	-2.111	0.043	[-2.197,-2.026]
	high	-13.077	0.171	[-13.405,-12.731]	-4.519	0.074	[-4.664,-4.375]	-3.093	0.044	[-3.179,-3.005]
2 $(Z_i \perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind}, \mathbf{X}_i^{neigh})$	low	-13.083	0.056	[-13.192,-12.973]	-2.413	0.071	[-2.553, -2.273]	-0.815	0.046	[-0.904,-0.725]
	medium	-13.111	0.073	[-13.254,-12.968]	-4.097	0.072	[-4.239,-3.957]	-2.447	0.044	[-2.535,-2.360]
	high	-13.156	0.081	[-13.314,-12.997]	-5.138	0.082	[-5.303,-4.982]	-3.506	0.046	[-3.598,-3.412]
3 $(Z_i \perp\!\!\!\perp G_i \vert \mathbf{X}_i^{ind}, \texttt{degree}_i)$	low	-13.234	0.178	[-13.582,-12.885]	-1.140	0.100	[-1.333,-0.942]	0.581	0.033	[0.518,0.646]
	medium	-13.121	0.254	[-13.619,-12.623]	-1.979	0.097	[-2.180,-1.800]	-0.452	0.031	[-0.514,-0.392]
	high	-13.163	0.228	[-13.609,-12.716]	-3.661	0.111	[-3.885,-3.451]	-2.090	0.031	[-2.152,-2.029]
4 $(Z_i \not\perp\!\!\!\perp G_i \vert \mathbf{X}_i^{z}, \mathbf{X}_i^{neigh})$	low	-13.085	0.073	[-13.229,-12.943]	-1.695	0.109	[-1.922,-1.493]	-1.289	0.089	[-1.462,-1.114]
	medium	-13.061	0.071	[-13.201,-12.924]	-3.913	0.117	[-4.151,-3.694]	-3.273	0.092	[-3.456,-3.095]
	high	-13.119	0.079	[-13.271,-12.959]	-5.231	0.120	[-5.479,-5.008]	-4.761	0.088	[-4.934,-4.588]

sampling mechanism is not independent our solution is not ideal and further investigation is needed. In the case of clustered sampling, which sample separated clusters of data where there are no links between clusters, a clustered bootstrap could be used. However, when individuals within clusters are organized in networks, smaller clusters with higher correlation can also be defined using a community detection algorithm. We will further investigate these issues in future work.

D Proofs

D.1 Proof of Theorem 1

We want to prove that the adjusted observed mean $\overline{Y}_{z,g}^{obs}$ is equal to the quantity $\mu(z,g)$.

$$\overline{Y}_{z,g}^{obs} = E[E\left[\,Y_i \vert Z_i = z, G_i = g, \mathbf{X}_i = \mathbf{x}, i \in V_g\right] \vert\, i \in V_g]$$
$$= E[E\left[\,Y_i(z,g) \vert Z_i = z, G_i = g, \mathbf{X}_i = \mathbf{x}, i \in V_g\right] \vert\, i \in V_g]$$
$$= E[E\left[\,Y_i(z,g) \vert \mathbf{X}_i = \mathbf{x}, i \in V_g\right] \vert\, i \in V_g] = E\left[Y_i(z,g) \vert\, i \in V_g\right] = \mu(z,g)$$

where the fist equality is the definition of the adjusted observed mean, expressed in terms of conditional mean, the second equality holds by both Assumption 1 (No Multiple Versions of Treatment) and Assumption 2 (Neighborhood Interference), and the third equality holds

by the Assumption 3 (Unconfoundedness).

D.2   Proof of Theorem 2.A

$$\tau_{X^\star}^{obs} = \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

$$= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right.$$

$$\left. - E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

by iterated equations and by definition of V_g leading to $P(G_i = g | Z_i = z, \mathbf{X}_i^\star = \mathbf{x}, i \notin V_g) = 0$

$$= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i(1, g) | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right.$$

$$\left. - E[Y_i(0, g) | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

by consistency

$$= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i(1, g) | \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right.$$

$$\left. - E[Y_i(0, g) | \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

by the unconfoundedness assumption.

