Identification of Causal Diffusion Effects Using Placebo Outcomes Under Structural Stationarity*

Naoki Egami†

First Version: August 29, 2018
This Version: January 22, 2022

Abstract

Social and biomedical scientists have long been interested in the process through which ideas and behaviors diffuse. In this article, we study an urgent social problem, the spatial diffusion of hate crimes against refugees in Germany, which has admitted more than 1 million asylum seekers since the 2015 refugee crisis. Despite its importance, identification of causal diffusion effects, also known as peer and contagion effects, remains challenging because the commonly used assumption of no omitted confounders is often untenable due to contextual confounding and homophily bias. To address this long-standing problem, we examine causal identification using placebo outcomes under a new assumption of *structural stationarity*, which formalizes the underlying diffusion process with a class of nonparametric structural equation models with recursive structure. We show under structural stationarity that a lagged dependent variable is a general, valid placebo outcome for detecting a wide range of biases, including the two types mentioned above. We then propose a difference-in-differences style estimator that can directly correct biases under an additional causal assumption. Analyzing fine-grained geo-coded hate crime data from Germany, we show when and how the proposed methods can detect and correct unmeasured confounding in spatial causal diffusion analysis.

Keywords: Contagion effects, Difference-in-differences, Homophily bias, Peer effects, Social influence

*I thank Peter Aronow, Eytan Bakshy, Matt Blackwell, Dean Eckles, Justin Grimmer, Erin Hartman, Zhichao Jiang, Gary King, Dean Knox, James Robins, Ilya Shpitser, Dustin Tingley, Tyler VanderWeele, Soichiro Yamauchi, and participants of the 2019 Atlantic Causal Inference Conference, for helpful comments and discussions. I am particularly grateful to Kosuke Imai, Rafaela Dancygier, and Brandon Stewart for their detailed feedback. The earlier draft of this article was entitled, "Identification of Causal Diffusion Effects Using Stationary Causal Directed Acyclic Graphs," (Egami, 2018), arXiv: `https://arxiv.org/abs/1810.07858v1`

†Assistant Professor, Department of Political Science, Columbia University, New York NY 10027. Email: naoki.egami@columbia.edu, URL: `https://naokiegami.com`

1 Introduction

Scientists have long been interested in how ideas and behaviors diffuse across space, networks, and time. For example, social scientists have studied the diffusion of policies and voting behaviors in political science (Sinclair, 2012; Graham *et al.*, 2013; Jones *et al.*, 2017), educational outcomes and crimes in economics (Glaeser *et al.*, 1996; Sacerdote, 2001; Duflo *et al.*, 2011), and innovations and job attainment in sociology (Rogers, 1962; Granovetter, 1973). Epidemiologists and researchers in public health have focused on the spread of infectious disease (Halloran and Struchiner, 1995; Morozova *et al.*, 2018; Cai *et al.*, 2019) and health behavior (Christakis and Fowler, 2013). In each of these research areas, a growing number of scholars aim to estimate the causal impact of diffusion dynamics, that is, how much an outcome of one unit causes, not just correlates with, an outcome of another unit.

In this paper, we study the spatial diffusion of hate crimes against refugees in Germany. Facing the biggest refugee crisis since the Second World War, Germany has recently registered more than 1 million asylum applications, making them the largest refugee-hosting country in Europe (United Nations High Commissioner for Refugees., 2017). During this time period, the number of hate crimes against refugees has substantially increased, a close to 200% increase from 2015 to 2016. A clear, *descriptive* pattern is that the incidence of hate crimes was spatially clustered and the number grew over time as waves (see Section 2). However, what is the *causal* process behind this dynamic spatial pattern? Understanding the causal impact of hate crime diffusion is of policy and scientific interest to prevent the further spread of hate crimes.

Despite its importance, identification of causal diffusion effects, also known as peer effects, contagion effects, or social influence, is challenging (Manski, 1993; VanderWeele and An, 2013). Although commonly-used statistical methods, including spatial econometric models (e.g., Anselin, 2013), require the assumption of no omitted confounders, this assumption is often untenable due to two well-known types of confounding; contextual confounding and homophily bias (Ogburn, 2018). When there exist some unobserved contextual factors that affect multiple units, we suffer from *contextual confounding* — we cannot distinguish whether units affect one another through diffusion processes or units are jointly affected by the shared unobserved contextual variables. *Homophily bias* arises when the spatial or network proximity is affected by some unobserved characteristics. We cannot discern whether units close to one another exhibit similar outcomes because of diffusion or because they selectively become closer in space or networks with others who have similar unobserved characteristics. Emphasizing concerns over these biases, influential papers across disciplines criticize existing observational diffusion studies (e.g., Cohen-Cole and Fletcher, 2008; Lyons, 2011; Angrist, 2014). In fact,

causal diffusion effects are often found to be overestimated by a large amount, for example, by 300 – 700% (Aral *et al.*, 2009; Eckles and Bakshy, 2017). Shalizi and Thomas (2011) argue that it is nearly impossible to credibly estimate causal diffusion effects from observational studies by relying on the conventional assumption of no omitted confounders.

