Estimating treatment effects from observational network data

STAT 548 Qualifying Paper

Kenny Chiu

January 16, 2022

Abstract. Forastiere et al. (2021) examine the problem of estimating causal treatment effects from observational network data in the presence of interference. We review their proposed method and discuss its place in the literature. We present simple examples that illustrate the relevance of their theoretical results. We replicate part of their simulation study on a different dataset (the Twitch Social Networks dataset), obtain findings mostly consistent with theirs, and discuss the possible reasons for the few inconsistent results. We also provide a critique of their work.

1 Introduction

Forastiere et al. (2021) examine the problem of estimating causal treatment effects from observational network data in the presence of interference. The problem is formulated under a potential outcome framework in which they propose a propensity score-based estimator for the treatment effect that is unbiased under certain assumptions. Forastiere et al. also derive the biases of naive estimators that ignore interference under their framework. In this report, we review the work by Forastiere et al. (2021), present simple examples that illustrate their theoretical results, and replicate part of their simulation study on a different dataset.

This report is organized as follows: Section 2 introduces the notation used in this report; Section 3 summarizes their proposed methodology in the context of the literature; Section 4 presents examples that illustrate the relevance of their theoretical results; Sections 5 and 6 discusses our attempts at replicating part of their simulation study on a different dataset; and Section 7 concludes the report with a critique of their work. Appendix A includes our technical derivations from Section 4, and Appendix B contains additional information about the dataset and our simulations from Sections 5 and 6.

2 Notation

Throughout this report, we closely follow the notation used by Forastiere et al. (2021). Let $G = (\mathcal{N}, \mathcal{E})$ be an undirected network where $\mathcal{N}$ is a set of N units (nodes) and $\mathcal{E} = \{\{i, j\} : i, j \in \mathcal{N}\}$ is a set of edges. For a unit $i \in \mathcal{N}$, a partition $(i, \mathcal{N}_i, \mathcal{N}_{-i})$ of $\mathcal{N}$ describes unit i's neighbourhood $\mathcal{N}_i$ (the set of N_i units connected to unit i) and the set $\mathcal{N}_{-i}$ of all other units that are not i and that are not in $\mathcal{N}_i$. Let $Z_i \in \{0, 1\}$ be the treatment assigned to unit i and $Y_i \in \mathcal{Y}$ the observed outcome of unit i. Denote the treatment vector for the population $\mathcal{N}$ as $\mathbf{Z}$ and the corresponding vectors for partition $(i, \mathcal{N}_i, \mathcal{N}_{-i})$ as $(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}})$. Let $G_i = g_i(\mathbf{Z}_{\mathcal{N}_i}) \in \mathcal{G}_i$ be some known and well-specified summary g_i of the treatments in unit i's neighbourhood. Depending on the size of a unit's neighbourhood, the space $\mathcal{G}_i$ may differ between units. Let $V_g = \{i \in \mathcal{N} : g \in \mathcal{G}_i\}$ denote the subset containing v_g units that have g as a possible

value for the neighbourhood treatment. Let $\mathbf{X}_i \in \mathcal{X}$ be a vector of covariates for unit i that partitions into individual-level characteristics $\mathbf{X}_i^{\text{ind}} \in \mathcal{X}^{\text{ind}}$ and neighbourhood-level/aggregated individual-level characteristics $\mathbf{X}_i^{\text{neigh}} \in \mathcal{X}^{\text{neigh}}$.

We use $\mathbb{E}[\bullet]$ and $\mathbb{P}(\bullet)$ to denote expectations and probabilities, respectively. In the context of an observed network, these quantities should be viewed from the "super-population" perspective (Imbens & Rubin, 2015) and are defined in terms of averages over the finite network. Otherwise, expectations and probabilities are interpreted the usual way. We leave it up to the context to distinguish which interpretation is used throughout this report.

3 Review of proposed methodology

In this section, we explain the method proposed by Forastiere et al. (2021) and discuss its relevance in the context of the related literature. We also discuss its advantages over other methods used in similar contexts.

3.1 Setting, objective and method

Forastiere et al. (2021) focus on the problem of estimating treatment effects from observational network data under interference. The setting is challenging for causal inference because

1. the assignment mechanism of treatments is unknown with observational data and so estimated effects may be biased in the presence of unmeasured confounders, and
2. conventional inference methods typically ignore the effect of interference, which may also lead to biased estimates.

In the randomized controlled trial literature, these issues can generally be dealt with by designing the study in such a way that the influence of confounders is minimized and that inference methods that do account for interference can be used (e.g., Saveski et al., 2017; Doudchenko et al., 2020; Jagadeesan et al., 2020; Imai et al., 2021). In the observational setting, these considerations need to be addressed by the inference method in the analysis phase. Forastiere et al. specifically focus on the setting involving a binary treatment (e.g., an intervention and a control) and where the interference on a unit comes from only its treated neighbours. They formulate the inference problem under a potential outcome framework, and they propose an estimation procedure that yields unbiased treatment and spillover (interference) effect estimates under certain assumptions.

The procedure proposed by Forastiere et al. (2021) involves partitioning units into subclasses according to a joint propensity score $\psi(z; g; x)$ that they also define. The joint propensity score factorizes into a neighbourhood propensity score $\lambda(g; z; \mathbf{x}^g)$ (probability of being exposed to neighbourhood treatment g given individual treatment z and relevant covariates $\mathbf{x}^g$) and an individual propensity score $\phi(z; \mathbf{x}^z)$ (probability of being assigned treatment z given relevant covariates $\mathbf{x}^z$), which they exploit in order to break down the problem into subproblems with lower-dimensional spaces that are easier to work with. Note that $\mathbf{X}^g$ and $\mathbf{X}^z$ do not necessarily correspond to $\mathbf{X}^{\text{neigh}}$ and $\mathbf{X}^{\text{ind}}$, respectively, as they may not be disjoint. The steps for their propensity score-based estimation procedure are as follows:

1. **Subclassify units**.
   (a) Fit a logistic regression model on the individual treatments Z_i given covariates $\mathbf{X}_i^z$, and use the model to predict the individual propensity scores $\phi(1; \mathbf{X}_i^z)$.
   (b) Partition the units into J subclasses B_j, $j \in \{1, \ldots, J\}$, based on similar estimated individual propensity scores $\hat{\phi}(1; \mathbf{X}_i^z)$ and such that each subclass is approximately balanced in the number of treated and untreated units.
2. **Estimate potential outcomes**. Let $B_j^g = V_g \cap B_j$. For each subclass B_j:

(a) Fit some regression model on the neighbourhood treatments G_i given the individual treatments Z_i and covariates $\mathbf{X}_i^g$, and use the model to estimate the neighbourhood propensity scores $\lambda(g; z; \mathbf{X}_i^g)$.

(b) Fit some regression model on the potential outcomes $Y_i(z, g)$ given the individual and neighbourhood treatments Z_i and G_i and the estimated neighbourhood propensity scores $\hat{\lambda}(g; z; \mathbf{X}_i^g)$.

(c) Estimate the dose-response function by averaging over the estimated potential outcomes for a particular level of the joint treatment, i.e.,

$$\hat{\mu}_j(z, g; V_g) = \frac{\sum_{i \in B_j^g} \hat{Y}_i(z, g)}{\left|B_j^g\right|} .$$

3. **Estimate the average dose-response function** (ADRF) $\mu(z, g; V_g) = \mathbb{E}\left[Y_i(z, g) | i \in V_g\right]$ for a particular level of the joint treatment by taking the weighted average of the estimated dose-response functions over the subclasses, i.e.,

$$\hat{\mu}(z, g; V_g) = \sum_{j=1}^{J} \hat{\mu}_j(z, g; V_g) \left(\frac{\left|B_j^g\right|}{v_g} \right) .$$

4. **Estimate** the treatment effects $\tau(g)$, overall treatment effect τ, spillover effects $\delta(g; z)$ and overall spillover effects $\Delta(z)$ using the estimated ADRF by

$$\hat{\tau}(g) = \hat{\mu}(1, g; V_g) - \hat{\mu}(0, g; V_g) , \qquad \hat{\tau} = \sum_{g \in \mathcal{G}} \hat{\tau}(g) \mathbb{P}(G_i = g) ,$$

$$\hat{\delta}(g; z) = \hat{\mu}(z, g; V_g) - \hat{\mu}(z, 0; V_g) , \qquad \hat{\Delta}(z) = \sum_{g \in \mathcal{G}} \hat{\delta}(g; z) \mathbb{P}(G_i = g) .$$

Forastiere et al. (2021) show that their estimators for the treatment and spillover effects are unbiased under three assumptions, the first two of which form the Stable Unit Treatment on Neighbourhood Value Assumption (SUTNVA, a generalization of SUTVA that relaxes the no interference assumption to allow interference of immediate neighbours) and the third being an unconfoundedness assumption that says the treatment assignment mechanism is conditionally independent of the outcomes for the given set of covariates.

