Difference-in-Differences with Interference: A Finite Population Perspective

Ruonan Xu*

First draft

Abstract

By relaxing the stable unit treatment value assumption, I study the "difference-in-differences" (DID) type of estimators that allow interference. When spillover effects are of interest, we often sample the entire population. Thus, I adopt a finite population perspective in the sense that the estimands are defined as population averages and inference is conditional on the attributes of all population units. The general and unified approach in this paper relaxes common restrictions in the literature, such as partial interference and correctly specified spillover functions. I propose doubly robust estimators for the direct average treatment effect on the treated as well as spillover effects under a modified parallel trends assumption. Moreover, robust inference is discussed based on the asymptotic distribution of the proposed estimators. Using the two time period DID estimator as a building block, I then extend the setting to multiple time periods with constant treatment timing.

1 Introduction

In the fields of environmental economics, urban economics, criminal justice, and many other fields of social sciences, place-based policies often generate spillover effects. One example studied by Jardim, Long, Plotnick, Van Inwegen, Vigdor, and Wething (2022) is minimum

*ruonan.xu@rutgers.edu, Rutgers University

wage increase in Seattle. Through the channels of competition in the regional labor market for workers and the possibility of relocation of businesses, they find that significant spillover effects on wages and hours are seen up to a 40-minute drive from Seattle city limits.

When spillover effects are of interest, we often need to observe the entire population. For example, we can typically collect information about all counties in the United States. In the example above, Jardim et al. (2022) use administrative employment records in the state of Washington. As pointed out by Manski (1993), "it does not make sense in studies of neighborhood and other large-group social effects, where the sample members are randomly chosen individuals. Taken at face value, equation (7) implies that the sample members know who each other are and choose their outcomes only after having been selected into the sample." If we take sampling from the superpopulation approach literally, what we are estimating turns out to be the spillover effect in a researcher's sample rather than in the population from which the sample is drawn unless interactions are restricted within clusters of friends or household members.

By relaxing the stable unit treatment value assumption (SUTVA), I study the "difference-in-differences" (DID) type of estimators that allow interference from a finite population perspective, where the entire population is observed. Having said that, the approach I take is different from the design-based approach. In the literature of finite population causal inference, there are many choices of the conditioning variables. Here only the attributes of the entire population are conditioned on and hence the conditional means of potential outcomes are allowed to be flexibly modeled. As a result, our approach can be considered as a middle ground between superpopulation and design-based frameworks. The upside is that I can make the proposed estimators more robust to model specification with straightforward causal interpretation. Meanwhile, I still maintain the flavor of conditional inference in the design-based approach, which can provide guidance on robust inference.

The spatial setting here is different but close to that in Xu and Wooldridge (2022). A finite population in a spatial space is characterized by fixed attributes containing intrinsic locational information and neighborhood characteristics. Meanwhile, potential outcome functions can be stochastic, partly due to measurement errors. With the entire population observed, the sampling probability is essentially one, which shows up in the conditional asymptotic variance-covariance matrix derived below. The inference becomes more precise when the interest on the finite population is recognized. Our hybrid approach can be considered as an application of the conditional inference discussed by Abadie, Imbens, and Zheng (2014) to DID type estimators.

Most of the literature studies spillover effects in a single cross section with experimental data and assumes partial interference or limits interference to immediate neighbors. Additionally, they assume that the function of dependence on neighbors' treatments is known and correctly specified. See, for instance, Hudgens and Halloran (2008) and Aronow and Samii (2017). Delgado and Florax (2015), Clarke (2017), and Butts (2021) allow interference in DID estimation in a two-way fixed effects (TWFE) framework (often without covariates) from a superpopulation perspective. All three papers mentioned above share some or all limitations of the general interference literature. Design-based DID estimation has been studied by Athey and Imbens (2022), Rambachan and Roth (2022), and Arkhangelsky, Imbens, Lei, and Luo (2021) but they keep the SUTVA.

I work with observational data in this paper since it is the most common type of data in economics. I consider the expected direct treatment effect at certain neighborhood exposure levels or the expected spillover effect at different neighborhood exposure levels. As a result, the causal estimands are well defined even when the spillover function is misspecified. In terms of relaxing the assumption of a fixed neighborhood boundary, I apply the device of approximate neighborhood interference (ANI) in Leung (2022) to spatial data, in which

treatments assigned to units further from i have a smaller, but possibly nonzero, effect on i's response. In addition, the assignment variables are allowed to be spatially correlated as is often the case in practice with spatial data. To sum up, our approach is the most general one so far and is closest to empirical settings.

Our contribution is fourfold. First, I lay the basis for studying direct and spillover effects in a DID context with the entire population observed. Second, I study the identification of canonical DID estimators available in the literature. I provide conditions under which canonical estimators still identify meaningful causal estimands. This discussion alone would be of interest to practitioners. Third, I clarify what toolkit practitioners can use by comparing various dimension reduction approaches in the interference literature. Last and most importantly, I provide solutions to the question under study by proposing doubly robust estimators for the direct treatment effect and spillover effect. Our doubly robust estimator is a modified version of the augmented inverse probability weighting (AIPW) estimator, which only requires correct specification of either the propensity score of treatment or the conditional mean of the outcomes. Sant'Anna and Zhao (2020) has proposed AIPW estimators in the DID context, maintaining SUTVA and the superpopulation framework. Once interference is allowed for, one of the biggest challenges is incorporating it in a general and flexible manner. On top of the different estimands I target, our conditional inference approach also leads to a different variance-covariance matrix which requires a new variance estimator when necessary.

2 Setup

2.1 Environment

I start with the relatively simple setting of panel data with two time periods; $t = 1, 2$ stands for the time period before and after treatment respectively. Let $D \subseteq \mathbb{R}^d$, $d \geq 1$, be a lattice of (possibly) unevenly placed locations in $\mathbb{R}^d$. Consider a sequence of finite subsets of D, $\{D_M\}$, where M indexes the sequence of finite populations. $|D_M|$ diverges to infinity in deriving the asymptotic properties, where $|V|$ denotes the cardinality of a finite subset $V \subseteq D$.

For each unit i in the population, there is a stochastic assignment variable $W_i \in \{0, 1\}$, and a vector of fixed attributes z_i. The potential outcome function is defined to be a mapping from the treatment vector of the entire population $y_{it}(w_i, \boldsymbol{w}_{-i})$, where $\boldsymbol{w}_{-i} = \{w_j, j \in D_M, j \neq i\}$. The realized potential outcomes are denoted by $Y_{it} = y_{it}(\boldsymbol{W})$. Notice that $(\boldsymbol{W}, \boldsymbol{z}, \boldsymbol{Y}) = \{(W_i, z_i, Y_{it}(\cdot)), i \in D_M, M \geq 1\}$ are triangular arrays of random fields defined on a probability space $(\Omega, \mathcal{F}, P)$. Exposure mapping is defined by the function $G_i = G(i, \boldsymbol{w}_{-i}) \in \mathcal{G}$, where $\mathcal{G}$ is a discrete set.

The setup is closest to that in Xu and Wooldridge (2022). The main difference here is that I allow the potential outcome functions to be random. In addition, no sampling process is involved since the entire population is observed. In other words, the sampling probability is one. The key to the finite population perspective is to allow positive sampling probabilities.

It is worth explaining the construction of the $G(\cdot)$ function here. Given a fixed K, define the K-neighborhood of unit i as

$$\mathcal{N}(i, K) = \{j \in D_M : \rho(i, j) \leq K, j \neq i\}$$

Let $\boldsymbol{w}_{\mathcal{N}(i,K)} = (w_j : j \in \mathcal{N}(i,K))$ be the treatment vector of units within i's K-neighborhood. There exists $K < \infty$ such that for all $\boldsymbol{w}_{-i}$ and $\boldsymbol{w}'_{-i}$ such that $\boldsymbol{w}_{\mathcal{N}(i,K)} = \boldsymbol{w}'_{\mathcal{N}(i,K)}$, $G(i, \boldsymbol{w}_{-i}) = G(i, \boldsymbol{w}'_{-i})$. As a result, the specified exposure mapping function restricts spillover effects within the immediate K-neighborhood of each unit. Having said that, the actual potential outcome function places no restriction on the interference structure. Treatments of units outside of i's K-neighborhood can legitimately influence i's potential outcome as long as treatments assigned to units further from i have a smaller, but possibly nonzero, effect on i's response. A detailed description of the assumptions is given in Section 4 below. This way, the exposure mapping function is allowed to be arbitrarily misspecified. The $G(\cdot)$ function is allowed to be multidimensional, in which the K distance would be the largest distance that interference is allowed under the specification across the fixed dimensions of $G(\cdot)$.

I briefly summarize the notation used throughout the paper. I adopt the metric $\rho(i,j) = \max_{1 \le l \le d} |j_l - i_l|$ in space $\mathbb{R}^d$, where i_l is the l-th component of i. The distance between any subsets $K, V \subseteq D$ is defined as $\rho(K,V) = \inf\{\rho(i,j) : i \in K \text{ and } j \in V\}$. For any random vector W, $\|W\|_p = \left[\mathbb{E}(\|W\|^p \,|\boldsymbol{z})\right]^{1/p}$, $p \ge 1$, denotes its L_p-norm. Lastly, C denotes a generic positive constant that may vary under different circumstances.

2.2 Estimands of Interest

To allow for flexible modeling of the conditional mean of potential outcomes, I adopt a hybrid of model-based and design-based frameworks. This paper is interested in the expected finite population average, i.e., the average of the expected potential outcome across all units in the finite population. In other words, I focus on conditional inference given fixed attributes z_i; see Abadie et al. (2014) and Jin and Rothenhäusler (2023) for detailed discussion of conditional parameters and conditional inference. I take the finite population

perspective in the sense that the entire population is observed with fixed attributes.

There are two types of estimands of interest. The first parameter is the expected direct average treatment effect on the treated (EDATT) at exposure level g.

$$\tau(g) = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\big[y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i}) | W_i = 1, G_i = g, z_i\big] \tag{1}$$

The key ingredient of the definition is the expected potential outcome at exposure level g,

$$\begin{aligned} &\mathbb{E}\big[y_{i2}(1, \boldsymbol{w}_{-i}) | W_i = 1, G_i = g, z_i\big] \\ &= \sum_{\overline{\boldsymbol{W}}_{-i} \in \Omega} \mathbb{E}\big[y_{i2}(1, \overline{\boldsymbol{W}}_{-i}) | W_i = 1, \boldsymbol{W}_{-i} = \overline{\boldsymbol{W}}_{-i}, z_i\big] P(\boldsymbol{W}_{-i} = \overline{\boldsymbol{W}}_{-i} | G_i = g, W_i = 1, z_i), \end{aligned}$$

where the expectation is taken over all possible realizations of $\boldsymbol{W}_{-i}$ given the specified exposure mapping $G(i, \boldsymbol{w}_{-i})$ and $\Omega = \{0, 1\}^{|D_M|-1}$.

In terms of the interpretation of EDATT, if the spillover effect and the direct effect are additively separable, we can identify the exact direct ATT even if the spillover function is misspecified. Without additivity, we can still identify the direct ATT that would realize in expectation at the specified exposure level.

In addition to EDATT, empirical researchers might also be interested in spillover effects defined in equations (2) and (3).

$$\frac{1}{|D_M|} \sum_{i \in D_M} \Big(\mathbb{E}\big[y_{i2}(1, \boldsymbol{w}_{-i}) | W_i = 1, G_i = g, z_i\big] - \mathbb{E}\big[y_{i2}(1, \boldsymbol{w}'_{-i}) | W_i = 1, G_i = g', z_i\big] \Big) \tag{2}$$

$$\frac{1}{|D_M|} \sum_{i \in D_M} \Big(\mathbb{E}\big[y_{i2}(0, \boldsymbol{w}_{-i}) | W_i = 0, G_i = g, z_i\big] - \mathbb{E}\big[y_{i2}(0, \boldsymbol{w}'_{-i}) | W_i = 0, G_i = g', z_i\big] \Big) \tag{3}$$

The spillover effect contrasts the expected potential outcomes between levels g and g' and could differ with or without direct treatment. A leading case would be setting g' to 0. The

identification of the spillover effect is more straightforward because the potential outcomes under direct assignment and the specified exposure are observable.

3 Identification

The first question when relaxing SUTVA is whether it matters if spillover effects are ignored when estimating the treatment effect. Namely, will the canonical DID estimator consistently estimate ATT when interference occurs? To facilitate the discussion of identification, I impose the following assumptions.