To address this long-standing challenge, we examine identification of causal diffusion effects using placebo outcomes — variables known to be not causally related to the treatment variable. In this paper, we show that a lagged dependent variable is a general, valid placebo outcome under a new assumption of *structural stationarity*, which formalizes diffusion processes with a nonparametric structural equation model (NPSEM) and its corresponding causal directed acyclic graph (DAG) (Pearl, 2000; Ogburn and VanderWeele, 2014). In particular, by extending a class of dynamic causal DAGs (Dean and Kanazawa, 1989; Pearl and Russell, 2001) to the diffusion setting, we assume that the underlying NPSEM has recursive causal structure over time, while we can leave unspecified how effects of each variable change over time. That is, the structural stationarity assumption requires the existence of causal relationships among variables — not the effect or sign of such relationships — to be stable over time. Instead of simply assuming the validity of placebo outcomes, we clarify the importance of structural stationarity to transparently choose and justify placebo outcomes for identifying causal diffusion effects.

Under structural stationarity, we first develop a statistical test that uses a lagged dependent variable as a placebo outcome to detect a wide class of biases, including contextual confounding and homophily bias (Section 4.2). It assesses whether a lagged dependent variable is conditionally independent of the treatment variable. We prove statistical properties of the test based on a new theorem, which states that under structural stationarity, the no omitted confounders assumption is equivalent to the conditional independence of a lagged dependent variable and the treatment variable.

In addition, we propose a bias-corrected estimator that can directly remove biases under an additional causal assumption (Section 4.3). In its basic form, it subtracts the bias detected by the placebo test from a biased estimator. We prove unbiasedness of this estimator under an assumption that the effect and imbalance of unobserved confounders are constant over time. We describe its connection to the widely-used difference-in-differences estimator (Angrist and Pischke, 2008; Sofer *et al.*, 2016).

Applying the proposed methods to fine-grained geo-coded hate crime data, we estimate the causal diffusion effect of hate crimes against refugees in Germany (Section 5). In contrast to existing studies (Braun, 2011; Jäckle and König, 2016), we first find that the spatial diffusion

effect is small when averaging over all counties. By removing contextual confounding that previous studies have suffered from, we avoid overestimation of the causal diffusion effect. Then, we extend this analysis by considering types of counties that are more susceptible to the diffusion of hate crimes. This further investigation shows that the spatial diffusion of hate crimes is concentrated in counties with a higher proportion of school dropouts, if any.

Related Literature. This article builds on a growing literature of causal diffusion effects (Shalizi and Thomas, 2011; Goldsmith-Pinkham and Imbens, 2013; Ogburn, 2018).[1] In particular, several papers develop methods specifically for network data. Some studies (e.g., Bramoullé *et al.*, 2009; O'Malley *et al.*, 2014; An, 2015) propose to use instrumental variables to examine causal diffusion effects (a.k.a., peer effects) in a network. McFowland III and Shalizi (2021) propose a consistent estimator of causal peer effects, which adjusts for estimated latent homophilous attributes in settings where the data generating process is linear and the network grows according to either a stochastic block model or a continuous space model. While these papers are powerful for analyzing causal diffusion effects in networks, these methods are not directly applicable to our application of the spatial diffusion of hate crimes. In contrast, our approach is applicable to spatial data as well as to network data.

This paper also draws upon emerging literature of negative controls (Lipsitch *et al.*, 2010). This paper extends recent studies using negative controls in panel data settings (Sofer *et al.*, 2016; Miao and Tchetgen Tchetgen, 2017) to identification of causal diffusion effects. Our work is different from two recent papers utilizing negative controls. Egami and Tchetgen Tchetgen (2021) propose a framework for using double negative controls (negative control outcome and exposure variables) for identification and estimation of causal peer effects in the presence of uncontrolled network confounding, while taking into account network dependence. Liu and Tchetgen Tchetgen (2020) use a negative control exposure variable. Unlike these two papers, our paper relies on a placebo outcome (a.k.a., negative control outcome), and thus, both the placebo test and the bias-corrected estimator are different. Second, while both papers focus on the two-period network data, we focus on panel data with both network and spatial settings and analyze the spatial diffusion of hate crimes in our application. To accommodate this generality, we introduced structural stationarity, which is not exploited in the other work.