D.3 Proof of Corollary 2.A.1

$$\tau_{X^\star}^{obs} - \tau = \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

$$- \left\{ \sum_{g \in \mathcal{G}} \left(E[Y_i(1, g) - Y_i(0, g)| i \in V_g] \right) P(G_i = g) \right\}$$

$$= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right. \right.$$

$$\left. \left. - E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

$$- \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \left(E[Y_i(1, g) - Y_i(0, g) | \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

by iterated equations

$$= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right. \right.$$

$$\left. \left. - E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

$$- \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \left(E[Y_i(1, g) | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] - E[Y_i(0, g) Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right) \right.$$

$$\left. P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

by the unconfoundedness assumption

$$= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right. \right.$$

$$\left. \left. - E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

$$- \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \left(E[Y_i | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] - E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right) \right.$$

$$\left. P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} = 0$$

by consistency and by the assumption of independence between Z_i and G_i.

D.4 Proof of Corollary 2.A.2

We continue the previous proof without the assumption of independence between Z_i and G_i conditional on $\mathbf{X}^\star$. Then

$$\tau_{X^\star}^{obs} - \tau = \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] - E[Y_i | Z_i = 1, G_i = g', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right)$$

$$\left(P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) - P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

$$- \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} E[Y_i | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] - E[Y_i | Z_i = 0, G_i = g', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right)$$

$$\left(P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) - P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

since $E[Y_i | Z_i = 1, G_i = g', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]$ and $E[Y_i | Z_i = 0, G_i = g', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]$ do not depend on g. If there is no interaction between the individual treatment Z_i and the neighborhood treatment G_i, the simplified expression of the bias formula is straightforward.

D.5 Proof of Theorem 2.B

$$\tau_{X^\star}^{obs} - \tau = \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

$$- \left\{ \sum_{g \in \mathcal{G}} \left(E[Y_i(1, g) - Y_i(0, g) | i \in V_g] \right) P(G_i = g) \right\}$$

$$= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right. \right.$$

$$P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x})$$

$$- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]$$

$$\left. P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

$$- \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} \left(E[Y_i(1, g) | \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] - E[Y_i(0, g) | \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right) \right.$$

$$\left. P(\mathbf{U}_i = \mathbf{u} | G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

by iterated equations and by definition of V_g

$$
= \Bigg\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \Bigg(\sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \Bigg) P(\mathbf{X}_i^\star = \mathbf{x}) \Bigg\}
$$

$$
- \Bigg\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} \Bigg(E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \Bigg)
$$

$$
P(\mathbf{U}_i = \mathbf{u} | G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \Bigg\}
$$

by the unconfoundedness assumption conditional on $\mathbf{X}_i^\star$ and $\mathbf{U}_i$ and by consistency

$$
= \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \Bigg(E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = 1, G_i = g', \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \Bigg)
$$

$$
\Bigg(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
- P(\mathbf{U}_i = \mathbf{u} | G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) \Bigg) P(\mathbf{X}_i^\star = \mathbf{x})
$$

$$
- \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \Bigg(E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = 0, G_i = g', \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \Bigg)
$$

$$
\Bigg(P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
- P(\mathbf{U}_i = \mathbf{u} | G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) \Bigg) P(\mathbf{X}_i^\star = \mathbf{x})
$$

by adding and subtracting two quantities $E[Y_i | Z_i = 1, G_i = g', \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]$ and $E[Y_i | Z_i = 0, G_i = g', \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]$ that do not depend on g and $\mathbf{u}$. Now if there is no interaction between the individual treatment Z_i and the neighborhood treatment G_i,

the bias formula formula reduces to

$$
= \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \Bigg(E[Y_i | Z_i = z, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = z, G_i = g', \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \Bigg)
$$

$$
\Bigg(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
- P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \Bigg) P(\mathbf{X}_i = \mathbf{x})
$$