3.2 Relevant literature

The body of literature that involves estimating causal treatment effects from observational data under general forms of interference is still relatively new. We summarize the common approaches in this literature and discuss how the work by Forastiere et al. (2021) fits in. We also briefly highlight other work that examine similar problems under different but closely related contexts.

3.2.1 Similar contexts

As noted by Forastiere et al. (2021), the majority of works that look at similar problems under the same context involve either inverse probability-weighted (IPW) estimators (Liu et al., 2016) or targeted maximum likelihood estimators (TMLE) (van der Laan, 2014; Ogburn et al., 2017; Sofrygin & van der Laan, 2017) for the causal treatment effect. The main advantage of the method proposed by Forastiere et al. over these two approaches is the weaker assumptions that it requires.

IPW estimators are weighted averages of the outcomes where the weights are defined with respect to a hypothetical allocation strategy (an assumed distribution over the neighbourhood treatments) and a known or correctly modelled generalized propensity score. The Bernoulli allocation strategy (Tchetgen & VanderWeele, 2012) is commonly used, which assumes that each unit in the neighbourhood is treated independently

with probability α and that a unit's assignment is independent of its neighbours' assignment. This assumption rules out homophily—the tendency for units with similar characteristics to form ties—which is a strong assumption that is said to be unrealistic (Shalizi & Thomas, 2011). In contrast, the estimators proposed by Forastiere et al. (2021) only use the observed neighbourhood treatments, and therefore no assumptions that explicitly rule out homophily are made (though the issue may still manifest as an unmeasured confounder if the unconfoundedness assumption does not hold for the given set of covariates). Both the IPW estimators and the estimators proposed by Forastiere et al. rely on being able to correctly model the joint propensity score.

TMLEs are obtained by maximizing the likelihood of the outcomes defined on a structural equation model. Similar to the IPW estimators, TMLE generally involves a randomization assumption (van der Laan, 2014) on the model where the conditional joint distribution of the treatment assignments factorizes into independent conditionals given the covariates for all units, and similarly for the conditional joint distribution of the outcomes given the covariates and the treatment assignments. In comparison, the unconfoundedness assumption in the method proposed by Forastiere et al. (2021) makes a weaker assumption where the outcome and treatment assignment of each unit is conditionally independent given only the covariates of that unit. The significance of these assumptions again circle back to the argument of disregarding homophily and/or other venues of confounding. It is notable that extensions of TMLE that allow for limited forms of homophily were later introduced (Ogburn et al., 2017).

More recently, approaches that can be described as extensions of the method by Forastiere et al. (2021) have been proposed. Jackson et al. (2020) proposed estimators based on propensity score matching but which also explicitly account for homophily by modeling neighbourhood treatment assignments as an incomplete information game. Sánchez-Becerra (2021) questioned the justification of the unconfoundedness assumption with respect to the constructed propensity score (which would need to be estimated for every unit in the network and may be challenging to estimate accurately) and proposed a two-step method where a network propensity score is estimated and then used to as inverse weights to match units.

3.2.2 Other contexts

We briefly highlight other works in the literature that look at similar problems under different but closely related contexts. The difference in settings makes it difficult to directly compare these approaches to that proposed by Forastiere et al. (2021).

A large body of the literature examine the inference problem under the assumption of partial interference where units are partitioned into groups with no spillover effects between groups. The focus on partial interference settings seems to be primarily due to momentum of earlier works (e.g., Sobel, 2006; Hudgens & Halloran, 2008) that looked at inference in randomized control studies with interference in which group-randomization tends to be more practical. Examples of recent work that assume partial interference include the work by Liu et al. (2019), Barkley et al. (2020), and Qu et al. (2021). IPW estimators are commonly used in the partial interference setting as they were originally introduced for grouped observational data (Tchetgen & VanderWeele, 2012).

A small number of works consider more specific and niche contexts. For example, Toulis et al. (2018) explore the problem of treatment entanglement (where treatment assignments are assumed to satisfy certain restrictions) and proposes a propensity score-based estimator. Zigler and Papadogeorgou (2021) focus on the problem of bipartite causal inference with interference (where the treatment is applied to one unit and the outcome is measured on another) and propose a IPW estimator for this setting.

4 Potential bias of naive estimator

In this section, we describe a simple example that illustrates the relevance of Theorem 2.A, Corollary 2 and Corollary 3 in the paper by Forastiere et al. (2021). Specifically, we show for a simple network that

1. if the unconfoundedness assumption holds but the individual and neighbourhood treatments are not conditionally independent for a given set of covariates, then an unbiased naive estimator that assumes SUTVA for the treatment effect will be biased due to interference effects, and
2. if the individual and neighbourhood treatments are conditionally independent but the unconfoundedness assumption does not hold for a given set of covariates, then an unbiased SUTVA estimator will be biased due to unmeasured confounders.

Consider some undirected network $G = (\mathcal{N}, \mathcal{E})$ where every unit is paired (has an edge) with exactly one other unit. For simplicity, we index a unit by its pair $i \in \mathbb{N}$ and its position $j \in \{1, 2\}$ within the pair. Denote $Z_{ij} \in \{0, 1\}$ as the treatment assignment of unit j in pair i. Let the neighbourhood treatment G_{ij} denote whether unit (ij)'s paired counterpart is treated. Therefore, $\mathcal{G}_{ij} = \{0, 1\}$ for all units in the network and $V_g = \mathcal{N}$ for all $g \in \{0, 1\}$. For convenience of notation, we drop the dependence on V_g where applicable. Let $\mathbf{X}_{ij} = X_{ij} \in \{0, 1\}$ be some measured covariate for each unit, and assume that the covariates of the units in the network are generated independently with $\mathbb{P}(X_{ij} = 1) = \mathbb{P}(X_{ij} = 0) = \frac{1}{2}$. Let $Y_{ij} \in \mathbb{R}$ be the outcome measured for each unit. Other details, such as the size of $\mathcal{N}$, are assumed but left unspecified due to being irrelevant for the discussion. Figure 1 shows an example network.

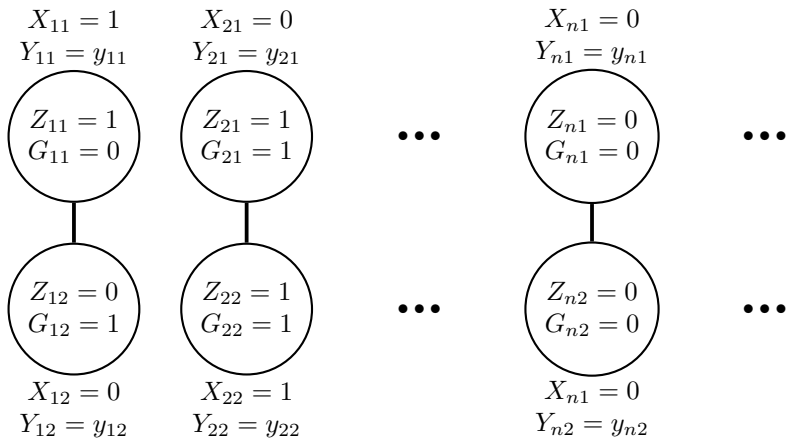

Figure 1: Example network of paired units with their observed individual treatment Z_{ij}, neighbourhood treatment G_{ij}, covariate X_{ij} and outcome Y_{ij}.

We examine the bias of a naive estimator on this network under two settings. In both settings, we assume that Assumption 1 (no multiple versions of treatment) and Assumption 2 (neighbourhood interference) stated by Forastiere et al. (2021) hold. Derivations of the quantities presented in this section can be found in Appendix A.