Finally, our approach based on causal DAGs and corresponding NPSEM is different and complementary to an alternative approach based on chain graphs. Recent papers (Tchet-

[1]Related but different literature is on causal inference with interference. The difference is that while interference focuses primarily on the causal effect of others' *treatments*, diffusion (a.k.a, peer and contagion effects) considers the causal effect of others' *outcomes* (Ogburn and VanderWeele, 2014). See Halloran and Hudgens (2016) for a review of the interference literature.

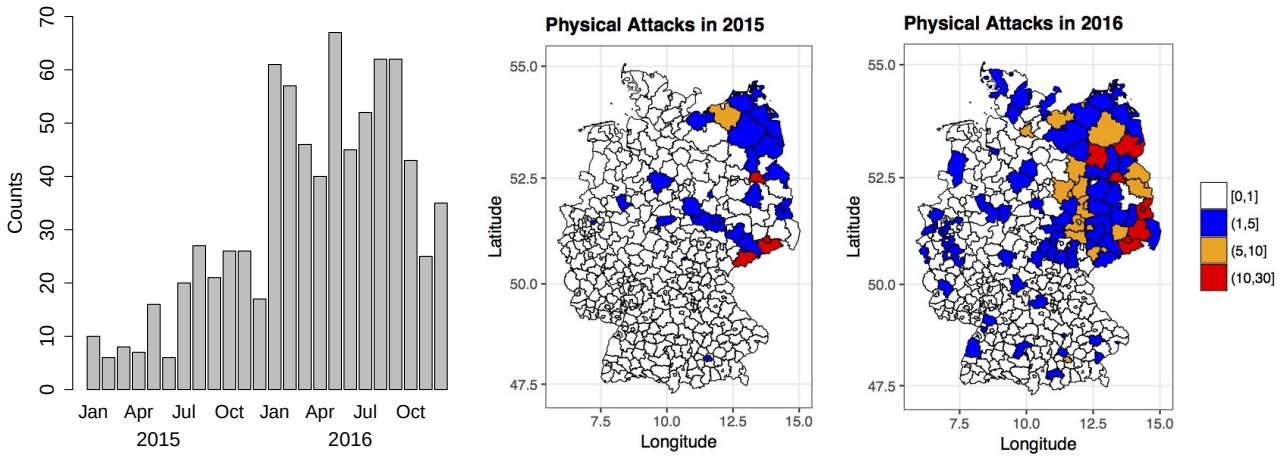
gen Tchetgen *et al.*, 2020; Ogburn *et al.*, 2020) discuss the difference between chain graphs and causal DAGs, and show the utility of chain graphs, especially when researchers are interested in characterizing equilibrium relationships between units in networks using cross-sectional data. Our approach is useful when we are interested in learning about causal diffusion effects — how units affect other units *over time step by step* — using panel data. This is exactly the setup of our motivating application, where we want to estimate how hate crimes spread across space over time in Germany.

2 A Motivating Empirical Application: Spatial Diffusion of Hate Crimes against Refugees

Research across the social sciences has shown that many types of violence are contagious (Wilson and Kelling, 1982; Myers, 2000). One small act of violence can trigger another act of violence, which again induces another, and can lead to waves of violence. Without taking into account how violent behaviors spread across space, it is difficult to explain when, where, and why some areas experience violence and to prevent the further spread of violence.

In this paper, we investigate the spatial diffusion of hate crimes against refugees in Germany, one of the most pressing problems in the country. Over the last few years, Germany has experienced a record influx of refugees, and during the same time period, the number of hate crimes against refugees has increased substantially. Our primary data source of hate crimes is a project, Mut gegen rechte Gewalt (courage against right-wing violence), by the Amadeu Antonio Foundation and the weekly magazine *Stern*, which has been documenting anti-refugee violence in Germany since the beginning of 2014. The dataset we analyze in this paper is compiled by Dancygier *et al.* (2020), who extended this hate crime data by merging in other variables, such as the number of refugees, the population size, a proportion of school dropouts and unemployment rates, collected from the Federal Statistical Office in Germany.