D.6   Proof of Corollary 2.B.1

Following the expression of the bias in the proof Theorem 2.B,

$$
\tau_{X^\star}^{obs} - \tau = \Bigg\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \Bigg(\sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x})
$$

$$
- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \Bigg) P(\mathbf{X}_i^\star = \mathbf{x}) \Bigg\}
$$

$$
- \Bigg\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} \Bigg(E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \Bigg)
$$

$$
P(\mathbf{U}_i = \mathbf{u} | G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \Bigg\}
$$

which results from iterated equations, the unconfoundedness assumption conditional on $\mathbf{X}_i^\star$ and $\mathbf{U}_i$ and consistency, the independence between Z_i and G_i given $\mathbf{X}_i^\star$ leads to

$$
= \Bigg\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \Bigg(\sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x})
$$

$$
- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g]
$$

$$
P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) \Bigg) P(\mathbf{X}_i^\star = \mathbf{x}) \Bigg\}
$$

$$- \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{g \in \mathcal{G}} \sum_{\mathbf{u} \in \mathcal{U}} \left(E[Y_i | Z_i = 1, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \right. \right.$$

$$- E[Y_i | Z_i = 0, G_i = g, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}, i \in V_g] \Big)$$

$$\left. P(\mathbf{U}_i = \mathbf{u} | G_i = g, \mathbf{X}_i^\star = \mathbf{x}) P(G_i = g | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

where we can just marginalize over the distribution of G_i, yielding

$$= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 1, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) \right. \right.$$

$$\left. - E[Y_i | Z_i = 0, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x}) \Big\}$$

$$- \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \sum_{\mathbf{u} \in \mathcal{U}} \left(E[Y_i | Z_i = 1, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = 0, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] \right) \right.$$

$$\left. P(\mathbf{U}_i = \mathbf{u} | \mathbf{X}_i^\star = \mathbf{x}) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}$$

Now we can just proceed as we would do to derive the bias given by unmeasured confounders. By adding and subtracting two quantities $E[Y_i | Z_i = 1, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}]$ and $E[Y_i | Z_i = 0, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}]$ we get

$$= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 1, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = 1, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}] \right)$$

$$\left(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) - P(\mathbf{U}_i = \mathbf{u} | \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

$$- \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = 0, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = 0, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}] \right)$$

$$\left(P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) - P(\mathbf{U}_i = \mathbf{u} | \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

If there is no interaction between the individual treatment Z_i and $\mathbf{U}_i$, we can have a simplified expression of the bias formula:

$$= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = z, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = z, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}] \right)$$

$$\left(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) - P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

D.7 Proof of Corollary 2.B.2

$$
\begin{aligned}
\tau_{X^\star}^{obs} - \tau &= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i|Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} \\
&\quad - \left\{ \sum_{g \in \mathcal{G}} \left(E[Y_i(1, g)Y_i(0, g)|i \in V_g] \right) P(G_i = g) \right\} \\
&= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i|Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} \\
&\quad - \left(E[Y_i(1)] - E[Y_i(0)] \right) \\
&= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i|Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} \\
&\quad - \left(E[Y_i(1)|Z_i = 1] - E[Y_i(0)Z_i = 0] \right)
\end{aligned}
$$

by SUTVA and by definition of V_g, which results in $P(G_i = g) = P(G_i = g| \in V_g)P(i \in V_g)$. Now we can just proceed as we would do to derive the bias given by unmeasured confounders. By iterated equations and applying the unconfoudedness assumption we get

$$
\begin{aligned}
&= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i|Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} \\
&\quad - \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i(1)Z_i = 1] - E[Y_i(0)|Z_i = 0] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} \\
&= \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i|Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\} \\
&\quad - \left\{ \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] \right) P(\mathbf{X}_i^\star = \mathbf{x}) \right\}
\end{aligned}
$$

By adding and subtracting two quantities $E[Y_i|Z_i = 1, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}]$ and $E[Y_i|Z_i = 0, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}]$ we get

$$
\begin{aligned}
&= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i|Z_i = 1, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 1, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}] \right) \\
&\qquad \left(P(\mathbf{U}_i = \mathbf{u}|Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) - P(\mathbf{U}_i = \mathbf{u}|\mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x}) \\
&\quad - \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i|Z_i = 0, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i|Z_i = 0, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}] \right) \\
&\qquad \left(P(\mathbf{U}_i = \mathbf{u}|Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) - P(\mathbf{U}_i = \mathbf{u}|\mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})
\end{aligned}
$$