4.1 Setting 1: bias due to interference

In our first setting, suppose that the treatment assignment follows

$$\mathbb{P}(Z_{ij} = 1 | X_{i1}, X_{i2}) = \frac{3}{4}\mathbb{1}\left[X_{i1} = X_{i2}\right] + \frac{1}{4}\mathbb{1}\left[X_{i1} \neq X_{i2}\right]$$

where $\mathbb{1}[\bullet]$ is the indicator function, and suppose that $Y_{ij}(z,g) \sim N(c_z z + c_g g, \sigma^2)$ for some constants $c_z, c_g \in \mathbb{R}$. This treatment assignment corresponds to a homophily situation where the units in a pair are more likely to be treated if they share the same value of the covariate. Thus, Z_{ij} and G_{ij} are not conditionally independent given X_{ij} as the probability of a unit being treated also depends on its counterpart. On the other hand, the unconfoundedness assumption holds for the outcome $Y_{ij}(z,g)$ given X_{ij} (or any other covariates) as the outcome only depends on the observed values of the treatments z and g. Therefore, we have

$$Z_{ij} \not\perp\!\!\!\perp G_{ij} | X_{ij} \qquad \text{and} \qquad Y_{ij}(z,g) \perp\!\!\!\perp Z_{ij}, G_{ij} | X_{ij}$$

and so by Corollary 2, an effect estimator that is unbiased under SUTVA will be biased in the presence of neighbourhood interference. A hypothetical study that corresponds to this setting may be one where the paired units correspond to a pair of spouses with potentially different socioeconomic backgrounds X_{ij}, and it is of interest to determine whether being comfortable discussing financial matters (Z_{ij}) has an effect on some measure of their own financial management Y_{ij}. The homophily context arises from the assumption that spouses who come from similar backgrounds will find it easier to discuss finances. The interference context comes from the assumption that one may learn better management by hearing others talk about their experiences or strategies.

The overall treatment effect τ in this setting is given by

$$\begin{aligned}\tau &= \mathbb{E}\left[Y_{ij} | Z_{ij} = 1, G_{ij} = g\right] - \mathbb{E}\left[Y_{ij} | Z_{ij} = 0, G_{ij} = g\right] \\ &= c_z \ .\end{aligned}$$

By Theorem 2.A, the conventional covariate-adjusted estimator τ_X^{obs} that assumes SUTVA from the literature is well-defined under the potential outcome framework and estimates the quantity

$$\begin{aligned}\tau_X^{\text{obs}} = &\sum_{g,x,x' \in \{0,1\}} \left(\mathbb{E}\left[Y_{ij} | Z_{ij} = 1, G_{ij} = g\right] \mathbb{P}(G_{ij} = g | X_{ij} = x, X_{ik} = x') \mathbb{P}(Z_{ij} = 1 | X_{ij} = x, X_{ik} = x')\right. \\ &\left. - \mathbb{E}\left[Y_{ij} | Z_{ij} = 0, G_{ij} = g\right] \mathbb{P}(G_{ij} = g | X_{ij} = x, X_{ik} = x') \mathbb{P}(Z_{ij} = 0 | X_{ij} = x, X_{ik} = x')\right) \mathbb{P}(X_{ij} = x) \\ = &\ c_z + \frac{1}{4} c_g \ .\end{aligned}$$

It then follows that the bias of an unbiased estimator for τ_X^{obs} is $\tau_X^{\text{obs}} - \tau = \frac{1}{4} c_g$, i.e., the bias due to interference. Therefore, an unbiased SUTVA estimator is biased for the treatment effect when the individual and neighbourhood treatments are not conditionally independent for the given set of covariates in the presence of interference.

The derivation for the overall treatment effect τ, the estimator τ_X^{obs} and the bias $\tau_X^{\text{obs}} - \tau$ can be found in Appendix A.1.

4.2 Setting 2: bias due to unmeasured confounders

In our second setting, suppose that the treatment assignment mechanism follows

$$\mathbb{P}(Z_{ij} = 1 | X_{ij}) = \frac{1}{4} + \frac{1}{2} X_{ij} \ .$$

Further suppose that there is some unmeasured confounder $U_{ij} \in \{0, 1\}$ for each unit that is generated based on the individual treatment Z_{ij} according to the distribution

$$\mathbb{P}(U_{ij} = 1 | Z_{ij}) = \frac{1}{4} + \frac{1}{2} Z_{ij} \ ,$$

and suppose that $Y_{ij}(z,g) | U_{ij} \sim N(c_z z + c_u U_{ij}, \sigma^2)$ for some constants $c_z, c_u \in \mathbb{R}$. Because Z_{ij} only depends on X_{ij} in this setting, Z_{ij} and G_{ij} are conditionally independent given X_{ij}. However, because the outcome

$Y_{ij}(z,g)$ now depends on U_{ij}, the unconfoundedness assumption no longer holds given only X_{ij}. Therefore, we have

$$Z_{ij} \perp\!\!\!\perp G_{ij}|X_{ij} \qquad \text{and} \qquad Y_{ij}(z,g) \perp\!\!\!\perp Z_{ij}, G_{ij}|U_{ij}$$

and so by Corollary 3, an effect estimator that is unbiased under SUTVA will be biased due to the unmeasured confounder U_{ij}. A hypothetical study corresponding to this setting may be one similar to the example in setting one except where the paired units are friends rather than spouses, and so it is assumed that financial discussions are less intimate and therefore have minimal interference effects. In addition, not accounted for by the researcher, those who are comfortable discussing financial matters (Z_{ij}) are more likely to make financial investments (U_{ij}), which also effects their financial management Y_{ij}.

The overall treatment effect τ in this setting is given by

$$\begin{aligned}\tau &= \sum_{u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1, U_{ij}=u\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0, U_{ij}=u\right]\right)\mathbb{P}(U_{ij}=u) \\ &= c_z \ .\end{aligned}$$

The naive estimator τ_X^{obs} that assumes SUTVA estimates the quantity

$$\begin{aligned}\tau_X^{\text{obs}} &= \sum_{u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1, U_{ij}=u\right]\mathbb{P}(U_{ij}=u|Z_{ij}=1) - \mathbb{E}\left[Y_{ij}|Z_{ij}=0, U_{ij}=u\right]\mathbb{P}(U_{ij}=u|Z_{ij}=0)\right) \\ &= c_z + \frac{1}{2}c_u\end{aligned}$$

It then follows that the bias is $\tau_X^{\text{obs}} - \tau = \frac{1}{2}c_u$, i.e., the bias due to the unmeasured covariate. Therefore, an unbiased estimator of the SUTVA estimator is biased for the treatment effect when the unconfoundedness assumption does not hold for the given set of covariates.

The derivation for the overall treatment effect τ, the estimator τ_X^{obs} and the bias τ_X^{obs} can be found in Appendix A.2.

5 Reproducing simulation study findings

In this section, we aim to reproduce the general findings of the simulation study conducted by Forastiere et al. (2021). Note that the complete Add Health dataset that Forastiere et al. work with is publicly inaccessible, and so we instead work with the Twitch Social Networks dataset (Rozemberczki et al., 2021) that is comparable in size. Twitch is an American live streaming website that focuses on video game streaming. The full dataset contains several networks of streamers and their mutual friendship relationships that were collected in May 2018. We consider a hypothetical study where we are interested in understanding the effect of streamer self-promotions and advertising (the individual treatment) on the number of subscribers (users who follow a particular streamer; the outcome). It is reasonable to assume that there is interference at play where promoting one's self would inadvertently promote the streamer's network due to the site's recommendation algorithms that suggest similar streams to a viewer.

We consider only the EN subnetwork of the full Twitch dataset, which includes a sample of streamers who stream in English. The network consists of 7126 streamers and 35,324 mutual friendships. A total of 3169 binary features (e.g., games liked and played, location, streaming habits, etc.) are collected from each streamer. However, the individual features are unnamed. For the purposes of this study, we consider only feature `224` and feature `569` due to their distribution (the second and third most represented features in the dataset, respectively). We pretend these features are covariates representing whether a streamer plays a particular `game1` and `game2`. We note that one major difference between the Twitch dataset and the Add Health dataset is that the degree of each unit is limited to at most ten in the Add Health dataset while there is no

contextual limit in the Twitch dataset. More details about the Twitch dataset are provided in Appendix B.1.