Figure 1 (a) reports the number of physical attacks against refugees each month, from the beginning of 2015 to the end of 2016. While there were about 15 hate crimes on average in each month of 2015, this rose to more than 40 in 2016, a close to 200% increase. Figure 1 (b) presents the spatial patterns over the two years. Two empirical patterns are worth noting. First, hate crimes were spatially clustered in East Germany. Second, the number of counties that experience hate crimes grew over time as waves. This dynamic spatial pattern is consistent with the spatial diffusion theory, which argues that hate crimes diffuse from one county to another spatially proximate county over time (Myers, 2000; Braun, 2011). Indeed, Jäckle and König (2016) found that the incidence of hate crimes in one county predicts that of hate crimes



(a) Time Trend (b) Spatial Patterns

Figure 1: Temporal and Spatial Patterns of Hate Crimes in Germany. Note: The left figure shows the number of physical attacks each month. In the middle and right figures, we show the number of physical attacks in each county in 2015 and 2016, respectively. Each of 402 counties is colored in white, blue, orange, or red if the number of hate crimes in a given year is less than or equal to 1, 5, 10, or greater than 10, respectively.

in its spatially proximate counties using the data from Germany in 2015.

However, it is challenging to estimate the causal impact of this spatial diffusion process because there exist well-known concerns of contextual confounding: many unobserved confounders can be spatially correlated. For example, the number of refugees increased substantially during this period and is also spatially correlated. Even if we collect a long list of covariates, it is difficult to assess whether a selected set of control variables is sufficient for removing contextual confounding. To address this type of pervasive concern over bias, we develop a placebo test to detect bias and a bias-corrected estimator to remove bias. The main empirical analysis appears in Section 5.

3 Setup for Causal Diffusion Analysis

Causal diffusion, also known as peer and contagion effects, refers to a process in which an outcome of one unit influences an outcome of another unit over time (Shalizi and Thomas, 2011; VanderWeele *et al.*, 2012). This section introduces a setup for analyzing such causal diffusion. We define the average causal diffusion effect and then describe challenges for its identification.

3.1 Notations and Definitions

Consider n units over T time periods. Let Y_{it} be the outcome for unit i at time t for $i \in \{1, \ldots, n\}$ and $t \in \{0, 1, \ldots, T\}$. Use $\mathbf{Y}_t$ to denote a vector $(Y_{1t}, \ldots, Y_{nt})$, which contains the

outcomes at time t for n units. To encode spatial or network connections between these n units, we follow the standard spatial statistics literature (Anselin, 2013) and use a distance matrix $\mathbf{W}$ where $\mathbf{W}$ can be an asymmetric, weighted matrix. In the motivating application, it is of interest to estimate how much hate crimes in one county diffuse to other spatially proximate counties. Here, the distance matrix $\mathbf{W}$ could encode physical distance between counties where W_{ij} might be an inverse of the distance between district i and j. In network diffusion settings, W_{ij} could represent a directed tie, e.g., whether unit i follows unit j in a Twitter network. Define *neighbors* $\mathcal{N}_i$ to be other units that are connected with a given unit i, i.e., $\mathcal{N}_i \equiv \{j : W_{ij} \neq 0\}$. In spatial diffusion analysis, researchers often assign 0 to W_{ij} when the distance between two units is greater than a certain threshold, e.g., 100 km. We denote the outcome variables at time t of unit i's neighbors as $\mathbf{Y}_{\mathcal{N}_i,t} \equiv \{Y_{jt} : j \in \mathcal{N}_i\}$.

In causal diffusion analysis, we are interested in how an outcome of one unit is affected by the outcomes of neighbors over time, that is, the causal effect of neighbors' outcomes at the previous time points $\mathbf{Y}_{\mathcal{N}_i,t-1}$ on Y_{it}. In principle, it is possible to perform causal inference by defining a multivariate treatment variable $\mathbf{Y}_{\mathcal{N}_i,t-1}$. However, in practice, we often make a dimension-reducing assumption, known as the exposure mapping (Aronow and Samii, 2017), to define the treatment variable. In particular, we define the treatment variable D_{it} at time t as a function of relevant neighbors' outcomes at time $t-1$, $D_{it} \equiv \phi(\mathbf{Y}_{\mathcal{N}_i,t-1}) \in \mathbb{R}$, where $\phi(\cdot)$ is a function specified by researchers based on their substantive interest. In the spatial statistics literature (Anselin, 2013), researchers have focused on the weighted average of the neighbors' outcomes $D_{it} = \mathbf{W}_i^\top \mathbf{Y}_{t-1}$ as the treatment variable. Following this practice, we examine the treatment variable $D_{it} = \mathbf{W}_i^\top \mathbf{Y}_{t-1}$ for concrete presentation throughout the paper, but the methodologies in this paper can be applied to other definitions of exposure mapping ϕ as well.