26

If there is no interaction between the individual treatment Z_i and $\mathbf{U}_i$, we can have a simplified expression of the bias formula:

$$= \sum_{\mathbf{x} \in \mathcal{X}^\star} \left(\sum_{\mathbf{u} \in \mathcal{U}} E[Y_i | Z_i = z, \mathbf{U}_i = \mathbf{u}, \mathbf{X}_i^\star = \mathbf{x}] - E[Y_i | Z_i = z, \mathbf{U}_i = \mathbf{u}', \mathbf{X}_i^\star = \mathbf{x}] \right)$$
$$\left(P(\mathbf{U}_i = \mathbf{u} | Z_i = 1, \mathbf{X}_i^\star = \mathbf{x}) - P(\mathbf{U}_i = \mathbf{u} | Z_i = 0, \mathbf{X}_i^\star = \mathbf{x}) \right) P(\mathbf{X}_i^\star = \mathbf{x})$$

D.8 Proof of Proposition 3

We have to show that $P(Z_i = z, G_i = g | \mathbf{X}_i, \psi(z; g; \mathbf{X}_i)) = P(Z_i = z, G_i = g | \psi(z; g; \mathbf{X}_i))$. First consider the left hand side:

$$P(Z_i = z, G_i = g | \mathbf{X}_i, \psi(z; g; \mathbf{X}_i)) = P(Z_i = z, G_i = g | \mathbf{X}_i) = \psi(z; g; \mathbf{X}_i)$$

where the first equality follows because the joint propensity score is a function of $\mathbf{X}_i$ and the second is by the definition of the joint propensity score. Second, consider the right hand side. By iterated equations we have that

$$P(Z_i = z, G_i = g | \psi(z; g; \mathbf{X}_i)) = E_{\mathbf{X}}[P(Z_i = z, G_i = g | \mathbf{X}_i, \psi(z; g; \mathbf{X}_i)) | \psi(z; g; \mathbf{X}_i)]$$
$$= E_{\mathbf{X}}[P(Z_i = z, G_i = g | \mathbf{X}_i) | \psi(z; g; \mathbf{X}_i)]$$
$$= E_{\mathbf{X}}[\psi(z; g; \mathbf{X}_i) | \psi(z; g; \mathbf{X}_i)] = \psi(z; g; \mathbf{X}_i)$$

where the second equality is also because the joint propensity score is a function of $\mathbf{X}_i$, the third equality is by the definition of the joint propensity score and the fourth is equality is just trivial.

D.9 Proof of Proposition 4

We have to show that if Assumption 3 holds then $P(Z_i = z, G_i = g | Y_i(z, g), \psi(z; g; \mathbf{X}_i)) = P(Z_i = z, G_i = g | \psi(z; g; \mathbf{X}_i))$. The proof proceeds by showing that both the left and the right hand sides of this equation are equal to the joint propensity score itself and, hence, they are also equal to each other. In doing so we will make use of the assumption of unconfoundedness 3. Notice that in the proof of the Balancing Property we have already shown that the right hand side of the equation is equal to the propensity score, i.e., $P(Z_i = z, G_i = g | \psi(z; g; \mathbf{X}_i)) = \psi(z; g; \mathbf{X}_i)$. Now to prove that $P(Z_i = z, G_i =$