Assumption 1 *(Overlap)* $\forall\ i \in D_M$, *there exists* $\epsilon > 0$ *such that* $\epsilon < p(z_i) < 1 - \epsilon$, $\pi_{1g}(z_i) > \epsilon$, *and* $\pi_{0g}(z_i) > \epsilon$, *where*

$$p(z_i) = P(W_i = 1|z_i), \tag{4}$$

$$\pi_{1g}(z_i) = P(G_i = g|W_i = 1, z_i), \tag{5}$$

and

$$\pi_{0g}(z_i) = P(G_i = g|W_i = 0, z_i). \tag{6}$$

To simplify the notation, I assume that the overlap assumption applies to every unit in the population. With certain exposure mapping specifications, this might not be plausible. An easy fix is to change the estimand by averaging over the subpopulation where G_i can take on the value g. Also, please see Man, Sant'Anna, Sasaki, and Ura (2023) for trimming propensity scores with bias correction when the overlap condition holds weakly.

Assumption 2 *(No Anticipation)*

$$\frac{1}{|D_M|}\sum_{i\in D_M}\mathbb{E}\big[y_{i1}(w_i, \boldsymbol{w}_{-i})|W_i, z_i\big] = \frac{1}{|D_M|}\sum_{i\in D_M}\mathbb{E}\big[y_{i1}(0, \underline{0})|W_i, z_i\big]$$

Assumption 2 requires that the expected potential outcome in the first time period before treatment is always equal to the expected potential outcome without treatment nor spillover. The no anticipation assumption is quite standard in the literature, sometimes implicitly.

To identify EDATT, I impose the following parallel trends assumption:

Assumption 3 *(Parallel Trends)*

$$\begin{aligned}&\frac{1}{|D_M|}\sum_{i\in D_M}\Big[\mathbb{E}\big(y_{i2}(0,\boldsymbol{w}_{-i})|W_i=1,G_i=g,z_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=1,z_i\big)\Big]\\=&\frac{1}{|D_M|}\sum_{i\in D_M}\Big[\mathbb{E}\big(y_{i2}(0,\boldsymbol{w}_{-i})|W_i=0,G_i=g,z_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=0,z_i\big)\Big]\end{aligned}\tag{7}$$

A sufficient condition for Assumption 3 is that for any $g^*\in\mathcal{G}^*$,

$$\begin{aligned}&\frac{1}{|D_M|}\sum_{i\in D_M}\Big[\mathbb{E}\big(y_{i2}(0,g^*)|W_i=1,G_i^*=g^*,z_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=1,z_i\big)\Big]\\=&\frac{1}{|D_M|}\sum_{i\in D_M}\Big[\mathbb{E}\big(y_{i2}(0,g^*)|W_i=0,G_i^*=g^*,z_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=0,z_i\big)\Big],\end{aligned}\tag{8}$$

where G^* stands for the unknown true exposure mapping and $\mathcal{G}^*$ is the set of values that G^* can take. If equation (8) holds, then Assumption 3 is satisfied by the law of iterated expectations invariant of the specified exposure mapping function.

If we remove the outer average, and assume that equality holds for each unit $i\in D_M$, then Assumption 3 becomes the conditional parallel trends. Further notice that in the parallel trends assumption, we do not condition on the exposure level g in the first time period as no one is treated at $t=1$. Accordingly, there is no spillover in the potential outcome function in the first time period. Assumptions 2 and 3 can be relaxed if we observe multiple time periods before treatment or one is willing to model the differential time trends among the control and treatment groups. However, to fix idea I keep them in

the standard form in the literature.

There is a growing literature on justification and falsification of the parallel trends assumption under SUTVA; see, for instance, Roth and Sant'Anna (2023) and Ghanem, Sant'Anna, and Wüthrich (2022). When parallel trends might be violated, Rambachan and Roth (2023) present confidence sets for the identified set of treatment effects. The extension of these analyses to parallel trends with interference is out of the scope of this paper. Readers can refer to the references above for intuition. Since no units are treated prior to the treatment, there is no spillover effect before $t = 2$. Therefore, the feasible classical pre-trends test here is reduced to the standard case without interference. The interpretation of the pre-trends test requires caution, though; see Roth (2022).

3.1 Canonical DID

The usual ATT under the SUTVA is

$$\tilde{\tau} = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\big(y_{i2}(1) - y_{i2}(0) | W_i = 1, z_i\big).$$

Here, the potential outcomes are determined solely by unit i's own treatment. Suppose the canonical DID estimator consistently estimates

$$\tau_{canonic} = \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}(Y_{i2} - Y_{i1} | W_i = 1, z_i) - \mathbb{E}(Y_{i2} - Y_{i1} | W_i = 0, z_i)\Big].$$

Examples include the TWFE linear estimating equation in Remark 1 in Sant'Anna and Zhao (2020) under the additional restrictions of the data generating process there and the inverse probability weighting (IPW) estimator in Abadie (2005). If the usual (conditional) parallel trends assumption holds without interference, $\tau_{canonic}$ would be equivalent to $\tilde{\tau}$.

If SUTVA is violated, EDATT is generally determined by the specific exposure level.

As a result, I use the overall direct effect as a benchmark for comparison.

$$\tau = \sum_{g \in \mathcal{G}} \tau(g) P(G_i = g | W_i = 1, z_i)$$

The overall direct effect is comparable to the expected average treatment effect (EATE) studied by Sävje, Aronow, and Hudgens (2021), which in our notation is equal to

$$\begin{aligned}
\tau_{EATE} =& \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(1, \boldsymbol{w}_{-i})\big) - \mathbb{E}\big(y_{i2}(0, \boldsymbol{w}_{-i})\big) \Big] \\
=& \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\sum_{g \in \mathcal{G}} \mathbb{E}\big(y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i}) | G_i = g\big) P(G_i = g) \Big]
\end{aligned}$$

The difference between τ and τ_{EATE} is that the expected potential outcome and the propensity score is further conditional on $W_i = 1$ and z_i because of the DID setting. With abuse of notation, I get rid of the finite population average for the parameters τ and $\tau_{canonic}$ for now as it does not affect the comparison.

I first suppose that the parallel trends assumption (7) holds. Using the law of iterated expectations, τ and $\tau_{canonic}$ can be decomposed in the following way:

$$\begin{aligned}
\tau =& \sum_{g \in \mathcal{G}} \mathbb{E}(Y_{i2} | W_i = 1, G_i = g, z_i) P(G_i = g | W_i = 1, z_i) - \mathbb{E}(Y_{i1} | W_i = 1, z_i) \\
& - \Big[\sum_{g \in \mathcal{G}} \mathbb{E}(Y_{i2} | W_i = 0, G_i = g, z_i) P(G_i = g | W_i = 1, z_i) - \mathbb{E}(Y_{i1} | W_i = 0, z_i) \Big]
\end{aligned}$$

$$\begin{aligned}
\tau_{canonic} =& \sum_{g \in \mathcal{G}} \mathbb{E}(Y_{i2} | W_i = 1, G_i = g, z_i) P(G_i = g | W_i = 1, z_i) - \mathbb{E}(Y_{i1} | W_i = 1, z_i) \\
& - \Big[\sum_{g \in \mathcal{G}} \mathbb{E}(Y_{i2} | W_i = 0, G_i = g, z_i) P(G_i = g | W_i = 0, z_i) - \mathbb{E}(Y_{i1} | W_i = 0, z_i) \Big]
\end{aligned}$$

Proposition 1 *Under Assumptions 1-3,* $\tau_{canonic} \neq \tau$ *in general unless* $P(G_i = g|W_i = 0, z_i) = P(G_i = g|W_i = 1, z_i)$.

Notice that if the direct effect and the spillover effect are separately additive, the canonical DID estimator can still identify the direct effect. For a generic potential outcome function, a sufficient condition for equality would be $G_i \perp\!\!\!\perp W_i \mid z_i$. However, conditional independence can be easily violated if either of the following is true: (i) G_i and W_i are linked through covariates not included in z_i; (ii) neighbors' behavior affects unit i's treatment uptake; (iii) similar neighborhood characteristics drive the assignment mechanism; see Forastiere, Airoldi, and Mealli (2021) for a parallel discussion allowing interference on networks under unconfoundedness.

Secondly, suppose parallel trends (7) fails but a modified version holds conditional on additional attributes.

$$\begin{aligned} &\frac{1}{|D_M|}\sum_{i\in D_M}\Big[\mathbb{E}\big(y_{i2}(0,\boldsymbol{w}_{-i})|W_i=1,G_i=g,z_i,u_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=1,z_i,u_i\big)\Big] \\ =&\frac{1}{|D_M|}\sum_{i\in D_M}\Big[\mathbb{E}\big(y_{i2}(0,\boldsymbol{w}_{-i})|W_i=0,G_i=g,z_i,u_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=0,z_i,u_i\big)\Big], \end{aligned} \tag{9}$$

where u_i are additional attributes. Similarly, τ and $\tau_{canonic}$ can be rewritten as

$$\begin{aligned} \tau =& \sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i2}|W_i=1,G_i=g,z_i,u_i=u)P(u_i=u|W_i=1,G_i=g,z_i) \\ &\cdot P(G_i=g|W_i=1,z_i)-\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i1}|W_i=1,z_i,u_i=u)P(u_i=u|W_i=1,z_i) \\ &-\Big[\sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i2}|W_i=0,G_i=g,z_i,u_i=u)P(u_i=u|W_i=1,G_i=g,z_i) \\ &\cdot P(G_i=g|W_i=1,z_i)-\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i1}|W_i=0,z_i,u_i=u)P(u_i=u|W_i=1,z_i)\Big] \end{aligned}$$

$$
\begin{aligned}
\tau_{canonic} = & \sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \mathbb{E}(Y_{i2} | W_i = 1, G_i = g, z_i, u_i = u) P(u_i = u | W_i = 1, G_i = g, z_i) \\
& \cdot P(G_i = g | W_i = 1, z_i) - \sum_{u \in \mathcal{U}} \mathbb{E}(Y_{i1} | W_i = 1, z_i, u_i = u) P(u_i = u | W_i = 1, z_i) \\
& - \Big[\sum_{g \in \mathcal{G}} \sum_{u \in \mathcal{U}} \mathbb{E}(Y_{i2} | W_i = 0, G_i = g, z_i, u_i = u) P(u_i = u | W_i = 0, G_i = g, z_i) \\
& \cdot P(G_i = g | W_i = 0, z_i) - \sum_{u \in \mathcal{U}} \mathbb{E}(Y_{i1} | W_i = 0, z_i, u_i = u) P(u_i = u | W_i = 0, z_i) \Big]
\end{aligned}
$$

Proposition 2 *Under Assumption 1, the modified no anticipation assumption conditional on (W_i, z_i, u_i), and the modified parallel trends (9), $\tau_{canonic} \neq \tau$ unless*

$$
P(G_i = g | W_i = 0, z_i) = P(G_i = g | W_i = 1, z_i)
$$

and

$$
P(u_i = u | W_i = 1, G_i = g, z_i) = P(u_i = u | W_i = 0, G_i = g, z_i).
$$

As a result, even if $G_i \perp\!\!\!\perp W_i \mid z_i$ or SUTVA holds, the canonical DID estimator is still biased because of the exclusion of u_i. Assuming away interference, the omitted attributes often can be the attributes of unit i's neighbors.

3.2 Modified Two-Way Fixed Effects

Another approach to estimate the spillover effect suggested in the literature is to augment the TWFE DID regression with another binary indicator S_i equal to one if a unit is close to the treated unit; see, for instance, Di Tella and Schargrodsky (2004) and Butts (2021). Using our notation, the estimating equation becomes

$$
Y_{it} = \beta_1 W_{it} + \beta_2 (1 - W_{it}) S_i + \beta_3 W_{it} S_i + \alpha_i + \lambda_t + \epsilon_{it}, \tag{10}
$$

where $W_{it} = W_i * \mathbb{1}\{t = 2\}$. $\hat{\beta}_1$ estimated from equation (10) would be consistent for the EDATT defined by

$$\bar{\tau}(0) = \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(1, \underline{0}) - y_{i2}(0, \underline{0}) | W_i = 1, S_i = 0\big) \Big]$$

under the parallel trends assumption

$$\begin{aligned} &\frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(0, \underline{0}) - y_{i1}(0, \underline{0}) | W_i = 1, S_i = 0\big) \Big] \\ =&\frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(0, \underline{0}) - y_{i1}(0, \underline{0}) | W_i = 0, S_i = 0\big) \Big]. \end{aligned}$$

Similarly $\hat{\beta}_1 + \hat{\beta}_3 - \hat{\beta}_2$ would be consistent for the EDATT defined by

$$\bar{\tau}(1) = \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i}) | W_i = 1, S_i = 1\big) \Big]$$

under the parallel trends assumption

$$\begin{aligned} &\frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(0, \boldsymbol{w}_{-i}) - y_{i1}(0, \underline{0}) | W_i = 1, S_i = 1\big) \Big] \\ =&\frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(0, \boldsymbol{w}_{-i}) - y_{i1}(0, \underline{0}) | W_i = 0, S_i = 1\big) \Big]. \end{aligned}$$

$\bar{\tau}(0)$ and $\bar{\tau}(1)$ are the direct ATT without neighborhood exposure and EDATT with neighborhood exposure respectively only if the distance cutoff, $\bar{d}$, for the interference structure is correctly chosen. Namely, units with $S_i = 1$ indeed receive spillover and those with $S_i = 0$ indeed receive no spillover at all. If the cutoff is chosen too small, then $\bar{\tau}(0)$ becomes

the EDATT,

$$\bar{\tau}(0) = \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i}) | W_i = 1, G_i = 0\big) \Big],$$

where $G(i, \boldsymbol{w}_{-i}) = \mathbb{1}\{A_s \boldsymbol{W} > 0\}$ and A_s is the adjacency matrix with units being neighbors if their distance is less than or equal to $d_s < \bar{d}$. Analogously,

$$\bar{\tau}(1) = \frac{1}{|D_M|} \sum_{i \in D_M} \Big[\mathbb{E}\big(y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i}) | W_i = 1, G_i = 1\big) \Big]$$

with the exposure mapping $G(i, \boldsymbol{w}_{-i}) = \mathbb{1}\{A_s \boldsymbol{W} > 0\}$ no matter the cutoff d_s is chosen too small or too large.