In the following sections, we describe our efforts to translate the simulation procedure and (partial) findings of Tables 1 and 2 in the work by Forastiere et al. (2021) to our Twitch dataset. Under various individual and neighbourhood treatment generation scenarios, Table 1 compares the theoretical bias of estimators that adjust for different sets of covariates, while Table 2 compares the observed bias and RMSE of several estimators on simulated data. In our work, we focus specifically on Scenario 1 where the unconfoundedness assumption holds given the individual-level covariates.

5.1 Treatment and outcome generation models

To study the bias of different estimators, we simulate both the individual treatment Z and the outcome Y based on the individual covariates $\mathbf{X}^{\text{ind}} = (X^{\text{game1}}, X^{\text{game2}})$. Our treatment generation model is given by

$$\text{logit}(P(Z_i = 1)) = \text{logit}(\phi(1; \mathbf{X}_i^{\text{ind}})) = -3 + 3X_i^{\text{game1}} + 4X_i^{\text{game2}} \, .$$

Because Z_i is generated based on only $\mathbf{X}_i^{\text{ind}}$, it follows that the unconfoundedness assumption holds given $\mathbf{X}_i^{\text{ind}}$. For the neighbourhood treatment G_i in our simulations, we consider both the proportion of treated neighbours (following Forastiere et al. (2021)) and the number of treated neighbours (which may make more sense in our context as there should be a greater effect the more marketing there is). Additional details about the simulated treatments can be found in Appendix B.2.

Our outcome generation model is given by

$$\begin{aligned} Y_i(z, g)|\mathbf{X}_i^{\text{ind}} &\sim N\left(\mu(z, g, \mathbf{X}_i^{\text{ind}}), 1\right) \, , \\ \mu(z, g, \mathbf{X}_i^{\text{ind}}) &= 5 + 6\mathbb{1}[\phi(1; \mathbf{X}_i^{\text{ind}}) \geq 0.7] + 10z - 3z\mathbb{1}[\phi(1; \mathbf{X}_i^{\text{ind}}) \geq 0.7] + \gamma g \end{aligned}$$

where $\gamma \in \{4, 6, 8\}$ is the low, medium and high interference effect, respectively, for proportion neighbourhood treatment (for sum neighbourhood treatment, we take $\gamma \in \{0.4, 0.6, 0.8\}$). It follows that the treatment effect $\tau(g; \mathbf{X}^{\text{ind}})$ and overall treatment effect τ are then

$$\begin{aligned} \tau(g; \mathbf{X}_i^{\text{ind}}) &= \mu(1, g, \mathbf{X}_i^{\text{ind}}) - \mu(0, g, \mathbf{X}_i^{\text{ind}}) = 10 - 3\mathbb{1}[\phi(1; \mathbf{X}_i^{\text{ind}}) \geq 0.7] \, , \\ \tau &= \sum_{x \in \mathcal{X}^{\text{ind}}} \tau(g; x)\mathbb{P}(\mathbf{X}^{\text{ind}} = x) \, . \end{aligned}$$

5.2 Theoretical bias of estimators

We examine the theoretical bias of estimators that assume SUTVA and that adjust for differing sets of covariates. Following Forastiere et al. (2021), we consider the covariate sets $\mathbf{X}_i = \{\emptyset, \mathbf{X}_i^{\text{ind}}, \mathbf{X}_i^z = \mathbf{X}_i^{\text{ind}} \cup \mathbf{X}_i^{\text{neigh}}\}$ where

$$\mathbf{X}_i^{\text{neigh}} = \left(\text{Ngame1} = \frac{\sum_{k \in \mathcal{N}_i} \text{game1}_k}{N_i}, \text{Ngame2} = \frac{\sum_{k \in \mathcal{N}_i} \text{game2}_k}{N_i}, N_i\right) \, .$$

When no covariates are adjusted for ($\mathbf{X}_i = \emptyset$), we use Equation 12 from Theorem 2.B in the work by Forastiere et al. (2021) to compute the bias. For the other two covariate sets, we use Equation 11 from Corollary 2. Table 1 shows the computed biases for both proportion and sum neighbourhood treatments on one simulated dataset (the same dataset is used across estimators, and the dataset was regenerated for each interference level). Our results for proportion and sum neighbourhood treatments are similar. Our findings for the estimator that does not adjust for covariates and for the estimator that adjusts for $\mathbf{X}_i^{\text{ind}}$ are consistent with what is reported in Table 1 in the work of Forastiere et al. (2021).

For the estimator that adjusts for $\mathbf{X}_i^z$, we report biases that are larger than expected (the bias should be the same as the one that adjusts for $\mathbf{X}_i^{\text{ind}}$ in theory). Our investigation suggests that these larger values are

Interference (γ)	Bias($\emptyset$)	Bias($\mathbf{X}_i^{\text{ind}}$)	Bias($\mathbf{X}_i^z$)
Proportion G_i			
Low (4)	2.937	0.057	0.541
Medium (6)	3.009	0.085	0.812
High (8)	3.080	0.113	1.083
Sum G_i			
Low (0.4)	3.440	-0.144	0.736
Medium (0.6)	3.763	-0.217	1.104
High (0.8)	4.085	-0.289	1.472

Table 1: Bias of covariate-adjusted SUTVA estimators of τ when the unconfoundedness assumption holds given $\mathbf{X}_i^{\text{ind}}$. Each bias is computed from one simulated dataset.

arising due to poor estimates of the theoretical probabilities in Equation 11, which themselves are caused by severe imbalances in treated and untreated units when looking at a particular joint level of the covariates. Table 2 shows the ten joint levels—as well as the number of treated and untreated units with those levels—that contributed the largest positive and negative values to the bias of the estimator that adjusts for $\mathbf{X}_i^z$ for the dataset simulated with high proportion interference in Table 1. There are a total of 2384 distinct observed joint levels on this simulated dataset despite there being only 7126 units. Hence, for a particular level that heavily skews towards one treatment status, the other treatment status will generally be underrepresented or absent in this relatively small dataset. Forastiere et al. likely did not encounter this problem due to the restricted maximum degree (and therefore the limited possible values of G_i) in the Add Health data, which is also larger in size. It may also be the case that Forastiere et al. used the same computation for both the $\mathbf{X}_i^{\text{ind}}$ and $\mathbf{X}_i^z$-adjusted estimators as their reported biases are identical.

game1	game2	Ngame1	Ngame2	N_i	$n_{Z=1}$	$n_{Z=0}$	bias contrib.
Largest positive contributors							
1	1	0.5	1	2	47	0	0.0370
1	1	0.5	0.5	2	58	0	0.0309
0	0	0	1	1	6	222	0.0231
1	1	0.333	1	3	26	0	0.0228
1	1	0.667	0.667	3	27	0	0.0228
Largest (in magnitude) negative contributors							
0	0	0.5	1	2	0	21	-0.0202
0	0	1	0	2	0	26	-0.0168
0	0	0	1	3	0	20	-0.0157
0	0	1	1	1	9	130	-0.0149
0	0	1	0.5	2	1	37	-0.0110

Table 2: The ten joint levels of game1, game2, Ngame1, Ngame2 and N_i that contributed the largest (in magnitude) five positive and five negative values to the theoretical bias (Equation 11) of the $\mathbf{X}_i^z$-adjusted estimator for one dataset simulated with high proportion interference. The columns $n_{Z=1}$ and $n_{Z=0}$ show the number of treated and untreated units that have the joint level. The theoretical bias of 1.083 reported in Table 1 is obtained by summing the bias contribution column over all 2384 distinct joint levels.