$g|Y_i(z,g), \psi(z;g;\mathbf{X}_i)) = \psi(z;g;\mathbf{X}_i)$ we use the law of iterated equations:

$$
\begin{aligned}
P(Z_i = z, G_i = g|Y_i(z,g), \psi(z;g;\mathbf{X}_i)) &= E_{\mathbf{X}}[P(Z_i = z, G_i = g|\mathbf{X}_i, Y_i(z,g), \psi(z;g;\mathbf{X}_i))|Y_i(z,g), \psi(z;g;\mathbf{X}_i)] \\
&= E_{\mathbf{X}}[P(Z_i = z, G_i = g|Y_i(z,g), \mathbf{X}_i)|Y_i(z,g), \psi(z;g;\mathbf{X}_i)] \\
&= E_{\mathbf{X}}[P(Z_i = z, G_i = g|\mathbf{X}_i)|Y_i(z,g), \psi(z;g;\mathbf{X}_i)] \\
&= E_{\mathbf{X}}[\psi(z;g;\mathbf{X}_i)|Y_i(z,g), \psi(z;g;\mathbf{X}_i)] = \psi(z;g;\mathbf{X}_i)
\end{aligned}
$$

where the second equality holds because $\psi(z;g;\mathbf{X}_i)$ is a function of $\mathbf{X}_i$ and the third follows from assumption 3.

D.10 Proof of Proposition 5

We have to show that if Assumption 3 holds given $\mathbf{X}_i$ then $P(Z_i = z, G_i = g|Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)) = P(Z_i = z, G_i = g|\lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z))$. First, by definition of $\mathbf{X}_i^z$ and $\mathbf{X}_i^g$ we can write $\phi(1;\mathbf{X}_i^z) = P(Z_i = 1|\mathbf{X}_i) = P(Z_i = 1|\mathbf{X}_i^z)$ and $\lambda(z;g;\mathbf{X}_i^g) = P(G_i = g|Z_i = z, \mathbf{X}_i) = P(G_i = g|Z_i = z, \mathbf{X}_i^g)$. Now, the proof proceeds by showing that both the left and the right hand sides of this equation are equal to the joint propensity score itself and, hence, they are also equal to each other. Let us consider the right hand side of the equation. By iterated equations we have that

$$
\begin{aligned}
P(Z_i = z, G_i = g|\lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)) & \\
&= E_{\mathbf{X}}[P(Z_i = z, G_i = g|\mathbf{X}_i, \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z))|\lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)] \\
&= E_{\mathbf{X}}[P(Z_i = z, G_i = g|\mathbf{X}_i)|\lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)] \\
&= E_{\mathbf{X}}[\psi(z;g;\mathbf{X}_i)|\lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)] = \psi(z;g;\mathbf{X}_i)
\end{aligned}
$$

where we have used the result that both $\lambda(z;g;\mathbf{X}_i^g)$ and $\phi(1;\mathbf{X}_i^z)$ can be seen as functions of $\mathbf{X}_i$ and the second equality is given Assumption 3. The third equality is straightforward, given that by factorization $\psi(z;g;\mathbf{X}_i) = \lambda(z;g;\mathbf{X}_i^g)\phi(z;\mathbf{X}_i^z)$ and that the value of the propensity score $\phi(z;\mathbf{X}_i^z)$ for the binary treatment Z_i is defined given $\phi(1;\mathbf{X}_i^z) = 1-\phi(0;\mathbf{X}_i^z)$. Finally, to prove that the left hand side $P(Z_i = z, G_i = g|Y_i(z,g), \psi(z;g;\mathbf{X}_i)) = \psi(z;g;\mathbf{X}_i)$

we use the law of iterated equations:

$$P(Z_i = z, G_i = g | Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z))$$

$$= E_{\mathbf{X}}[P(Z_i = z, G_i = g | \mathbf{X}_i, Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)) | Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)]$$

$$= E_{\mathbf{X}}[P(Z_i = z, G_i = g | Y_i(z,g), \mathbf{X}_i) | Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)]$$

$$= E_{\mathbf{X}}[P(Z_i = z, G_i = g | \mathbf{X}_i) | Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)]$$

$$= E_{\mathbf{X}}[\psi(z;g;\mathbf{X}_i) | Y_i(z,g), \lambda(z;g;\mathbf{X}_i^g), \phi(1;\mathbf{X}_i^z)] = \psi(z;g;\mathbf{X}_i)$$

where the second equality holds because $\lambda(z;g;\mathbf{X}_i^g)$ and $\phi(1;\mathbf{X}_i^z)$ are functions of $\mathbf{X}_i$, the third follows from assumption 3, and the last equality is given by the same argument as above.