We can see that given the estimating equation of the augmented TWFE, the specified exposure mapping is fixed as $\mathbb{1}\{A_s \boldsymbol{W} > 0\}$. Only when the interference structure coincides with the indicator function $\mathbb{1}\{A_s \boldsymbol{W} > 0\}$ along with the correct distance cutoff, can we identify the exact direct ATT. In contrast, our approach instead can identify the EDATT, $\tau(g)$, with varying levels of neighborhood exposure g allowing for misspecification of the spillover structure. We can also identify the exact direct ATT when the exposure mapping is correctly specified allowing for various interference structure. Meanwhile, we can flexibly account for covariates by assuming the conditional parallel trends. Furthermore, the basic augmented TWFE regression linear in covariates,

$$Y_{it} = \beta_0 + \beta_1 W_{it} + \beta_2 (1 - W_{it}) S_i + \beta_3 W_{it} S_i + \beta_4 W_i + z_i \gamma + \lambda_t + \epsilon_{it},$$

suffers from the same drawbacks of the usual canonical TWFE regression for DID estimation as pointed out by Remark 1 in Sant'Anna and Zhao (2020). Adding interactions of the covariates with the treatment and time indicators can help.

3.3 Doubly Robust Estimand

Since ignoring the spillover effect is only harmless under special scenarios, we need to propose new estimators for the EDATT. Under parallel trends and overlap assumptions, the EDATT can be identified by inverse weighting using propensity scores.

$$\begin{aligned}\tau(g) =& \frac{1}{|D_M|}\sum_{i\in D_M}\mathbb{E}\left[\frac{W_i - p(z_i)}{p(z_i)(1-p(z_i))}\left(\frac{\mathbb{1}\{G_i = g\}}{W_i\pi_{1g}(z_i) + (1-W_i)\pi_{0g}(z_i)}Y_{i2} - Y_{i1}\right)\middle| z_i\right] \\ =& \mathbb{E}_D\left[\frac{W_i - p(z_i)}{p(z_i)(1-p(z_i))}\left(\frac{\mathbb{1}\{G_i = g\}}{W_i\pi_{1g}(z_i) + (1-W_i)\pi_{0g}(z_i)}Y_{i2} - Y_{i1}\right)\right]\end{aligned} \tag{11}$$

To simplify notation, I use $\mathbb{E}_D$ to denote the finite population average conditional on the attributes $\boldsymbol{z}$ from now on.

Without the indicator for G and the additional propensity scores for spillover, the IPW-DID estimand is the same as the estimand proposed in Abadie (2005). To allow for more robustness against misspecification of the propensity scores, the IPW-DID estimand can be extended to an AIPW estimand in the similar spirit of Ning, Peng, and Tao (2020).

Define the conditional means of the potential outcome as

$$\mu_{it,wg}(z_i) = \mathbb{E}(Y_{it}|W_i = w, G_i = g, z_i) \tag{12}$$

or

$$\mu_{it,w}(z_i) = \mathbb{E}(Y_{it}|W_i = w, z_i). \tag{13}$$

Let $m_{it,wg}(z_i)$ and $m_{it,w}(z_i)$ denote the model for equations (12) and (13), respectively. Denote $\Delta m_{i2,g}(z_i) = m_{i2,1g}(z_i) - m_{i2,0g}(z_i)$ and $\Delta m_{i1}(z_i) = m_{i1,1}(z_i) - m_{i1,0}(z_i)$. Furthermore, let $\eta(z_i)$, $\eta_{1g}(z_i)$, and $\eta_{0g}(z_i)$ be the models for the propensity scores in equations (4)-(6), respectively.

The doubly robust estimand is

$$\begin{aligned}\tau(g) =& \mathbb{E}_D \Bigg[\frac{W_i}{\eta(z_i)} \Bigg(\frac{\mathbb{1}\{G_i = g\}}{\eta_{1g}(z_i)} \big(Y_{i2} - m_{i2,1g}(z_i)\big) - \big(Y_{i1} - m_{i1,1}(z_i)\big) \Bigg) \\ & - \frac{1 - W_i}{1 - \eta(z_i)} \Bigg(\frac{\mathbb{1}\{G_i = g\}}{\eta_{0g}(z_i)} \big(Y_{i2} - m_{i2,0g}(z_i)\big) - \big(Y_{i1} - m_{i1,0}(z_i)\big) \Bigg) \\ & + \Delta m_{i2,g}(z_i) - \Delta m_{i1}(z_i) \Bigg]. \end{aligned} \tag{14}$$

Proposition 3 *Under Assumptions 1-3, equation (14) recovers the EDATT, $\tau(g)$, as long as either the propensity scores or the conditional means of the outcome are correctly specified.*

Although DID estimators identify the ATT, the doubly robust estimand here formulates the AIPW in the same way as the ATE rather than the ATT estimand. In addition to the extra weighting of the exposure level, this difference to the doubly robust estimand in Sant'Anna and Zhao (2020) is due to the fixed attributes.

It is worth explaining what we mean by correct specification of the propensity scores and the conditional means of outcomes. The specification of $p(z_i)$ and $\mu_{it,w}(z_i)$ is more straightforward; the only difference from usual practice without interference is the choice of z_i, which may include neighbors' attributes. As for the specification of $\pi_{wg}(z_i)$ and $\mu_{it,wg}(z_i)$, it is easier to fix ideas using a simple example.

Suppose the spatial units are located on a square grid at locations $\{(d_1, d_2) : d_1, d_2 = 1, 2, \ldots, l\}$. Units immediately to the left or right of i are classified as neighbors of i. Each unit is assigned to treatment independently according to a Bernoulli trial with probability $p(z_i)$. The potential outcome function is $y_{it}(w_i, \boldsymbol{w}_{-i}) = w_i + A_i \boldsymbol{W} + e_i$, where A_i is the i^{th} row of the adjacency matrix and e_i is the standard normal independent of everything else. Nevertheless, the spillover function is misspecified as $G_i = \mathbb{1}\{A_i \boldsymbol{W} > 0\}$. I use z_i^*

and z_i to differentiate neighborhood attributes for i and individual attributes for i alone. In this example, $z_i^* = \{z_j : j \in \mathcal{N}_i\}$. Then $\pi_{1g}(z_i^*) = \pi_{0g}(z_i^*) = 1 - \Pi_{j\in\mathcal{N}_i}\big(1 - p(z_j)\big)$ and $\mu_{it,wg}(z_i^*) = w + g\Big(\sum_{j\in\mathcal{N}_i} p(z_j)\Big)/\big[2\big(1 - \Pi_{j\in\mathcal{N}_i}\big(1 - p(z_j)\big)\big)\big]$ for units with two neighbors. We hope to correctly specify $\pi_{wg}(z_i^*)$ and $\mu_{it,wg}(z_i^*)$ along with the correct spillover function. Nonetheless, even if the spillover function is misspecified, we might still be able to correctly specify the propensity scores and conditional expected potential outcomes at exposure g.

Analogously, the doubly robust estimands for the spillover effects are

$$\begin{aligned}&\mathbb{E}_D\Bigg[\frac{W_i}{\eta(z_i)}\frac{\mathbb{1}\{G_i = g\}}{\eta_{1g}(z_i)}\big(Y_{i2} - m_{i2,1g}(z_i)\big) + m_{i2,1g}(z_i)\\ &-\frac{W_i}{\eta(z_i)}\frac{\mathbb{1}\{G_i = g'\}}{\eta_{1g'}(z_i)}\big(Y_{i2} - m_{i2,1g'}(z_i)\big) - m_{i2,1g'}(z_i)\Bigg]\end{aligned} \tag{15}$$

and

$$\begin{aligned}&\mathbb{E}_D\Bigg[\frac{1 - W_i}{1 - \eta(z_i)}\frac{\mathbb{1}\{G_i = g\}}{\eta_{0g}(z_i)}\big(Y_{i2} - m_{i2,0g}(z_i)\big) + m_{i2,0g}(z_i)\\ &-\frac{1 - W_i}{1 - \eta(z_i)}\frac{\mathbb{1}\{G_i = g'\}}{\eta_{0g'}(z_i)}\big(Y_{i2} - m_{i2,0g'}(z_i)\big) - m_{i2,0g'}(z_i)\Bigg].\end{aligned} \tag{16}$$

4 Asymptotic Properties of the Parametric Estimator

I focus on the estimation of the EDATT since the estimation of spillover effects would be similar. I propose a GMM estimator combining equation (14) with moment conditions for the propensity scores and conditional means of outcomes chosen by the empirical researcher. To make our estimator more robust to misspecification of these functions, one can use various moment conditions to identify the propensity scores. One option is the covariate balancing propensity scores (CBPS), which can be locally more robust than the propensity scores based on maximum likelihood estimation (MLE); see, for instance, Imai and Ratkovic (2014). The alternative would be estimating all functions nonparametrically, which is left

as future work.

I denote the generic moment condition for the propensity scores as

$$\mathbb{E}_D\big[q_{i1}(W_i, z_i, \gamma_1^*)\big] = 0 \tag{17}$$

and

$$\mathbb{E}_D\big[q_{i2}(W_i, G_i, z_i, \gamma_2^*)\big] = 0, \tag{18}$$

where z_i can contain neighbors' attributes within K-neighborhood. For instance, the moment conditions for CBPS are

$$\mathbb{E}_D\left[\frac{W_i}{P(W_i = 1|z_i)}z_i - \frac{(1 - W_i)}{1 - P(W_i = 1|z_i)}z_i\right] = 0 \tag{19}$$

and for $g = 1, 2, \ldots, G$, $G - 1$

$$\mathbb{E}_D\left[\frac{\mathbb{1}\{G_i = g\}}{P(G_i = g|W_i, z_i)}(W_i, z_i) - \frac{\mathbb{1}\{G_i = g - 1\}}{P(G_i = g - 1|W_i, z_i)}(W_i, z_i)\right] = 0, \tag{20}$$

where $P(W_i = 1|z_i)$ is some probability for a binary response, such as $\frac{exp(z_i\gamma_1)}{1+exp(z_i\gamma_1)}$, and $P(G_i = g|W_i, z_i)$ is some probability for discrete choices. Similarly, the generic conditional moment conditions are denoted by

$$\mathbb{E}_D\big[q_{i3}(Y_{i1}, W_i, z_i, \gamma_3^*)\big] = 0 \tag{21}$$

and

$$\mathbb{E}_D\big[q_{i4}(Y_{i2}, W_i, G_i, z_i, \gamma_4^*)\big] = 0. \tag{22}$$

Alternatively, one can model the conditional mean for $\Delta Y_i = Y_{i2} - Y_{i1}$ and formulate the

moment condition as

$$\mathbb{E}_D\big[\tilde{q}_{i3}(\Delta Y_i, W_i, G_i, z_i, \tilde{\gamma}_3^*)\big] = 0. \tag{23}$$

If there are only a few possible values that the exposure levels G_i can take, one can alternatively model the conditional outcomes for the subpopulation with $W_i = w$ and $G_i = g$ as a function of z_i, separately. Lastly, the moment condition for $\tau(g)$ is a restatement of equation (14)[1]. Denote $\theta_M^* = (\gamma_1^{*\prime}, \gamma_2^{*\prime}, \gamma_3^{*\prime}, \gamma_4^{*\prime}, \tau(g))'$.

$$\begin{aligned}
&\mathbb{E}_D\big[q_{i5}(Y_{it}, W_i, G_i, z_i, \theta_M^*)\big] \\
=&\mathbb{E}_D\Bigg[\frac{W_i}{\eta(z_i)}\bigg(\frac{\mathbb{1}\{G_i = g\}}{\eta_{1g}(z_i)}\big(Y_{i2} - m_{i2,1g}(z_i)\big) - \big(Y_{i1} - m_{i1,1}(z_i)\big)\bigg) \\
&- \frac{1 - W_i}{1 - \eta(z_i)}\bigg(\frac{\mathbb{1}\{G_i = g\}}{\eta_{0g}(z_i)}\big(Y_{i2} - m_{i2,0g}(z_i)\big) - \big(Y_{i1} - m_{i1,0}(z_i)\big)\bigg) \\
&+ \Delta m_{i2,g}(z_i) - \Delta m_{i1}(z_i) - \tau(g)\Bigg] = 0
\end{aligned} \tag{24}$$

Let $X_i = \{Y_i, W_i, G_i, z_i\}$, $q_i(X_i, \theta) = (q_{i1}'(\gamma_1), q_{i2}'(\gamma_2), q_{i3}'(\gamma_3), q_{i4}'(\gamma_4), q_{i5}(\theta))'$, and $\widehat{\Psi}$ as the weighting matrix with dimensions larger or equal to that of θ.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{|D_M|} \sum_{i \in D_M} q_i(X_i, \theta)' \widehat{\Psi} \frac{1}{|D_M|} \sum_{i \in D_M} q_i(X_i, \theta) \tag{25}$$

The GMM estimator is the solution to the finite population minimization problem in equation (25). And the estimator of $\tau(g)$ is the last element of $\hat{\theta}$.