5.3 Observed bias and RMSE of estimators

We examine the observed bias and RMSE of two estimators that assume SUTVA for the treatment effect τ on simulated datasets. The naive unadjusted estimator computes the simple contrast between treated and untreated units given by

$$\tau_{\text{naive}} = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$$

where $\bar{Y}_\bullet$ is the mean outcome across units with $\bullet$. The regression estimator τ_{reg} (Imbens & Rubin, 2015) is extracted from a fitted linear model given by

$$\mathbb{E}[Y|Z_i, \mathbf{X}_i^{\text{ind}}] = \beta_0 + \tau_{\text{reg}} Z_i + \beta_1 X_i^{\text{game1}} + \beta_2 X_i^{\text{game2}}$$

where β_j are the other parameters in the model. The observed bias is computed as the differences between estimates $\hat{\tau}_{\text{naive}}$ and $\hat{\tau}_{\text{reg}}$ and the expected value τ, which is computed using the formula provided in Section 5.1. Table 3 (estimators 1–2) reports the mean computed biases and RMSE for the two estimators over 500 simulated datasets. The individual and neighbourhood treatments are re-generated for each simulated dataset, and the same 500 simulated datasets are used for all estimators. The same generated treatments are also used across interference levels, though the outcomes are re-generated between levels.

Interference	τ_{naive}		$\tau_{\text{reg}} \sim Z_i, \mathbf{X}_i^{\text{ind}}$		$\tau_{\text{reg}} \sim Z_i, \mathbf{X}_i^{z}$		τ_{sub} $\hat{\phi}(1; \mathbf{X}_i^{\text{ind}})$		τ_{sub} $\hat{\phi}(1; \mathbf{X}_i^{z})$		τ_{GPS}	
(γ)	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Proportion G_i												
Low (4)	2.915	2.916	0.414	0.418	0.417	0.420	0.036	0.158	0.002	0.085	0.003	0.055
Med (6)	2.958	2.958	0.411	0.418	0.416	0.420	0.033	0.169	-0.001	0.112	0.000	0.059
High (8)	3.004	3.005	0.411	0.422	0.417	0.424	0.036	0.194	0.002	0.144	0.003	0.054
Sum G_i												
Low (0.4)	3.553	3.556	0.448	0.504	0.406	0.409	0.021	0.304	-0.310	0.584	-0.064	0.089
Med (0.6)	3.915	3.921	0.462	0.579	0.400	0.404	0.010	0.431	-0.469	0.877	-0.066	0.093
High (0.8)	4.280	4.289	0.479	0.664	0.395	0.402	0.006	0.563	-0.623	1.170	-0.062	0.088

Table 3: Mean bias and RMSE of estimators of τ when the unconfoundedness assumption holds given $\mathbf{X}_i^{\text{ind}}$ over 500 simulated datasets.

For both proportion and sum neighbourhood treatments, our findings are generally consistent with the results of Forastiere et al. (2021) reported in Table 2 (Unadjusted and Regression $\sim Z_i, \mathbf{X}_i^{\text{ind}}$ estimators under Scenario 1) where a greater interference effect leads to a greater RMSE. The bias of the regression estimator is smaller than that of the unadjusted estimator, though it is still non-trivial and is presumably due to the misspecification of the linear model.

6 Extending the simulation study

In this section, we extend our simulation study from Section 5.3 to the other estimators considered in Table 2 of the work by Forastiere et al. (2021). These estimators include

1. a second regression estimator $\tau_{\text{reg}}(Z_i, \mathbf{X}_i^{z})$ that adjusts for all covariates, i.e.,

$$\mathbb{E}[Y|Z_i, \mathbf{X}_i^{z}] = \beta_0 + \tau_{\text{reg}} Z_i + \beta_1 X_i^{\text{game1}} + \beta_2 X_i^{\text{game2}} + \beta_3 X_i^{\text{Ngame1}} + \beta_4 X_i^{\text{Ngame2}} + \beta_5 N_i \ ;$$

2. an estimator $\tau_{\text{sub}}(\mathbf{X}_i^{\text{ind}})$ that takes averages within subclasses (Imbens & Rubin, 2015), which are defined according to a propensity score that adjusts for individual covariates. The procedure to compute the estimate is as follows:

   (a) Fit a logistic regression model on the individual treatment Z_i with covariates $\mathbf{X}_i^{\text{ind}}$.

(b) Estimate the individual propensity score $\hat{\phi}(1; \mathbf{X}_i^{\text{ind}})$ for each unit.

(c) Partition the units into J roughly equal-sized subclasses based on $\hat{\phi}(1; \mathbf{X}_i^{\text{ind}})$. In our context, $J = 4$ is the maximum number of subclasses possible given the two binary covariates. We take $J = 4$ in our simulations.

(d) Let $\bar{Y}^{(j)}$ denote the average outcome over subclass B_j. The estimate is computed as

$$\hat{\tau}_{\text{sub}} = \frac{1}{N} \sum_{j=1}^{J} \left(\bar{Y}_{Z=1}^{(j)} - \bar{Y}_{Z=0}^{(j)} \right) |B_j| \; ;$$

3. a second subclassification estimator $\tau_{\text{sub}}(\mathbf{X}_i^z)$ where the subclasses are defined according to a propensity score that adjusts for all covariates. The estimation procedure is the same as the first subclassification estimator, except that $\mathbf{X}_i^z$ is used instead of $\mathbf{X}_i^{\text{ind}}$ and that we take $J = 5$ in our simulations following the recommendations of Rosenbaum and Rubin (1984) (also see our investigation in Appendix B.3); and

4. the estimator τ_{GPS} proposed by Forastiere et al. (2021) with the estimation procedure outline in Section 3.1. For proportion neighbourhood treatments, logistic regression models are used for both the individual and neighbourhood propensity scores (the latter in which N_i are used as weights). For sum neighbourhood treatments, a negative binomial model is used to estimate the neighbourhood propensity scores[1]. We again take $J = 5$ for the number of subclasses.

Note that where applicable, we take $V_g = \mathcal{N}$ in the case of proportion neighbourhood treatment for convenience of computation. Table 3 (estimators 3–6) shows the observed bias and RMSE of the above estimators. In the case of proportion neighbourhood treatments, our obtained biases and RMSEs are consistent with that reported by Forastiere et al. (2021). The regression estimators have relatively larger biases and RMSE due to misspecification. The subclassification estimators based on individual propensity scores have smaller biases that are comparable to the analytical bias of an unbiased SUTVA estimator that adjusts for $\mathbf{X}_i^{\text{ind}}$ reported in Table 1. The estimator proposed by Forastiere et al. based on individual and neighbourhood propensity scores achieves the smallest RMSE compared to the other estimators that ignore the effect of interference.

In the case of sum neighbourhood treatments, the results are similar to the proportion case except for the subclassification estimators based on individual propensity scores. The RMSE of these estimators are relatively large compared to the other estimators in this case. Note that the propensity score models and estimates are unchanged from the proportion neighbourhood treatment case, and so these larger RMSEs are a direct result of the sum interference effects which are unbounded. Specifically, the outlier units with relatively large degrees are likely the main contributors to the bias. It is also worth noting that the estimator proposed by Forastiere et al. still achieves the smallest RMSE out of all the compared estimators.

7 Critical appraisal and concluding remarks

We conclude this report with a critical appraisal of the paper and proposed method by Forastiere et al. (2021). The main contributions of the paper include a potential outcome formulation of the causal interference problem and a propensity score-based estimator for the treatment effect that adjusts for possible interference. The formulation is useful as it provides a framework under which properties (e.g., bias) of estimators that ignore interference can be derived and quantified. The proposed estimator appears to successfully achieve its purpose in adjusting for interference effects (at least when assumptions hold) based on our simulation results in Table 3. Note that due to the nature of the simulation study, our results were obtained with minimal

[1] In some instances, a particular subclass would end up with very few units and the model would fail to be fit. The units in this subclass would generally be the ones with the largest degrees in the network. In these cases, we took the estimated neighbourhood propensity scores to be 0.

tuning of the logistic and negative binomial models, and so we expect that this estimator would perform even better in practice with proper diagnostics and tuning of model fit.

One practical consideration that seems to be missing in the paper by Forastiere et al. (2021) is the computational performance of the procedure to compute the estimator. The complexity scales at least linearly with both the number of units and the number of distinct values of the neighbourhood treatment in the network. For the Add Health dataset that Forastiere et al. worked with, the latter was less of an issue as every unit was limited to having at most ten neighbours. For the Twitch dataset considered in this project, units could have many neighbours and so the number of distinct neighbourhood treatments was not trivial. With such datasets, it may be reasonable to expect that the max degree in the network is increases proportional to the total number of units, and so the computational complexity of the procedure is overall closer to quadratic in the number of units in the network. For large networks with millions of units, the procedure would likely be infeasible without further modifications.