I impose the following assumptions to study the asymptotic distribution of the GMM estimator.

[1] In practice, it is recommended to normalize the weights for IPW type of estimators. Changing the moment condition with normalized propensity scores – where the weights sum to unity – does not affect asymptotic normality of the GMM estimator. In fact, estimators with normalized weights consistently show better finite sample performance in the simulations below.

Assumption 4 *Suppose* $\{D_M\}$ *is a sequence of finite subsets of* D *such that* $|D_M| \to \infty$ *as* $M \to \infty$*, where the lattice* $D \subseteq \mathbb{R}^d$*,* $d \geq 1$*, is infinitely countable. All elements in* D *are located at distances of at least* $\rho_0 > 0$ *from each other, i.e., for all* $i, j \in D$*:* $\rho(i,j) \geq \rho_0$*; w.l.o.g. we assume that* $\rho_0 > 1$*.*

Consistent with the increasing domain asymptotics, the assumption of the minimum distance ensures the expansion of the finite population region.

Assumption 5 *(Approximate Neighborhood Interference) Let* $\boldsymbol{W}^{(i,s)} = (\boldsymbol{W}_{\mathcal{N}(i,s)}, \boldsymbol{W}'_{D_M \setminus \mathcal{N}(i,s)})$*, where* $\boldsymbol{W}'$ *is an independent copy of* $\boldsymbol{W}$*,* $\boldsymbol{W}^{(i,s,0)} = (\boldsymbol{W}_{\mathcal{N}(i,s)}, \underline{0})$*, i.e.,* $\boldsymbol{W}'_{D_M \setminus \mathcal{N}(i,s)} = \underline{0}$*, and*

$$\kappa_M(s) = \max_{i \in D_M} \mathbb{E}\Big[y_{i2}(\boldsymbol{W}) - y_{i2}\big(\boldsymbol{W}^{(i,s,0)}\big) \big| \boldsymbol{z} \Big].$$

Suppose that $\sup_M \kappa_M(s) \to 0$ *as* $s \to \infty$*.*

Assumption 5 is a modified version of Assumption 4 in Leung (2022). Essentially, treatments of units from s distance away from i should become minimal as the distance s gets larger. This way, we can allow interference outside the immediate K-neighborhood while still being able to derive the asymptotic properties of our estimator. Leung (2022) has shown that several interference structures satisfy the ANI assumption, including the linear-in-means model with endogenous peer effects. Section 5 gives an overview of the different approaches to model interference taken by the literature and compares them to ANI.

I adopt ψ-dependence in Kojevnikov, Marmer, and Song (2021) as our notion of weak dependence throughout the paper. Notice that α-mixing is a special case of ψ-dependence. Let $\mathcal{L}_{\nu,h}$ denote the collection of bounded Lipschitz real functions $f(\cdot)$ on $\mathbb{R}^{\nu \times h}$ with the Lipschitz constant $\text{Lip}(f) < \infty$ and $\|f\|_\infty < \infty$, where $\|f\|_\infty = \sup_x |f(x)|$. Denote the

collection of subset pairs as

$$\mathcal{P}_M(h,h';s)=\{(H,H'):H,H'\subseteq D_M,|H|=h,|H'|=h',\rho(H,H')\geq s\}.$$

Definition 1 *A triangular array* $\{V_i,i\in D_M,M\geq 1\},V_i\in\mathbb{R}^\nu$, *is called* ψ*-dependent if there exist uniformly bounded constants* $\{\tilde{\kappa}_{M,s}\}_{s\geq 0}$ *with* $\tilde{\kappa}_{M,0}=1$, *and a collection of nonrandom functions* $\{\psi_{h,h'}\}_{h,h'\in\mathbb{N}}$ *with* $\psi_{h,h'}:\mathcal{L}_{\nu,h}\times\mathcal{L}_{\nu,h'}\to[0,\infty)$ *such that for all* $(H,H')\in\mathcal{P}_M(h,h';s)$ *with* $s>0$ *and all* $f\in\mathcal{L}_{\nu,h}$ *and* $f'\in\mathcal{L}_{\nu,h'}$,

$$\big|Cov\big(f(V_H),f'(V_{H'})\big)\big|\leq\psi_{h,h'}(f,f')\tilde{\kappa}_{M,s},\tag{26}$$

where $V_H=(V_i:i\in H)$.

I require $\tilde{\kappa}_{M,s}$ to approach zero as s grows. ψ-dependence is used to bound the covariances of any two subsets of observations distant from each other.

Assumption 6 *Let* $y_{it}=\phi(W_i,\boldsymbol{W}_{-i},z_i,U_i)$, *where* $\phi(\cdot)$ *is some generic function and* U_i *denotes the unobservables. Let* $\epsilon_i=(W_i,U_i)$. *The random field* $\epsilon=\{\epsilon_i,i\in D_M,M\geq 1\}$ *is* α*-mixing under Definition 2 in Jenish and Prucha (2012). The mixing coefficient is denoted by* $\alpha^\epsilon(u,v,r)\leq(u+v)\widehat{\alpha}^\epsilon(r)$.

On top of possible interference, Assumption 6 allows assignment variables to be spatially correlated as well.

Lemma 4.1 *Under Assumptions 4, 5, 6, and Assumption A.1 in Appendix A, for each* $\theta\in\Theta$, *each element of* $q_i(X_i,\theta)$ *and* $\nabla_\theta q_i(X_i,\theta)$ *is* ψ*-dependent with* $\tilde{\kappa}_{M,s}=\big(\kappa_M(s/3)+s^d\widehat{\alpha}^\epsilon_M(s/3)\big)\mathbb{1}(s>3\max\{K,1\})+\mathbb{1}(s\leq 3\max\{K,1\})$. *Equation (26) holds with* $h=h'=1$ *and* $f=f'$ *being the identity function.*

To adapt the limit theorems in Kojevnikov et al. (2021) to spatial data, I replace the network denseness with the cardinality of the spatial sets implied by Lemma A.1 in Jenish and Prucha (2009). As a result, Assumption 3.2 in Kojevnikov et al. (2021) is modified as

Assumption 7

$$\sum_{s=1}^{\infty} s^{d-1}\tilde{\kappa}_{M,s} < \infty$$

Assumption 7 is in the similar spirit of Assumption 3(b) in Jenish and Prucha (2009) for α-mixing random fields.

Let $\sigma_M^2 = \mathbb{V}\big[\sum_{i\in D_M}\lambda' q_i(U_i,\theta)|\boldsymbol{z}\big]$ for a nonzero vector λ. Similarly, Assumption 3.4 in Kojevnikov et al. (2021) is modified as

Assumption 8 *There exists a positive sequence $r_M \to \infty$ such that for $k = 1, 2$*

$$\frac{1}{\sigma_M^{2+k}}\sum_{i\in D_M}\sum_{s=1}^{\infty} s^{d-1}\max_{j\in D_M, s\le\rho(i,j)<s+1}\big|\mathcal{N}(i;r_M)\setminus\mathcal{N}(j;s-1)\big|^k\tilde{\kappa}_{M,s}^{1-\frac{2+k}{p}} \to 0$$

and

$$\frac{|D_M|^2\tilde{\kappa}_{M,r_M}^{1-(1/p)}}{\sigma_M} \to 0$$

as $M\to\infty$, where $p > 4$ is that appears in Assumption A.1 in Appendix A.

The rate of $\tilde{\kappa}_{M,s}$ is implicitly implied by Assumption 8. A sufficient condition for the first part of the assumption is

$$\frac{|D_M|}{\sigma_M^{2+k}} r_M^{kd}\sum_{s=1}^{\infty} s^{d-1}\tilde{\kappa}_{M,s}^{1-\frac{2+k}{p}} \to 0,$$

Analogous conditions – equations (B.18) and (B.19) – can be found in Jenish and Prucha (2009). Leung (2022) provides an example data generating process of spatial networks that satisfies Assumption 8.

Define

$$\Omega_M = \Delta_{ehw,M} + \Delta_{spatial,M} - \Delta_{E,M} - \Delta_{ES,M}, \tag{27}$$

where

$$\Delta_{ehw,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\big[q_i(X_i, \theta_M^*) q_i(X_i, \theta_M^*)' | \boldsymbol{z}\big], \tag{28}$$

$$\Delta_{E,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\big[q_i(X_i, \theta_M^*) | \boldsymbol{z}\big] \mathbb{E}\big[q_i(X_i, \theta_M^*) | \boldsymbol{z}\big]', \tag{29}$$

$$\Delta_{spatial,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M, j \neq i} \mathbb{E}\big[q_i(X_i, \theta_M^*) q_j(X_j, \theta_M^*)' | \boldsymbol{z}\big], \tag{30}$$

$$\Delta_{ES,M} = \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M, j \neq i} \mathbb{E}\big[q_i(X_i, \theta_M^*) | \boldsymbol{z}\big] \mathbb{E}\big[q_j(X_j, \theta_M^*) | \boldsymbol{z}\big]'. \tag{31}$$

Denote

$$R_M^* = \mathbb{E}_D\big[\nabla_\theta q_i(X_i, \theta_M^*)\big]$$

and

$$V_M = \big(R_M^{*\prime} \Psi_M R_M^*\big)^{-1} R_M^{*\prime} \Psi_M \Omega_M \Psi_M R_M^* \big(R_M^{*\prime} \Psi_M R_M^*\big)^{-1}, \tag{32}$$

where $\widehat{\Psi} - \Psi_M \xrightarrow{p} \mathbf{0}$.

Theorem 4.2 *Under Assumptions 1-8, and Assumption A.1 in Appendix A, if either equations (4)-(6) or equations (12) and (13) are correctly modeled,*

$$V_M^{-1/2} \sqrt{|D_M|} (\hat{\theta} - \theta_M^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, I_k).$$

Let us compare Ω_M with the middle term of the variance-covariance matrix in Xu and

Wooldridge (2022):[2]

$$\begin{aligned} S_M =& \Delta_{ehw,M}(\theta_M^*) + \rho_{uM}\Delta_{cluster,M}(\theta_M^*) + \rho_{uM}\rho_{cM}\Delta_{spatial,M}(\theta_M^*) \\ & - \rho_{uM}\rho_{cM}\Delta_{E,M} - \rho_{uM}\rho_{cM}\Delta_{EC,M} - \rho_{uM}\rho_{cM}\Delta_{ES,M} \end{aligned} \tag{33}$$

Ω_M echos S_M without explicit cluster partition. The key difference is that the composite sampling probabilities $\rho_{uM}\rho_{cM}$ are equal to one, since we acquire the entire population here to estimate the population spillover effect and the direct treatment effect. In addition, the extra terms, $\Delta_{E,M}$ and $\Delta_{ES,M}$ are only conditional on observed attributes but not potential outcomes. According to the guidance in Xu and Wooldridge (2022), with the consideration of interference, we need to make inference robust to spatial correlation.

As a common approach to adjust the variance estimator for spatial correlation, the usual spatial heteroskedasticity and autocorrelation consistent (SHAC) variance estimator is defined as

$$\tilde{V} = \big(\hat{R}'\widehat{\Psi}\hat{R}\big)^{-1}\hat{R}'\widehat{\Psi}\tilde{\Omega}(\hat{\theta})\widehat{\Psi}\hat{R}\big(\hat{R}'\widehat{\Psi}\hat{R}\big)^{-1},$$

where

$$\hat{R} = \frac{1}{|D_M|}\sum_{i\in D_M}\nabla_\theta q_i(X_i,\hat{\theta})$$

and

$$\tilde{\Omega}(\theta) = \frac{1}{|D_M|}\sum_{s=0}^{\infty}\omega\Big(\frac{s}{b_M}\Big)\sum_{i\in D_M}\sum_{j\in D_M, s\le\rho(i,j)<s+1} q_i(X_i,\theta)q_j(X_j,\theta)'.$$

I impose the following assumption for the estimation of the variance-covariance matrix.

Assumption 9 *The weights satisfy:*

(i) $\omega(0) = 1$, $\omega\big(\frac{s}{b_M}\big) = 0$ *for any* $s > b_M$, $\big|\omega\big(\frac{s}{b_M}\big)\big| < \infty$, $\forall$ M*;*

[2]Please see Xu and Wooldridge (2022) for detailed explanation of notation.

(ii)

$$\sum_{s=1}^{\infty} \Big|\omega\Big(\frac{s}{b_M}\Big) - 1\Big| s^{d-1} \tilde{\kappa}_{M,s}^{1-2/p} \to 0;$$

(iii)

$$\frac{1}{|D_M|^2} \sum_{i \in D_M} \sum_{s=1}^{\infty} s^{d-1} \max_{j \in D_M, s \leq \rho(i,j) < s+1} \big|\mathcal{N}(i; b_M)\big|^2 \tilde{\kappa}_{M,s}^{1-4/p} \to 0$$

as $M \to \infty$*, where* $b_M = o\big(|D_M|^{1/2d}\big)$ *and* $p > 4$ *is that appears in Assumption A.1 in Appendix A.*

Assumption 9(ii) is a high-level condition, which requires that the kernel weights $\omega\big(\frac{s}{b_M}\big)$ converge to one sufficiently fast as $M \to \infty$. Assumption 9(iii) regulates the growth rate of the bandwidth $\{b_M\}$.

Theorem 4.3 *Under Assumptions 4-9, and Assumption A.1 in Appendix A,*

$$\tilde{V} - (V_M + V_E) \xrightarrow{p} \boldsymbol{0},$$

where

$$V_E = \big(R_M^{*\prime} \Psi_M R_M^*\big)^{-1} R_M^{*\prime} \Psi_M \Omega_E \Psi_M R_M^* \big(R_M^{*\prime} \Psi_M R_M^*\big)^{-1}$$

and

$$\Omega_E = \frac{1}{|D_M|} \sum_{s=0}^{\infty} \omega\big(\frac{s}{b_M}\big) \sum_{i \in D_M} \sum_{j \in D_M, s \leq \rho(i,j) < s+1} \mathbb{E}\big[q_i(X_i, \theta_M^*) | \boldsymbol{z}\big] \mathbb{E}\big[q_j(X_j, \theta_M^*) | \boldsymbol{z}\big]'.$$

Remark 1 *The usual SHAC variance estimator is generally conservative for the conditional SHAC variance-covariance matrix.*

Remark 1 is similar to Remark 3 in Xu and Wooldridge (2022) without sampling consideration. The conservativeness of the usual variance estimator for conditional variance has

also been investigated in Abadie et al. (2014) under the independence assumption for the heteroskedasticity-robust variance matrix. I extend it to the case with spatial correlation here.

That said, I would like to highlight a few points. First, because $\tilde{\Omega}(\hat{\theta})$ is a conservative estimator for Ω_M, even if we choose Ψ_M as the optimal weighting matrix Ω_M^{-1}, using $\widehat{\Psi} = \tilde{\Omega}(\hat{\theta})$ in estimation is not going to achieve the most efficient GMM estimator. The usual variance estimator is therefore conservative not only because of the neglect of the additional terms in the variance-covariance matrix but also because the optimal weighting matrix is not consistently estimated. Of course, when the model is just identified, the weighting matrix choice is irrelevant.

Second, unlike the finite population variance-covariance matrix in Xu and Wooldridge (2022), the conditional SHAC variance matrix is consistently estimable because it is no longer conditional on the unobserved potential outcomes. There are different approaches we can take. However, since the usual SHAC variance estimator is known to suffer from downward bias especially when the spatial correlation is high, it is not always necessary to estimate the smaller conditional variance matrix.

5 Different Approaches to Dimension Reduction

Manski (2013) and Basse and Airoldi (2018) formally point out that there exist no consistent treatment effect estimators under arbitrary interference. It is therefore necessary to make dimension reduction assumptions about the interference structure in order to identify meaningful treatment effect parameters. There are different approaches to dimension reduction in the literature; see, for instance, Auerbach and Tabord-Meehan (2021), Agarwal, Cen, Shah, and Yu (2022), Emmenegger, Spohn, and Bühlmann (2022), and Qu, Xiong, Liu, and Imbens (2022). In this paper, I provide an overview of some of the leading

approaches in the literature and show how the recent literature development relates to our general framework. Each article referenced proposes different estimation methods for various causal effect estimands. Our focus here is to compare the different approaches to modeling spillover effect.[3]

5.1 Partial Interference

The most popular approach to dimension reduction of the interference structure is partial interference restricted within disjoint clusters. In Qu et al. (2022), their potential outcome function is modeled as[4]

$$y_{c,i}(w_{c,i}, \boldsymbol{w}_{c,(i),1}, \cdots, \boldsymbol{w}_{c,(i),m}) \equiv y_{c,i}(w_{c,i}, g_{c,1}, \cdots, g_{c,m}), \tag{34}$$

where c is the index of a cluster, $y_{c,i}$ and $w_{c,i}$ is the potential outcome and treatment assignment of unit i in cluster c, and $\boldsymbol{w}_{c,(i),j}$ is the treatment assignment of unit i's neighbors in the disjoint subset j of cluster c. Units within each of the m disjoint subsets are exchangeable. As a result, the impact of $\boldsymbol{w}_{c,(i),j}$ can be summarized by $g_{c,j}$, which measures the number of treated neighbors in subset j of cluster c. Compared with the assumption of fully exchangeable neighbors in cluster c, the partition of m subsets allows for more heterogeneity of neighbors' influence and hence a more flexible interference structure.

If (34) is correctly specified, we can choose K to be $\max_{c=1,\ldots,C} \max_{i,j \in c} \rho(i,j)$. Given bounded cluster sizes, K is finite. For all $s > K$ and any i, $y_i(\boldsymbol{W}) - y_i(\boldsymbol{W}^{i,s}) = 0$. Therefore, potential outcomes in the form of (34) can be accommodated in the approach we take. A trickier question is how to partition the m subsets within each cluster c. On top of that, partial interference might be too strong an assumption. If either the exchangeability or the

[3] It is not supposed to be a comprehensive survey.

[4] The potential outcome is defined for a single cross section.

partial interference assumption does not hold, our approach can still identify the expected exposure effect as long as the interference from units further away is increasingly negligible.

5.2 Immediate Neighbors

A slightly different approach to dimension reduction is to restrict interference within immediate neighbors. For instance, in Emmenegger et al. (2022), the spillover function is specified as

$$\left(f^1(\{w_j\}_{j\in D_M, j\neq i}), \cdots, f^r(\{w_j\}_{j\in D_M, j\neq i})\right) \tag{35}$$

of fixed dimensions r. Each such function is specified by empirical researchers and describes a one-dimensional spillover effect that unit i receives from its neighbors. In Example 2.1 in Emmenegger et al. (2022), the functions f^l has been specified as the average number of treated neighbors of unit i and the average number of treated neighbors of neighbors of i, respectively, for $r = 2$. In this case, if we define neighbors of i as units within distance $\bar{K}$ from i, then ANI holds for any $s > 2\bar{K}$.

In Agarwal et al. (2022), they impose network SUTVA, i.e., the potential outcome only depends on treatment assigned to the unit's neighbors but not other units outside the neighborhood. In addition, they assume interference is additive across neighbors. Their potential outcome is therefore modeled as

$$y_{it}(\{w_j\}_{j\in\mathcal{N}(i,K)}) = \sum_{k\in\mathcal{N}(i,K)} \boldsymbol{u}'_{k,i}\boldsymbol{l}_{w_k,t} + \epsilon_{it}, \tag{36}$$

where $\boldsymbol{u}$ and $\boldsymbol{l}$ represents r-dimensional latent factors and ϵ is the error term. Factor analysis is apparently different from our estimation methods. Nevertheless, ANI holds for any $s > K$.

Equations (34) and (35) have recently been proposed in the literature allowing for a

more flexible interference structure. The purpose of the discussion is to show that if empirical researchers assume these specifications of the spillover function are correct, they can be well accommodated in our framework. Even if some dimension reduction assumptions fail, applied researchers are still able to identify causal estimands as long as ANI is true.

5.3 Local Configuration

A more interesting discussion is the comparison of the local configuration approach proposed by Auerbach and Tabord-Meehan (2021) and ANI. In a spatial setting, unit i's local configuration of radius r, denoted by G_i^r, refers to the units within the distance r of i and their characteristics. Units within a local configuration remain anonymous, similar to the exchangeability assumption. ANI and the expected exposure mapping are initially proposed to allow for misspecification of the spillover function. The local configuration approach instead assumes the correct specification of the spillover function, but uses local configurations of various radiuses r to approximate the effective treatment according to the spillover function. Below, I provide another interpretation of the ANI assumption. Under correct specification of the spillover function, the ANI approach is not too different from the local configuration approach.

According to the metric definition in Auerbach and Tabord-Meehan (2021), for effective treatment g and $\tilde{g}$, if $d(g,\tilde{g}) \leq \frac{1}{1+r}$ then $G_i^r = \tilde{G}_i^r$. Under Assumption 4.5 there,

$$\big|h(g_0) - h(\tilde{g})\big| \leq \phi\big(d(g_0,\tilde{g})\big), \tag{37}$$

where $\phi(x) \to 0$ as $x \to 0$, $h(g) = \mathbb{E}[h(g,U_i)]$, and $Y_i = h(G_i,U_i)$. Therefore, we can see that (37) goes to 0 as $r \to \infty$, which is analogous to the ANI assumption in Leung (2022).

$$\sup_M \max_{i\in D_M} \mathbb{E}\Big[\big|Y_i(\boldsymbol{W}) - Y_i\big(\boldsymbol{W}^{(i,r)}\big)\big|\Big] \to 0, \text{ as } r \to \infty \tag{38}$$

Examples 2.1 and 2.2 in Auerbach and Tabord-Meehan (2021) are essentially examples of Sections 5.1 and 5.2, and hence I focus on their Example 2.3 – the linear-in-means peer effects model. Assuming correct specification,

$$Y_i = \alpha + \delta \frac{1}{n_i} \sum_{j \in P_i} Y_j + W_i \gamma + e_i,$$

where P_i is the peer group of unit i with size n_i. As usual, $|\delta| < 1$. The reduced form of the potential outcome is solved to be

$$Y_i = \lim_{S \to \infty} \sum_{s=1}^{S} h_s(G_i^s, U_i) = h(G_i, U_i).$$

for some functions h_s and h. Hence, for $d(g, \tilde{g}) \leq \frac{1}{1+r}$,

$$\left|h(g) - h(\tilde{g})\right| \leq C|\beta|^r \text{ for some } |\beta| < 1,$$

which is exactly the ANI coefficient given in Proposition 1 in Leung (2022).[5]

Therefore, under correct specification of the spillover function, if we choose a large enough r neighborhood, the ANI approach can be thought of as using the units with the effective treatment closest to the actual effective treatment g to estimate the policy effect.

6 Multiple Time Periods with Common Treatment Timing

Extension to multiple time periods is straightforward. With common treatment timing, the simplest approach is to aggregate the time periods prior to and post treatment into a single time period, again denoted $t = 1, 2$. With the aggregated data, we can directly apply

[5] I refer the readers to Auerbach and Tabord-Meehan (2021) for the introduction to notation and more detailed derivation.

the results above. Alternatively, we might be interested in the EDATT at different time periods. Denote the time periods by $\{-\underline{T}, \ldots, -1, 0, 1, \ldots, \overline{T}\}$. Without loss of generality, suppose treatment starts at $t = 2$. For any $t \geq 2$, the EDATT at time period t at exposure level g is defined as

$$\tau_t(g) = \frac{1}{|D_M|} \sum_{i \in D_M} \mathbb{E}\big[y_{it}(1, \boldsymbol{w}_{-i}) - y_{it}(0, \boldsymbol{w}_{-i}) | W_i = 1, G_i = g, z_i\big] \tag{39}$$

Spillover effects at time period t can be defined analogously.