As discussed by Forastiere et al. (2021), other limitations of their proposed method include its reliance on fully knowing the network structure and assuming it is fixed, and its dependence on models for the propensity scores that are likely to be misspecified. Sánchez-Becerra (2021) also questions the unconfoundedness assumption with respect to the propensity score. In practice, the unconfoundedness assumption holding with respect to a set of covariates may already be challenging to verify. Further assuming that unconfoundedness holds with respect to a (factorized) propensity score constructed on top of that set of covariates adds another layer of complexity that is difficult to justify.

In conclusion, Forastiere et al. (2021) present a formulation of the interference problem with observational data under which the properties of effect estimators can be derived. We discussed simple network examples that illustrate the relevance of their theoretical results where the bias of naive estimators can be assessed under their framework. When the assumed conditions hold, their proposed effect estimator shows promising results based on the results of our simulation study of the Twitch Social Networks dataset. The assumptions the estimator makes are also relatively weaker compared to that of other methods in the literature. However, there are practical concerns such as computational performance on large datasets that are not fully addressed in the original paper.

References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

Barkley, B. G., Hudgens, M. G., Clemens, J. D., Ali, M., & Emch, M. E. (2020). Causal inference from observational studies with clustered interference, with application to a cholera vaccine study. *The Annals of Applied Statistics*, *14*(3), 1432–1448.

Doudchenko, N., Zhang, M., Drynkin, E., Airoldi, E. M., Mirrokni, V., & Pouget-Abadie, J. (2020). Causal inference with bipartite designs. *Available at SSRN 3757188*.

Forastiere, L., Airoldi, E. M., & Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, *116*(534), 901–918. https://doi.org/10.1080/01621459.2020.1768100

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

Imai, K., Jiang, Z., & Malani, A. (2021). Causal inference with interference and noncompliance in two-stage randomized experiments. *Journal of the American Statistical Association*, *116*(534), 632–644. https://doi.org/10.1080/01621459.2020.1775612

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jackson, M. O., Lin, Z., & Yu, N. N. (2020). Adjusting for peer-influence in propensity scoring when estimating treatment effects. *Available at SSRN 3522256*.

Jagadeesan, R., Pillai, N. S., & Volfovsky, A. (2020). Designs for estimating the treatment effect in networks with interference. *The Annals of Statistics*, *48*(2), 679–712.

Liu, L., Hudgens, M. G., & Becker-Dreps, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika*, *103*(4), 829–842. https://doi.org/10.1093/biomet/asw047

Liu, L., Hudgens, M. G., Saul, B., Clemens, J. D., Ali, M., & Emch, M. E. (2019). Doubly robust estimation in observational studies with partial interference. *Stat*, *8*(1), e214.

Normand, S.-L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *Journal of Clinical Epidemiology*, *54*(4), 387–398.

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

Qu, Z., Xiong, R., Liu, J., & Imbens, G. (2021). Efficient treatment effect estimation in observational studies under heterogeneous partial interference. *arXiv preprint arXiv:2107.12420*.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524.

Rozemberczki, B., Allen, C., & Sarkar, R. (2021). Multi-scale attributed node embedding. *Journal of Complex Networks*, *9*(2), cnab014.

Sánchez-Becerra, A. (2021). Spillovers, homophily, and selection into treatment: The network propensity score. https://economics.sas.upenn.edu/system/files/2021-03/AlejandroSanchez_JMP_March2021_0.pdf

Saveski, M., Pouget-Abadie, J., Saint-Jacques, G., Duan, W., Ghosh, S., Xu, Y., & Airoldi, E. M. (2017). Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239.

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*(476), 1398–1407.

Sofrygin, O., & van der Laan, M. J. (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).

Tchetgen, E. J. T., & VanderWeele, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, *21*(1), 55–75.

Toulis, P., Volfovsky, A., & Airoldi, E. M. (2018). Propensity score methodology in the presence of network entanglement between treatments. *arXiv preprint arXiv:1801.07310*.

van der Laan, M. J. (2014). Causal inference for a population of causally connected units. *Journal of Causal Inference*, *2*(1), 13–74.

Zigler, C. M., & Papadogeorgou, G. (2021). Bipartite causal inference with interference. *Statistical Science: a review journal of the Institute of Mathematical Statistics*, *36*(1), 109.

A Technical derivations from example

This appendix section contains derivations of the overall treatment effects, estimators and their biases presented in Section 4.

A.1 Setting 1

In Setting 1, the overall treatment effect τ is given by

$$\begin{aligned}
\tau &= \sum_{x\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g,X_{ij}=x\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g,X_{ij}=x\right]\right)\mathbb{P}(X_{ij}=x)\\
&= \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g\right]\right)\sum_{x\in\{0,1\}}\mathbb{P}(X_{ij}=x)\\
&= \mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g\right]\\
&= c_z
\end{aligned}$$

where the first equality follows from Theorem 1 and the second from the model assumptions.

The naive estimator τ_X^{obs} that assumes SUTVA is given by

$$\begin{aligned}
\tau_X^{\text{obs}} &= \sum_{x,g\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,X_{ij}=x,G_{ij}=g\right]\mathbb{P}(G_{ij}=g|Z_{ij}=1,X_{ij}=x)\right.\\
&\quad\left.-\,\mathbb{E}\left[Y_{ij}|Z_{ij}=0,X_{ij}=x,G_{ij}=g\right]\mathbb{P}(G_{ij}=g|Z_{ij}=0,X_{ij}=x)\right)\mathbb{P}(X_{ij}=x)\\
&= \sum_{x,g\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g\right]\mathbb{P}(G_{ij}=g|Z_{ij}=1,X_{ij}=x)\right.\\
&\quad\left.-\,\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g\right]\mathbb{P}(G_{ij}=g|Z_{ij}=0,X_{ij}=x)\right)\mathbb{P}(X_{ij}=x)\\
&= \sum_{x,g\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g\right]\sum_{x'\in\{0,1\}}\mathbb{P}(G_{ij}=g|X_{ij}=x,X_{ik}=x')\mathbb{P}(X_{ik}=x'|Z_{ij}=1,X_{ij}=x)\right.\\
&\quad\left.-\,\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g\right]\sum_{x'\in\{0,1\}}\mathbb{P}(G_{ij}=g|X_{ij}=x,X_{ik}=x')\mathbb{P}(X_{ik}=x'|Z_{ij}=0,X_{ij}=x)\right)\mathbb{P}(X_{ij}=x)\\
&= \sum_{g\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g\right]\sum_{x,x'\in\{0,1\}}\mathbb{P}(G_{ij}=g|X_{ij}=x,X_{ik}=x')\mathbb{P}(Z_{ij}=1|X_{ij}=x,X_{ik}=x')\mathbb{P}(X_{ij}=x)\right.\\
&\quad\left.-\,\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g\right]\sum_{x,x'\in\{0,1\}}\mathbb{P}(G_{ij}=g|X_{ij}=x,X_{ik}=x')\mathbb{P}(Z_{ij}=0|X_{ij}=x,X_{ik}=x')\mathbb{P}(X_{ij}=x)\right)\\
&= \frac{1}{2}\left(\left(2\left(\frac{3}{4}\right)^2+2\left(\frac{1}{4}\right)^2\right)\left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=1\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=0\right]\right)\right.\\
&\quad\left.+\,4\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=0\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=1\right]\right)\right)\\
&= \frac{5}{8}\left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=1\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=0\right]\right)+\frac{3}{8}\left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=0\right]-\mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=1\right]\right)\\
&= \frac{5}{8}(c_z+c_g)+\frac{3}{8}(c_z-c_g)\\
&= c_z+\frac{1}{4}c_g
\end{aligned}$$

where the first equality follows from Theorem 2.A, consistency and the unconfoundedness assumption. The second equality follows from model assumptions, and the third from iterated conditioning on the counterpart unit X_{ik} of X_{ij}, $j \neq k$, and the fact that $Z_{ij} \perp\!\!\!\perp G_{ij} | X_{i1}, X_{i2}$. The fourth equality follows from Bayes' theorem manipulation where

$$\begin{aligned}
\mathbb{P}(X_{ik} = x' | Z_{ij} = 1, X_{ij} = x) &= \frac{\mathbb{P}(Z_{ij} = 1 | X_{ij} = x, X_{ik} = x')\mathbb{P}(X_{ik} = x')}{\mathbb{P}(Z_{ij} = 1 | X_{ij} = x)} \\
&= \frac{\mathbb{P}(Z_{ij} = 1 | X_{ij} = x, X_{ik} = x')\mathbb{P}(X_{ik} = x')}{\mathbb{P}(Z_{ij} = 1 | X_{ij} = X_{ik})P(X_{ik} = x) + \mathbb{P}(Z_{ij} = 1 | X_{ij} \neq X_{ik})P(X_{ik} = x')} \\
&= \left(\frac{\frac{1}{2}}{\left(\frac{3}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{2}\right)} \right) \mathbb{P}(Z_{ij} = 1 | X_{ij} = x, X_{ik} = x') \\
&= \mathbb{P}(Z_{ij} = 1 | X_{ij} = x, X_{ik} = x')
\end{aligned}$$

with the first equality following from Bayes' theorem and the assumption that $X_{i1} \perp\!\!\!\perp X_{i2}$.