It is worth discussing different ways to formalize the parallel trends assumption. We can either pick one time period before treatment, say $t = 1$, as the comparison time period. Or, we can use the average potential outcomes across the time periods prior to treatment as a comparison. The latter can potentially improve efficiency since data from more time periods are used in estimation. On the other hand, if the parallel trends assumption only holds for the time periods closest to the treatment period, the second approach is less robust. Hence, there is a typical robustness and efficiency tradeoff.

Other than the slight modification of the estimands of interest, the estimation and asymptotic properties remain the same as long as we contrast the appropriate time periods, for instance, using data from any $t \geq 2$ and $t = 1$. This way, we can estimate the dynamic treatment effects when the duration under treatment progresses.

7 Conclusion

I propose doubly robust estimators for the expected direct treatment effect and spillover effect in a DID context. Our approach is general in the sense that misspecification of exposure mapping is allowed and interference is not restricted within a fixed boundary of neighborhoods. Given arbitrary spillover effect, one needs to account for spatial correla-

tion when conducting inference. With the entire population observed, the usual spatial correlation robust variance estimator could be conservative.

If one is interested in estimating the spillover effect in the sample or the spillover is restricted within clusters, the current framework can be extended to incorporate sampling from a finite population, which is the setup adopted by Xu and Wooldridge (2022). With sampling, we need to consider pooled cross sections along with panel data, which are the two types of datasets DID can be applied to. Another difference would be inference, since now sampling probabilities also play a role.

Since the limit theorems from Kojevnikov et al. (2021) are applied to derive the asymptotic distribution, our analysis can be extended to DID with interference in network data. Given the inclusion of neighbors' treatments and attributes in the propensity score and the conditional mean functions, nonparametric estimation is attractive to allow for arbitrary functional forms. This is left as future work. In a follow-up research, I extend the framework to multiple time periods with uncommon timing of treatment adoption.

References

Abadie, A. (2005), Semiparametric difference-in-differences estimators. *The review of economic studies* 72(1), 1–19.

Abadie, A., Imbens, G.W., and Zheng, F. (2014), Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association* 109(508), 1601–1614.

Agarwal, A., Cen, S., Shah, D., and Yu, C.L. (2022), Network synthetic interventions: A framework for panel data with network interference. Tech. rep., arXiv preprint arXiv:2210.11355.

Arkhangelsky, D., Imbens, G.W., Lei, L., and Luo, X. (2021), Double-robust two-way-fixed-effects regression for panel data. Tech. rep., arXiv preprint arXiv:2107.13737.

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

Athey, S. and Imbens, G.W. (2022), Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics* 226(1), 62–79.

Auerbach, E. and Tabord-Meehan, M. (2021), The local approach to causal inference under network interference. Tech. rep., arXiv preprint arXiv:2105.03810.

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

Butts, K. (2021), Difference-in-differences estimation with spatial spillovers. Tech. rep., arXiv preprint arXiv:2105.03737.

Clarke, D. (2017), Estimating difference-in-differences in the presence of spillovers. Tech. rep., MPRA Paper No. 81604.

Delgado, M.S. and Florax, R.J. (2015), Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction. *Economics Letters* 137, 123–126.

Di Tella, R. and Schargrodsky, E. (2004), Do police reduce crime? estimates using the allocation of police forces after a terrorist attack. *American Economic Review* 94(1), 115–133.

Emmenegger, C., Spohn, M.L., and Bühlmann, P. (2022), Treatment effect estimation from observational network data using augmented inverse probability weighting and machine learning. Tech. rep., arXiv preprint arXiv:2206.14591.

Forastiere, L., Airoldi, E.M., and 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.

Gallant, A.R. and White, H. (1988), *A unified theory of estimation and inference for nonlinear dynamic models*. Blackwell.

Ghanem, D., Sant'Anna, P.H., and Wüthrich, K. (2022), Selection and parallel trends. Tech. rep., arXiv preprint arXiv:2203.09001.

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

Imai, K. and Ratkovic, M. (2014), Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 243–263.

Jardim, E., Long, M.C., Plotnick, R., Van Inwegen, E., Vigdor, J., and Wething, H. (2022), Minimum-wage increases and low-wage employment: Evidence from seattle. *American Economic Journal: Economic Policy* 14(2), 263–314.

Jenish, N. and Prucha, I.R. (2009), Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of econometrics* 150(1), 86–98.

Jenish, N. and Prucha, I.R. (2012), On spatial processes and asymptotic inference under near-epoch dependence. *Journal of Econometrics* 170(1), 178–190.

Jin, Y. and Rothenhäusler, D. (2023), Tailored inference for finite populations: conditional validity and transfer across distributions. *Biometrika* p. asad022.

Kojevnikov, D., Marmer, V., and Song, K. (2021), Limit theorems for network dependent random variables. *Journal of Econometrics* 222(2), 882–908.

Leung, M.P. (2022), Causal inference under approximate neighborhood interference. *Econometrica* 90(1), 267–293.

Man, Y., Sant'Anna, P.H., Sasaki, Y., and Ura, T. (2023), Doubly robust estimators with weak overlap. Tech. rep., arXiv preprint arXiv:2304.08974.

Manski, C.F. (1993), Identification of endogenous social effects: The reflection problem. *The review of economic studies* 60(3), 531–542.

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

Newey, W.K. (1991), Uniform convergence in probability and stochastic equicontinuity. *Econometrica* 59, 1161–1167.

Newey, W.K. and McFadden, D. (1994), Large sample estimation and hypothesis testing. *Handbook of Econometrics* 4, 2111–2245.

Ning, Y., Peng, S., and Tao, J. (2020), Doubly robust semiparametric difference-in-differences estimators with high-dimensional data. Tech. rep., arXiv preprint arXiv:2009.03151.

Qu, Z., Xiong, R., Liu, J., and Imbens, G. (2022), Efficient treatment effect estimation in observational studies under heterogeneous partial interference. Tech. rep., arXiv preprint arXiv:2107.12420.

Rambachan, A. and Roth, J. (2022), Design-based uncertainty for quasi-experiments. Tech. rep., arXiv preprint arXiv:2008.00602.

Rambachan, A. and Roth, J. (2023), A more credible approach to parallel trends. *Review of Economic Studies* p. rdad018.

Roth, J. (2022), Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights* 4(3), 305–22.

Roth, J. and Sant'Anna, P.H. (2023), When is parallel trends sensitive to functional form? *Econometrica* 91(2), 737–747.

Sant'Anna, P.H. and Zhao, J. (2020), Doubly robust difference-in-differences estimators. *Journal of Econometrics* 219(1), 101–122.

Sävje, F., Aronow, P., and Hudgens, M. (2021), Average treatment effects in the presence of unknown interference. *Annals of statistics* 49(2), 673.

Xu, R. and Wooldridge, J.M. (2022), A design-based approach to spatial correlation. Tech. rep., arXiv preprint arXiv:2211.14354.

A Regularity Conditions for the GMM Estimator

Definition 2 *The random function $g_i(X_i, \theta)$ is said to be Lipschitz in the parameter θ on Θ if there is $h(u) \downarrow 0$ as $u \downarrow 0$ and $b(\cdot) : \mathcal{W} \to R$ such that $\sup_{M, i \in D_M} \mathbb{E}\big[|b_i(X_i)|\big] < \infty$, and for all $\tilde{\theta}, \theta \in \Theta$, $\big|g_i(X_i, \tilde{\theta}) - g_i(X_i, \theta)\big| \leq b_i(X_i) h(\|\tilde{\theta} - \theta\|)$, $i \in D_M, M \geq 1$.*

Assumption A.1 *(i) $\widehat{\Psi} - \Psi_M \xrightarrow{p} \boldsymbol{0}$, where Ψ_M is positive semidefinite; (ii) Θ is compact; (iii) let $Q_M(\theta) = \mathbb{E}_D\big[q_i(X_i, \theta)\big]' \Psi \mathbb{E}_D\big[q_i(X_i, \theta)\big]$. $\{Q_M(\theta)\}$ has identifiably unique minimizers $\{\theta_M^*\}$ on Θ as in Definition 3.2 in Gallant and White (1988); (iv) $q_i(X_i, \theta)$*

is continuously differentiable on $int(\Theta)$, $\forall\ i, M$; *(v)* $q_i(X_i, \theta)$ *is Lipschitz in* θ *on* Θ; *(vi)* $\sup_{M,i\in D_M} \mathbb{E}\Big[\sup_{\theta\in\Theta} \|q_i(X_i,\theta)\|^p \,\Big|\boldsymbol{z}\Big] < \infty$ *for some* $p > 4$; *(vii)* $\theta^*_M \in int(\Theta)$ *uniformly in* M, *and* $\mathbb{E}_D\big[q_i(X_i, \theta^*_M)\big] = 0$; *(viii)* $\inf_M \lambda_{min}(\Omega_M) > 0$, *where* $\lambda_{min}(\cdot)$ *is the smallest eigenvalue; (ix)* $\nabla_\theta q_i(X_i, \theta)$ *is Lipschitz in* θ *on* Θ; *(x)* $\sup_{M,i\in D_M} \mathbb{E}\Big[\sup_{\theta\in\Theta} \|\nabla_\theta q_i(X_i,\theta)\|^2 \,\Big|\boldsymbol{z}\Big] < \infty$; *(xi)* ${R^*_M}'\Psi_M R^*_M$ *is nonsingular; (xii) for each* $\theta \in \Theta$, *let* $f_i(X_i, \theta)$ *be a generic function standing for each element of either* $q_i(X_i, \theta)$ *or* $\nabla_\theta q_i(X_i, \theta)$. $f_i(X_i, \theta)$ *is Lipschitz in* X_i *on the domain of* X_i *such that* $\sup_{M,i\in D_M} Lip(f_i) < \infty$ *and* $\sup_{M,i\in D_M} \|f_i\|_\infty < \infty$.

Notice that a necessary condition for Assumption A.1(xii) is $\sup_{M,i\in D_M} |Y_{it}| \le C < \infty$ and $\sup_{M,i\in D_M} \|z_i\| \le C < \infty$, which can often imply Assumption A.1(vi) and (x). Hence, in the proofs below, I maintain the assumption that Y_{it} and z_i are bounded.

B Proofs

Proof of Proposition 3:

Identification of the doubly robust estimand:

When the propensity scores are correctly specified, $\eta(z) = p(z)$, $\eta_{1g}(z) = \pi_{1g}(z)$, and $\eta_{0g}(z) = \pi_{0g}(z)$.

$$
\begin{aligned}
&\mathbb{E}\left[\frac{W_i}{p(z_i)}\left(\frac{\mathbb{1}\{G_i = g\}}{\pi_{1g}(z_i)}\big(Y_{i2} - m_{i2,1g}(z_i)\big) - \big(Y_{i1} - m_{i1,1}(z_i)\big)\right)\middle| z_i\right] \\
=&\mathbb{E}\left[\frac{W_i}{p(z_i)}\left(\frac{\mathbb{1}\{G_i = g\}}{\pi_{1g}(z_i)}\big(Y_{i2} - m_{i2,1g}(z_i)\big) - \big(Y_{i1} - m_{i1,1}(z_i)\big)\right)\middle| z_i, W_i = 1\right] P(W_i = 1|z_i) \\
=&\mathbb{E}\left[\frac{\mathbb{1}\{G_i = g\}}{\pi_{1g}(z_i)}\big(Y_{i2} - m_{i2,1g}(z_i)\big)\middle| z_i, W_i = 1, G_i = g\right] P(G_i = g|W_i = 1, z_i) \\
&- \mathbb{E}\big(Y_{i1} - m_{i1,1}(z_i)\big)\big| z_i, W_i = 1\big) \\
=&\mathbb{E}(Y_{i2}|z_i, W_i = 1, G_i = g) - \mathbb{E}(Y_{i1}|z_i, W_i = 1) - \big[m_{i2,1g}(z_i) - m_{i1,1}(z_i)\big] \qquad \text{(B.1)}
\end{aligned}
$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1-W_i}{1-p(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\pi_{0g}(z_i)}\big(Y_{i2}-m_{i2,0g}(z_i)\big)-\big(Y_{i1}-m_{i1,0}(z_i)\big)\right)\middle| z_i\right] \\
=&\mathbb{E}(Y_{i2}|z_i,W_i=0,G_i=g)-\mathbb{E}(Y_{i1}|z_i,W_i=0)-\big[m_{i2,0g}(z_i)-m_{i1,0}(z_i)\big]
\end{aligned} \tag{B.2}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[\frac{W_i}{p(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\pi_{1g}(z_i)}\big(Y_{i2}-m_{i2,1g}(z_i)\big)-\big(Y_{i1}-m_{i1,1}(z_i)\big)\right)\middle| z_i\right] \\
&-\mathbb{E}\left[\frac{1-W_i}{1-p(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\pi_{0g}(z_i)}\big(Y_{i2}-m_{i2,0g}(z_i)\big)-\big(Y_{i1}-m_{i1,0}(z_i)\big)\right)\middle| z_i\right]+\Delta m_{i2,g}(z_i)-\Delta m_{i1}(z_i) \\
=&\mathbb{E}(Y_{i2}|z_i,W_i=1,G_i=g)-\mathbb{E}(Y_{i1}|z_i,W_i=1)-\big[\mathbb{E}(Y_{i2}|z_i,W_i=0,G_i=g)-\mathbb{E}(Y_{i1}|z_i,W_i=0)\big] \\
&-\Big(\big[m_{i2,1g}(z_i)-m_{i1,1}(z_i)\big]-\big[m_{i2,0g}(z_i)-m_{i1,0}(z_i)\big]\Big)+\Delta m_{i2,g}(z_i)-\Delta m_{i1}(z_i) \\
=&\mathbb{E}\big[y_{i2}(1,\boldsymbol{w}_{-i})-y_{i2}(0,\boldsymbol{w}_{-i})|W_i=1,G_i=g,z_i\big]
\end{aligned} \tag{B.3}$$