The bias of the naive estimator can be verified using Equation 11 from Corollary 2, which gives

$$\begin{aligned}
\tau_X^{\text{obs}} - \tau &= \sum_{x \in \{0,1\}} \left(\mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 1, X_{ij} = x\right] - \mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 0, X_{ij} = x\right] \right) \\
&\quad \times \left(\mathbb{P}(G_{ij} = 1 | Z_{ij} = 1, X_{ij} = x) - \mathbb{P}(G_{ij} = 1 | Z_{ij} = 0, X_{ij} = x) \right) \mathbb{P}(X_{ij} = x) \\
&= \left(\mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 1\right] - \mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 0\right] \right) \\
&\quad \times \sum_{x, x' \in \{0,1\}} \left(\mathbb{P}(G_{ij} = 1 | X_{ij} = x, X_{ik} = x')\mathbb{P}(X_{ik} = x' | Z_{ij} = 1, X_{ij} = x) \right. \\
&\quad \left. - \mathbb{P}(G_{ij} = 1 | X_{ij} = x, X_{ik} = x')\mathbb{P}(X_{ik} = x' | Z_{ij} = 0, X_{ij} = x) \right) \mathbb{P}(X_{ij} = x) \\
&= \left(\mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 1\right] - \mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 0\right] \right) \\
&\quad \times \sum_{x, x' \in \{0,1\}} \left(\mathbb{P}(G_{ij} = 1 | X_{ij} = x, X_{ik} = x')\mathbb{P}(Z_{ij} = 1 | X_{ij} = x, X_{ik} = x') \right. \\
&\quad \left. - \mathbb{P}(G_{ij} = 1 | X_{ij} = x, X_{ik} = x')\mathbb{P}(Z_{ij} = 0 | X_{ij} = x, X_{ik} = x') \right) \mathbb{P}(X_{ij} = x) \\
&= \frac{1}{2} \left(2\left(\frac{3}{4}\right)^2 + 2\left(\frac{1}{4}\right)^2 - 4\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) \right) \left(\mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 1\right] - \mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 0\right] \right) \\
&= \frac{1}{4} \left(\mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 1\right] - \mathbb{E}\left[Y_{ij} | Z_{ij} = z, G_{ij} = 0\right] \right) \\
&= \frac{1}{4} c_g
\end{aligned}$$

where the second equality follows from model assumptions and iterated conditioning on the counterpart unit X_{ik}, $j \neq k$, and the third from the above Bayes' theorem manipulation.

A.2 Setting 2

In Setting 2, the overall treatment effect τ is given by

$$\begin{aligned}
\tau &= \sum_{g\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}(1,g)\right] - \mathbb{E}\left[Y_{ij}(0,g)\right]\right)\mathbb{P}(G_{ij}=g) \\
&= \sum_{g,x,u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,G_{ij}=g,X_{ij}=x,U_{ij}=u\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0,G_{ij}=g,X_{ij}=x,U_{ij}=u\right]\right) \\
&\quad \times \mathbb{P}(U_{ij}=u,X_{ij}=x)\mathbb{P}(G_{ij}=g) \\
&= \sum_{u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,U_{ij}=u\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0,U_{ij}=u\right]\right) \sum_{x\in\{0,1\}} \mathbb{P}(U_{ij}=u,X_{ij}=x) \\
&= \sum_{u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,U_{ij}=u\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0,U_{ij}=u\right]\right)\mathbb{P}(U_{ij}=u) \\
&= \frac{1}{2}\sum_{u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,U_{ij}=u\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0,U_{ij}=u\right]\right) \\
&= c_z
\end{aligned}$$

where the first equality follows from definition (Equation 3) and the second from the fact that $X_{ij}, U_{ij} \perp\!\!\!\perp G_{ij}$.

The naive estimator τ_X^{obs} that assumes SUTVA is given by

$$\begin{aligned}
\tau_X^{\text{obs}} &= \sum_{x\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,X_{ij}=x\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0,X_{ij}=x\right]\right)\mathbb{P}(X_{ij}=x) \\
&= \mathbb{E}\left[Y_{ij}|Z_{ij}=1\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=0\right] \\
&= \sum_{u\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=1,U_{ij}=u\right]\mathbb{P}(U_{ij}=u|Z_{ij}=1) - \mathbb{E}\left[Y_{ij}|Z_{ij}=0,U_{ij}=u\right]\mathbb{P}(U_{ij}=u|Z_{ij}=0)\right) \\
&= \frac{1}{4}c_z - 0 + \cdot\frac{3}{4}(c_z+c_u) - \frac{1}{4}c_u \\
&= c_z + \frac{1}{2}c_u
\end{aligned}$$

where the first equality follows from definition (Equation 8) and the second from model assumptions.

The bias of the naive estimator can be verified using Equation 14 from Corollary 3, which gives

$$\begin{aligned}
\tau_X^{\text{obs}} - \tau &= \sum_{x\in\{0,1\}} \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=z,X_{ij}=x,U_{ij}=1\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=z,X_{ij}=x,U_{ij}=0\right]\right) \\
&\quad \times \left(\mathbb{P}(U_{ij}=1|Z_{ij}=1,X_{ij}=x) - \mathbb{P}(U_{ij}=1|Z_{ij}=0,X_{ij}=x)\right)\mathbb{P}(X_{ij}=x) \\
&= \left(\mathbb{E}\left[Y_{ij}|Z_{ij}=z,U_{ij}=1\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=z,U_{ij}=0\right]\right)\left(\mathbb{P}(U_{ij}=1|Z_{ij}=1) - \mathbb{P}(U_{ij}=1|Z_{ij}=0)\right) \\
&= \frac{1}{2}\left(\mathbb{E}\left[Y_{ij}|Z_{ij}=z,U_{ij}=1\right] - \mathbb{E}\left[Y_{ij}|Z_{ij}=z,U_{ij}=0\right]\right) \\
&= \frac{1}{2}c_u
\end{aligned}$$

where the second equality follows from model assumptions.

B Additional details of simulation study

This appendix section contains additional details from the simulation study in Sections 5 and 6.

B.1 Twitch dataset

Table 4 shows the distribution of features `224` (game1) and `569` (game2) in the Twitch dataset.

	game2		
game1	0	1	Total
0	1582 (22%)	1728 (24%)	3310 (46%)
1	1834 (26%)	1982 (28%)	3816 (54%)
Total	3416 (48%)	3710 (52%)	7216

Table 4: Distribution of game1 and game2 in the Twitch dataset.

Figure 2 shows the distribution of the log degree in the Twitch dataset. The (untransformed) degrees range from 1 to 720 with a median of 5, a mean of 9.9 and a standard deviation of 22.2. The first and third quartiles are 2 and 11, respectively.