When conditional means are correctly specified, $m_{it,wg}(z)=\mu_{it,wg}(z)$ and $m_{it,w}(z)=\mu_{it,w}(z)$.

$$\begin{aligned}
&\mathbb{E}\left[\frac{W_i}{\eta(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\eta_{1g}(z_i)}\big(Y_{i2}-\mu_{i2,1g}(z_i)\big)-\big(Y_{i1}-\mu_{i1,1}(z_i)\big)\right)\middle| z_i\right] \\
=&\mathbb{E}\left[\frac{W_i}{\eta(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\eta_{1g}(z_i)}\big(Y_{i2}-\mu_{i2,1g}(z_i)\big)-\big(Y_{i1}-\mu_{i1,1}(z_i)\big)\right)\middle| z_i,W_i=1\right]P(W_i=1|z_i) \\
=&\frac{p(z_i)}{\eta(z_i)}\mathbb{E}\left[\frac{\mathbb{1}\{G_i=g\}}{\eta_{1g}(z_i)}\big(Y_{i2}-\mu_{i2,1g}(z_i)\big)\middle| z_i,W_i=1,G_i=g\right]P(G_i=g|z_i,W_i=1) \\
&-\frac{p(z_i)}{\eta(z_i)}\mathbb{E}\big(Y_{i1}-\mu_{i1,1}(z_i)\big|z_i,W_i=1\big) \\
=&\frac{p(z_i)}{\eta(z_i)}\frac{\pi_{1g}(z_i)}{\eta_{1g}(z_i)}\big[\mathbb{E}(Y_{i2}|z_i,W_i=1,G_i=g)-\mu_{i2,1g}(z_i)\big]-\frac{p_i(z_i)}{\eta(z_i)}\big[\mathbb{E}(Y_{i1}|z_i,W_i=1)-\mu_{i1,1}(z_i)\big]=0
\end{aligned} \tag{B.4}$$

Analogously,

$$\mathbb{E}\left[\frac{1-W_i}{1-\eta(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\eta_{0g}(z_i)}\big(Y_{i2}-\mu_{i2,0g}(z_i)\big)-\big(Y_{i1}-\mu_{i1,0}(z_i)\big)\right)\middle| z_i\right]=0 \tag{B.5}$$

As a result,

$$\begin{aligned}
&\mathbb{E}\left[\frac{W_i}{\eta(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\eta_{1g}(z_i)}\big(Y_{i2}-\mu_{i2,1g}(z_i)\big)-\big(Y_{i1}-\mu_{i1,1}(z_i)\big)\right)\middle| z_i\right]\\
&-\mathbb{E}\left[\frac{1-W_i}{1-\eta(z_i)}\left(\frac{\mathbb{1}\{G_i=g\}}{\eta_{0g}(z_i)}\big(Y_{i2}-\mu_{i2,0g}(z_i)\big)-\big(Y_{i1}-\mu_{i1,0}(z_i)\big)\right)\middle| z_i\right]\\
&+\Delta\mu_{i2,g}(z_i)-\Delta\mu_{i1}(z_i)\\
=&\Delta\mu_{i2,g}(z_i)-\Delta\mu_{i1}(z_i)\\
=&\mathbb{E}\big[y_{i2}(1,\boldsymbol{w}_{-i})-y_{i2}(0,\boldsymbol{w}_{-i})|W_i=1,G_i=g,z_i\big]
\end{aligned} \tag{B.6}$$

Proof of Lemma 4.1:

Let $f_i(\cdot)$ be a generic function standing for each element of either $q_i(X_i,\theta)$ or $\nabla_\theta q_i(X_i,\theta)$. Denote $f_i^{(r)}=f_i\big(X_i^{(r)}\big)=f_i\big(y_{it}\big(\boldsymbol{W}^{(i,r,0)}\big),G\big(i,\boldsymbol{W}_{-i}^{(i,r,0)}\big),W_i,z_i\big)$. First, for $s\le 3\max\{K,1\}$, we have

$$|Cov(f_i,f_j)|\le 2\sup_{M,i\in D_M}\|f_i\|_\infty^2\le C_1<\infty \tag{B.7}$$

Next, consider $s>3\max\{K,1\}$.

$$\begin{aligned}
&|Cov(f_i,f_j)|=|Cov(f_i-f_i^{(s/3)}+f_i^{(s/3)},f_j)|\\
\le&|Cov(f_i-f_i^{(s/3)},f_j)|+|Cov(f_i^{(s/3)},f_j-f_j^{(s/3)})|+|Cov(f_i^{(s/3)},f_j^{(s/3)})|\\
\le&2\|f_j\|_\infty\mathbb{E}\Big[\big\|f_i-f_i^{(s/3)}\big\|\Big|\boldsymbol{z}\Big]+2\|f_i\|_\infty\mathbb{E}\Big[\big\|f_j-f_j^{(s/3)}\big\|\Big|\boldsymbol{z}\Big]+\Big|Cov\big(f_i^{(s/3)},f_j^{(s/3)}\big)\Big|
\end{aligned} \tag{B.8}$$

For the first two terms in equation (B.8),

$$\begin{aligned}&\|f_j\|_\infty \mathbb{E}\Big[\big\|f_i - f_i^{(s/3)}\big\|\Big|\boldsymbol{z}\Big] + \|f_i\|_\infty \mathbb{E}\Big[\big\|f_j - f_j^{(s/3)}\big\|\Big|\boldsymbol{z}\Big] \\ \leq & 2 \sup_{M,i\in D_M} \|f_i\|_\infty \sup_{M,i\in D_M} \mathrm{Lip}(f_i) \sup_{M,i\in D_M} \mathbb{E}\Big[\big\|X_i - X_i^{(s/3)}\big\|\Big|\boldsymbol{z}\Big]. \end{aligned} \tag{B.9}$$

Since $s/3 \geq K$,

$$\big(Y_{i1}, y_{i2}\big(\boldsymbol{W}^{(i,s/3,0)}\big), G\big(i, \boldsymbol{W}_{-i}^{(i,s/3,0)}\big), W_i, z_i\big) = \big(Y_{i1}, y_{i2}\big(\boldsymbol{W}^{(i,s/3,0)}\big), G\big(i, \boldsymbol{W}_{-i}\big), W_i, z_i\big).$$

As a result,

$$\mathbb{E}\Big[\big\|X_i - X_i^{(s/3)}\big\|\Big|\boldsymbol{z}\Big] = \mathbb{E}\Big[y_{i2}(\boldsymbol{W}) - y_{i2}\big(\boldsymbol{W}^{(i,s/3,0)}\big)\big|\boldsymbol{z}\Big] \leq \kappa_M(s/3). \tag{B.10}$$

For any fixed s, $f_i^{(s/3)}$ is α-mixing under Assumption 6. By Proposition 2.2 in Kojevnikov et al. (2021), the last term in equation (B.8) is bounded by

$$C_2 \alpha^{f^{(s/3)}}(1,1,s) \leq C_2 \alpha_M^\epsilon\Big(C_3\big(\frac{s}{3}\big)^d, C_3\big(\frac{s}{3}\big)^d, \frac{s}{3}\Big). \tag{B.11}$$

Putting these together, equation (B.8) is bounded by

$$C\big(\kappa_M(s/3) + s^d \widehat{\alpha}_M^\epsilon(s/3)\big). \tag{B.12}$$

Proof of Theorem 4.2:

I prove the theorem by verifying Theorem 2.1 and Theorem 3.2 in Newey and McFadden (1994). I first show $\hat{\theta} - \theta_M^* \xrightarrow{p} \mathbf{0}$.

Under Assumption A.1(vi) and Assumption 7

$$\frac{1}{|D_M|}\sum_{i\in D_M} q_i(X_i,\theta) - \mathbb{E}_D\big[q_i(X_i,\theta)\big] \xrightarrow{p} \mathbf{0} \tag{B.13}$$

follows from Lemma 4.1 and Theorem 3.1 in Kojevnikov et al. (2021). Next,

$$\sup_{\theta\in\Theta}\left\|\frac{1}{|D_M|}\sum_{i\in D_M} q_i(X_i,\theta) - \mathbb{E}_D\big[q_i(X_i,\theta)\big]\right\| \xrightarrow{p} \mathbf{0} \tag{B.14}$$

follows from Corollary 3.1 in Newey (1991) and equation (B.13) under condition (v). Also, $\mathbb{E}_D\big[q_i(X_i,\theta)\big]$ is uniformly equicontinuous. Let

$$\hat{Q}(\theta) = \frac{1}{|D_M|}\sum_{i\in D_M} q_i(X_i,\theta)'\widehat{\Psi}\frac{1}{|D_M|}\sum_{i\in D_M} q_i(X_i,\theta).$$

Finally, we need to show

$$\sup_{\theta\in\Theta}|\hat{Q}(\theta) - Q_M(\theta)| \xrightarrow{p} 0 \tag{B.15}$$

and $Q_M(\theta)$ is uniformly equicontinuous. The proof of equation (B.15) and the equicontinuity is standard. One can follow, for instance, the proof of Theorem 3 in Jenish and Prucha (2012).

Next, I prove the asymptotic normality. The key steps are to prove

$$\Omega_M^{-1/2}\frac{1}{\sqrt{|D_M|}}\sum_{i\in D_M} q_i(X_i,\theta_M^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I_k) \tag{B.16}$$

and

$$\sup_{\theta\in\Theta}\left\|\frac{1}{|D_M|}\sum_{i\in D_M} \nabla_\theta q_i(X_i,\theta) - \mathbb{E}_D\big[\nabla_\theta q_i(X_i,\theta)\big]\right\| \xrightarrow{p} \mathbf{0}. \tag{B.17}$$

Equation (B.16) is implied by Theorem 3.2 in Kojevnikov et al. (2021), Lemma 4.1, and

the Cramer-Wold device under Assumption A.1(vi) and (viii) and Assumption 8. By analogous argumentation for the proof of consistency, equation (B.17) holds under Assumption A.1(ix) and (x).

Proof of Theorem 4.3:

Using analogous arguments in the proof of Theorem 4.2, $\hat{R} - R_M^* \xrightarrow{p} \mathbf{0}$. The key step is to show that $\tilde{\Omega}(\hat{\theta}) - \Omega_M - \Omega_E \xrightarrow{p} \mathbf{0}$.

Notice that

$$\begin{aligned}\Omega_M =& \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M} \mathbb{E}\Big\{ \Big(q_i(X_i, \theta_M^*) - \mathbb{E}\big[q_i(X_i, \theta_M^*)|\boldsymbol{z}\big] \Big) \cdot \Big(q_j(X_j, \theta_M^*) - \mathbb{E}\big[q_j(X_j, \theta_M^*)|\boldsymbol{z}\big] \Big)' \Big| \boldsymbol{z} \Big\} \\ =& \frac{1}{|D_M|} \sum_{i \in D_M} \sum_{j \in D_M} \mathbb{E}\big(\tilde{q}_i(X_i, \theta_M^*) \tilde{q}_j(X_j, \theta_M^*)'\big), \end{aligned} \tag{B.18}$$

where

$$\tilde{q}_i(X_i, \theta_M^*) = q_i(X_i, \theta_M^*) - \mathbb{E}\big[q_i(X_i, \theta_M^*)|\boldsymbol{z}\big] \tag{B.19}$$

with $\mathbb{E}\big[\tilde{q}_i(X_i, \theta_M^*)|\boldsymbol{z}\big] = \mathbf{0}$.