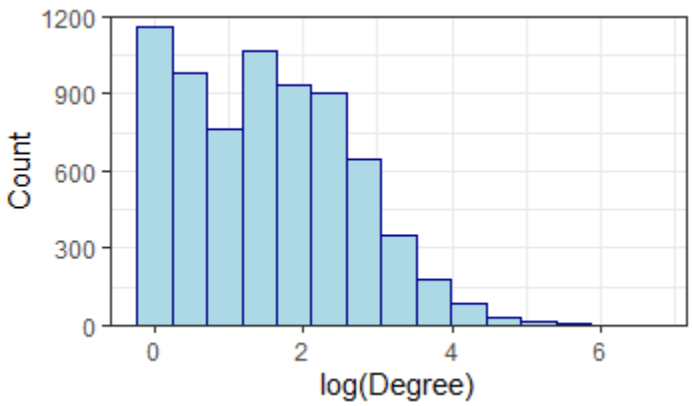

Figure 2: Distribution of log degree in the Twitch dataset.

B.2 Simulated treatment

Figure 3 shows the distribution of proportion and log sum neighbourhood treatments in one simulated dataset. For proportion G_i, the mean and median are 0.59 and 0.63, respectively. The standard deviation is 0.30, and the first and third quartiles are 0.44 and 0.80, respectively. For (untransformed) sum G_i, the values range from 0 to 489 with a median of 3, a mean of 6.3, and a standard deviation of 14.7. The first and third quartiles are 1 and 7, respectively.

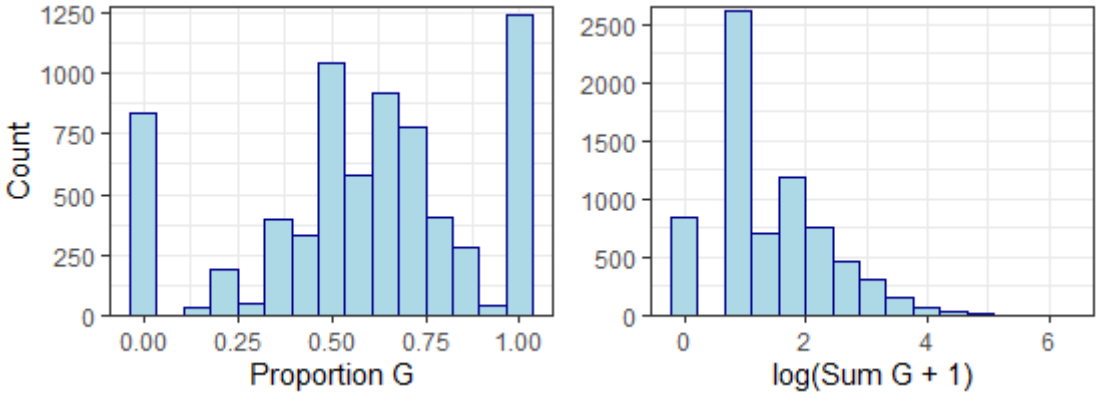

Figure 3: Distribution of proportion and log sum neighbourhood treatments in one simulated dataset. Note that for the sum plot, 1 is added to every G_i before taking the log due to the presence of zeros.

Tables 5 and 6 show the covariate balance across treatment arms in one simulated dataset. The standardized difference (Austin, 2011) is computed as

$$\text{Stand. Diff.} = \frac{\bar{X}_{Z=1} - \bar{X}_{Z=0}}{\sqrt{\frac{s^2_{Z=1}+s^2_{Z=0}}{2}}}$$

for continuous variables (all but game1, game2, and Z_i) where s^2 is the variance for the corresponding group, and otherwise computed as

$$\text{Stand. Diff.} = \frac{\bar{X}_{Z=1} - \bar{X}_{Z=0}}{\sqrt{\frac{\bar{X}_{Z=1}(1-\bar{X}_{Z=1})+\bar{X}_{Z=0}(1-\bar{X}_{Z=0})}{2}}}$$

for binary variables. Some works in the literature interpret a standardized difference of less than 0.1 (10%) as the covariate being balanced across treatment arms (Normand et al., 2001).

Variable	$\bar{X}_{Z=1}$	$\bar{X}_{Z=0}$	Stand. Diff.
game1	0.683	0.326	0.763
game2	0.763	0.176	1.454
Ngame1	0.486	0.466	0.061
Ngame2	0.620	0.586	0.114
Degree	10.757	8.717	0.088
Proportion G_i	0.604	0.569	0.118
Sum G_i	6.979	5.365	0.106

Table 5: Covariate balance across individual treatment arms.

Variable	$\bar{X}_{G\geq 0.5}$	$\bar{X}_{G<0.5}$	Stand. Diff.	$\bar{X}_{G\geq 3}$	$\bar{X}_{G<3}$	Stand. Diff.
game1	0.558	0.471	0.175	0.723	0.334	0.846
game2	0.542	0.461	0.162	0.555	0.483	0.144
Ngame1	0.553	0.267	0.938	0.564	0.385	0.559
Ngame2	0.669	0.430	0.818	0.631	0.580	0.174
Degree	11.694	4.897	0.371	16.905	2.432	0.701
Z_i	0.608	0.526	0.167	0.677	0.490	0.385

Table 6: Covariate balance across dichotomized neighbourhood treatment arms. For sum G_i, the median 3 is taken as the threshold.

B.3 Number of subclasses investigation

We briefly examine the effect of the number of subclasses on the RMSE of the $\mathbf{X}_i^{\text{ind}}$ and $\mathbf{X}_i^z$-adjusted estimators and of the GPS estimator (Forastiere et al., 2021). Figure 4 shows the effect of the number of subclasses on the mean RMSE over 25 datasets simulated with high interference and either proportion neighbourhood treatments (left) and sum neighbourhood treatments (right). The same 25 datasets were used to compute the RMSE across numbers of subclasses in each plot. Our results suggest that while the five subclass recommendation by Rosenbaum and Rubin (1984) does not necessarily lead to the smallest RMSE, the RMSE of most estimators do not change significantly with more than five subclasses (the one exception being the estimator that adjusts for $\mathbf{X}_i^z$ in the sum neighbourhood treatment case, for which the RMSE continues to decrease up to nine subclasses). It is also interesting that increasing the number of subclasses does not necessarily reduce the RMSE, notably with the $\mathbf{X}_i^z$-adjusted and the GPS estimators where the RMSE spikes at four subclasses. The reason for this is unclear. Given more time, possible reasons to look into include poor simulation setup (25 simulations may be too few), quirks of the Twitch dataset (does this also occur with more data or other datasets?), and incorrect code and/or numerical instabilities.

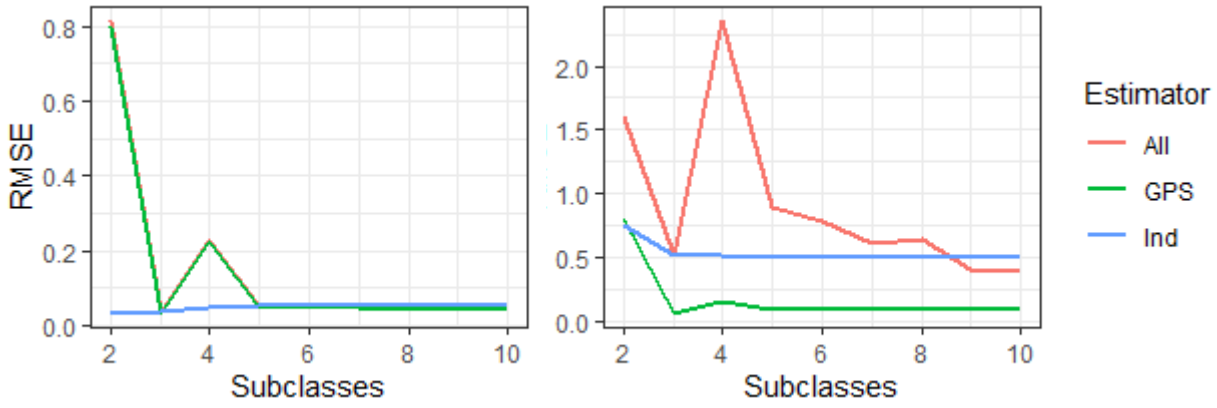

Figure 4: RMSE as the number of subclasses changes for three treatment effect estimators. The left plot shows the mean RMSE over 25 datasets simulated with proportion neighbourhood treatment and the right for those with sum neighbourhood treatments. The RMSE of the all-covariates and the GPS estimators follow similar trajectories in the left plot.