Since any sequence of symmetric matrices $\{A_N\}$ converges to a symmetric matrix $\{A_0\}$ if and only if $c'A_N c \to c'A_0 c$ for any vectors c, we can reach our conclusion by taking an arbitrary linear combination of $q_i(X_i, \theta)$. From now on, we focus on the case of scalar $q_i(X_i, \theta)$.

$$\left\| \tilde{\Omega}(\hat{\theta}) - \Omega_M - \Omega_E \right\| \leq \left\| \tilde{\Omega}(\hat{\theta}) - \tilde{\Omega}(\theta_M^*) \right\| + \left\| \tilde{\Omega}(\theta_M^*) - \Omega_M - \Omega_E \right\|. \tag{B.20}$$

For the first term in the right hand side of (B.20), take a mean value expansion of $\tilde{\Omega}(\hat{\theta})$ around θ_M^*. Let $\check{\theta}$ denote the mean value from this expansion.

$|\tilde{\Omega}(\hat{\theta}) - \tilde{\Omega}(\theta_M^*)|$

$$=\left|(\hat{\theta}-\theta_M^*)\frac{1}{|D_M|}\sum_{s=0}^{\infty}\omega\Big(\frac{s}{b_M}\Big)\sum_{i\in D_M}\sum_{j\in D_M, s\le\rho(i,j)<s+1}\big[\nabla_\theta q_i(X_i,\check{\theta})q_j(X_j,\check{\theta})+q_i(X_j,\check{\theta})\nabla_\theta q_j(X_j,\check{\theta})\big]\right|$$

$$\le C_1\big|\sqrt{|D_M|}(\hat{\theta}-\theta_M^*)\big|\frac{1}{|D_M|^{3/2}}\sum_{s=1}^{b_M}\sum_{i\in D_M}\sum_{j\in D_M, s\le\rho(i,j)<s+1}\sup_{\theta\in\Theta}\big|\nabla_\theta q_i(X_i,\theta)q_j(X_j,\theta)\big|$$

$$\le C\big|\sqrt{|D_M|}(\hat{\theta}-\theta_M^*)\big|\frac{1}{\sqrt{|D_M|}}\sum_{s=1}^{b_M}s^{d-1}\frac{1}{|D_M|}\sum_{i\in D_M}\sup_{\theta\in\Theta}\big|\nabla_\theta q_i(X_i,\theta)q_j(X_j,\theta)\big| \tag{B.21}$$

Since

$$\mathbb{E}\Big[\frac{1}{|D_M|}\sum_{i\in D_M}\sup_{\theta\in\Theta}\big|\nabla_\theta q_i(X_i,\theta)q_j(X_j,\theta)\big|\Big|\boldsymbol{z}\Big]\le\sup_{M,i\in D_M}\mathbb{E}\Big[\sup_{\theta\in\Theta}\big|\nabla_\theta q_i(X_i,\theta)q_j(X_j,\theta)\big|\Big|\boldsymbol{z}\Big]$$

$$\le\sup_{M,i\in D_M}\mathbb{E}\Big[\sup_{\theta\in\Theta}\big|\nabla_\theta q_i(X_i,\theta)\big|^2\Big|\boldsymbol{z}\Big]^{1/2}\cdot\sup_{M,i\in D_M}\mathbb{E}\Big[\sup_{\theta\in\Theta}\big|q_i(X_i,\theta)\big|^2\Big|\boldsymbol{z}\Big]^{1/2}<\infty, \tag{B.22}$$

$$\frac{1}{|D_M|}\sum_{i\in D_M}\sup_{\theta\in\Theta}\big|\nabla_\theta q_i(X_i,\theta)q_j(X_j,\theta)\big|=O_p(1) \tag{B.23}$$

by Markov's inequality. Given $b_M=o\big(|D_M|^{1/2d}\big)$, $\frac{1}{\sqrt{|D_M|}}\sum_{s=1}^{b_M}s^{d-1}=o(1)$. Also, $\sqrt{|D_M|}(\hat{\theta}-\theta_M^*)=O_p(1)$ by Theorem 4.2. Hence, $|\tilde{\Omega}(\hat{\theta})-\tilde{\Omega}(\theta_M^*)|=o_p(1)$.

Let

$$\check{\Omega}_M=\frac{1}{|D_M|}\sum_{s=0}^{\infty}\omega\Big(\frac{s}{b_M}\Big)\sum_{i\in D_M}\sum_{j\in D_M, s\le\rho(i,j)<s+1}\tilde{q}_i(X_i,\theta_M^*)\tilde{q}_j(X_j,\theta_M^*). \tag{B.24}$$

Applying Proposition 4.1 in Kojevnikov et al. (2021), we have

$$\big\|\check{\Omega}_M-\Omega_M\big\|=o_p(1). \tag{B.25}$$

What is left is to show

$$\begin{aligned}
&\left\|\tilde{\Omega}(\theta_M^*) - \Omega_E - \check{\Omega}_M\right\| \\
\leq& 2\left\|\frac{1}{|D_M|}\sum_{s=0}^{\infty}\omega\Big(\frac{s}{b_M}\Big)\sum_{i\in D_M}\sum_{j\in D_M, s\leq\rho(i,j)<s+1}\mathbb{E}\big[q_j(X_j,\theta_M^*)|\boldsymbol{z}\big]\tilde{q}_i(X_i,\theta_M^*)\right\| \qquad \text{(B.26)}\\
=& o_p(1).
\end{aligned}$$

Let $B_i = \sum_{s=0}^{\infty}\omega\big(\frac{s}{b_M}\big)\sum_{j\in D_M, s\leq\rho(i,j)<s+1}\mathbb{E}\big[q_j(X_j,\theta_M^*)|\boldsymbol{z}\big]$.

$$\begin{aligned}
&\left\|\frac{1}{|D_M|}\sum_{s=0}^{\infty}\omega\Big(\frac{s}{b_M}\Big)\sum_{i\in D_M}\sum_{j\in D_M, s\leq\rho(i,j)<s+1}\mathbb{E}\big[q_j(X_j,\theta_M^*)|\boldsymbol{z}\big]\tilde{q}_i(X_i,\theta_M^*)\right\|_1 \\
\leq& \left\|\frac{1}{|D_M|}\sum_{i\in D_M}\tilde{q}_i(X_i,\theta_M^*)B_i\right\|_2 \\
\leq& \Big[\frac{1}{|D_M|^2}\sum_{i\in D_M}\mathbb{E}\big(\tilde{q}_i(X_i,\theta_M^*)^2|\boldsymbol{z}\big)B_i^2 + \frac{1}{|D_M|^2}\sum_{i\in D_M}\sum_{j\in D_M, j\neq i}\mathbb{E}\big(\tilde{q}_i(X_i,\theta_M^*)\tilde{q}_j(X_j,\theta_M^*)|\boldsymbol{z}\big)B_iB_j\Big]^{1/2} \\
\leq& \Big[\frac{C_1}{|D_M|}b_M^{2d} + \frac{C_2}{|D_M|^2}\sum_{i\in D_M}\sum_{s=1}^{\infty}\sum_{j\in D_M, s\leq\rho(i,j)<s+1}\tilde{\kappa}_{M,s}B_iB_j\Big]^{1/2} \\
\leq& \Big[o(1) + \frac{C_2}{|D_M|}\sum_{s=1}^{\infty}s^{d-1}b_M^{2d}\tilde{\kappa}_{M,s}\Big]^{1/2} = o(1). \qquad \text{(B.27)}
\end{aligned}$$

Hence, equation (B.26) follows from Markov's inequality. Theorem 4.3 follows by continuity of matrix inversion and multiplication.

C Online Appendix

C.1 Additional Proofs

Proof of Proposition 1:

Compare the native DID estimand with EDATT:

$$
\begin{aligned}
\tau =& \sum_{g\in\mathcal{G}} \tau(g) P(G_i = g|W_i = 1, z_i) \\
=& \sum_{g\in\mathcal{G}} \mathbb{E}\big[y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i})|W_i = 1, G_i = g, z_i\big] P(G_i = g|W_i = 1, z_i) \\
=& \sum_{g\in\mathcal{G}} \Big\{ \mathbb{E}\big(y_{i2}(1, \boldsymbol{w}_{-i})|W_i = 1, G_i = g, z_i\big) - \mathbb{E}\big(y_{i1}(0, \underline{0})|W_i = 1, z_i\big) \\
& - \Big[\mathbb{E}\big(y_{i2}(0, \boldsymbol{w}_{-i})|W_i = 0, G_i = g, z_i\big) - \mathbb{E}\big(y_{i1}(0, \underline{0})|W_i = 0, z_i\big)\Big]\Big\} P(G_i = g|W_i = 1, z_i) \\
=& \sum_{g\in\mathcal{G}} \mathbb{E}(Y_{i2}|W_i = 1, G_i = g, z_i) P(G_i = g|W_i = 1, z_i) - \mathbb{E}(Y_{i1}|W_i = 1, z_i) \\
& - \Big[\sum_{g\in\mathcal{G}} \mathbb{E}(Y_{i2}|W_i = 0, G_i = g, z_i) P(G_i = g|W_i = 1, z_i) - \mathbb{E}(Y_{i1}|W_i = 0, z_i)\Big]
\end{aligned} \tag{C.1}
$$

$$
\begin{aligned}
\tau_{canonic} =& \mathbb{E}(Y_{i2} - Y_{i1}|W_i = 1, z_i) - \mathbb{E}(Y_{i2} - Y_{i1}|W_i = 0, z_i) \\
=& \sum_{g\in\mathcal{G}} \mathbb{E}(Y_{i2}|W_i = 1, G_i = g, z_i) P(G_i = g|W_i = 1, z_i) - \mathbb{E}(Y_{i1}|W_i = 1, z_i) \\
& - \Big[\sum_{g\in\mathcal{G}} \mathbb{E}(Y_{i2}|W_i = 0, G_i = g, z_i) P(G_i = g|W_i = 0, z_i) - \mathbb{E}(Y_{i1}|W_i = 0, z_i)\Big]
\end{aligned} \tag{C.2}
$$

Proof of Proposition 2:

$$
\begin{aligned}
\tau =& \sum_{g\in\mathcal{G}} \mathbb{E}\big[y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i})|W_i = 1, G_i = g, z_i\big] P(G_i = g|W_i = 1, z_i) \\
=& \sum_{g\in\mathcal{G}} \sum_{u\in\mathcal{U}} \mathbb{E}\big[y_{i2}(1, \boldsymbol{w}_{-i}) - y_{i2}(0, \boldsymbol{w}_{-i})|W_i = 1, G_i = g, z_i, u_i = u\big] \\
& \cdot P(u_i = u|W_i = 1, G_i = g, z_i) P(G_i = g|W_i = 1, z_i)
\end{aligned}
$$

$$
\begin{aligned}
=&\sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\Big\{\mathbb{E}\big(y_{i2}(1,\boldsymbol{w}_{-i})|W_i=1,G_i=g,z_i,u_i=u\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=1,z_i,u_i\big)\\
&-\Big[\mathbb{E}\big(y_{i2}(0,\boldsymbol{w}_{-i})|W_i=0,G_i=g,z_i,u_i\big)-\mathbb{E}\big(y_{i1}(0,\underline{0})|W_i=0,z_i,u_i\big)\Big]\Big\}\\
&\cdot P(u_i=u|W_i=1,G_i=g,z_i)P(G_i=g|W_i=1,z_i)\\
=&\sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i2}|W_i=1,G_i=g,z_i,u_i=u)P(u_i=u|W_i=1,G_i=g,z_i)\\
&\cdot P(G_i=g|W_i=1,z_i)-\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i1}|W_i=1,z_i,u_i=u)P(u_i=u|W_i=1,z_i)\\
&-\Big[\sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i2}|W_i=0,G_i=g,z_i,u_i=u)P(u_i=u|W_i=1,G_i=g,z_i)\\
&\cdot P(G_i=g|W_i=1,z_i)-\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i1}|W_i=0,z_i,u_i=u)P(u_i=u|W_i=1,z_i)\Big]
\end{aligned}
\tag{C.3}
$$

$$
\begin{aligned}
\tau_{canonic}=&\mathbb{E}(Y_{i2}-Y_{i1}|W_i=1,z_i)-\mathbb{E}(Y_{i2}-Y_{i1}|W_i=0,z_i)\\
=&\sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i2}|W_i=1,G_i=g,z_i,u_i=u)P(u_i=u|W_i=1,G_i=g,z_i)\\
&\cdot P(G_i=g|W_i=1,z_i)-\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i1}|W_i=1,z_i,u_i=u)P(u_i=u|W_i=1,z_i)\\
&-\Big[\sum_{g\in\mathcal{G}}\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i2}|W_i=0,G_i=g,z_i,u_i=u)P(u_i=u|W_i=0,G_i=g,z_i)\\
&\cdot P(G_i=g|W_i=0,z_i)-\sum_{u\in\mathcal{U}}\mathbb{E}(Y_{i1}|W_i=0,z_i,u_i=u)P(u_i=u|W_i=0,z_i)\Big]
\end{aligned}
\tag{C.4}
$$