With this definition of the treatment, we can define the potential outcome (Neyman, 1923; Rubin, 1974; Robins, 1986). $Y_{it}(d)$ is the potential outcome variable of unit i at time t if the unit receives the treatment $D_{it} = d$ where $d \in \mathcal{D}_t$ and $\mathcal{D}_t$ is the support of D_{it}. Throughout the paper, we assume the standard consistency assumption linking observed and potential outcomes: $Y_{it} = Y_{it}(D_{it})$.

We are interested in the *average causal diffusion effect* (ACDE) at time t, which is defined as the average causal effect of the treatment variable D_{it} on the outcome at time t (Ogburn and VanderWeele, 2014; Ogburn, 2018). It is the comparison between the potential outcome under a higher value of the treatment $D_{it} = d^H$ and the potential outcome under a lower value of the treatment $D_{it} = d^L$.

Definition 1 (Average Causal Diffusion Effect)

The average causal diffusion effect (ACDE) at time t is defined as,

$$\tau_t(d^H, d^L) \equiv \mathbb{E}[Y_{it}(d^H) - Y_{it}(d^L)], \tag{1}$$

where d^H and d^L are two constants specified by researchers.

For example, the ACDE could quantify how much the risk of having hate crimes in the next month changes if we see more hate crimes in neighboring counties this month. This captures how much hate crimes diffuse across space over time. An important related causal estimand is the time-average ACDE, defined as,

$$\tau(d^H, d^L) \equiv \frac{1}{T}\sum_{t=1}^{T} \tau_t(d^H, d^L). \tag{2}$$

Because identification of this time-average ACDE follows from the ACDE, we focus on the ACDE unless otherwise noted.

3.2 Identification under No Omitted Confounders Assumption

We now consider a widely used identification assumption of no omitted confounders and explain pervasive concerns about its violation.

The no omitted confounders assumption states that all relevant confounders are observed, and researchers select them as an adjustment set. Formally, the no omitted confounders assumption states that the potential outcomes at time t are independent of a joint distribution of neighbors' outcomes at time $t-1$ given an adjustment set.

Assumption 1 (No Omitted Confounders)

For $i = 1, 2, \ldots, n$,

$$Y_{it}(d) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i, t-1} \mid \overline{\mathbf{C}}_{it}, \tag{3}$$

for all $d \in \mathcal{D}_t$ where $\mathcal{D}_t$ is the support of D_{it}. $\overline{\mathbf{C}}_{it}$ can only include variables not affected $\mathbf{Y}_{\mathcal{N}_i, t-1}$. An overline clarifies that adjustment set $\overline{\mathbf{C}}_{it}$ can include variables not only measured at time t but also those measured before time t.

Under this assumption of no omitted confounders (Assumption 1) and the standard positivity assumption described below, the ACDE is identified as follows.

$$\tau_t(d^H, d^L) \;=\; \int_{\mathcal{C}} \Big\{ \mathbb{E}[Y_{it} | D_{it} = d^H, \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it} | D_{it} = d^L, \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}] \Big\} dF_{\overline{\mathbf{C}}_{it}}(\overline{\mathbf{c}}), \tag{4}$$

where $F_{\overline{\mathbf{C}}_{it}}(\overline{\mathbf{c}})$ is the cumulative distribution function of $\overline{\mathbf{C}}_{it}$. The standard positivity assumption states that $\Pr(D_{it} = d^H | \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}) > 0$ and $\Pr(D_{it} = d^L | \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}) > 0$ for $i = 1, \ldots, n$ and all $\overline{\mathbf{c}} \in \mathcal{C}$ where $\mathcal{C}$ is the support of $\overline{\mathbf{C}}_{it}$. We can estimate the ACDE by estimating the conditional expectation $\mathbb{E}[Y_{it} | D_{it}, \overline{\mathbf{C}}_{it}]$ and then averaging it over the empirical distribution of the adjustment set $\overline{\mathbf{C}}_{it}$.

Remark. Note that Assumption 1 is stronger than $Y_{it}(d) \perp\!\!\!\perp D_{it} \mid \overline{\mathbf{C}}_{it}$, which is sufficient for identification. The advantage of using Assumption 1 is twofold: (1) we can use the same assumption for other definitions of the treatment based on different ϕ, and (2) we can develop a formal placebo test, the central topic of this paper we discuss in Section 4. □

Although many empirical studies of diffusion make the assumption of no omitted confounders, it is widely known that the assumption is often questionable in practice (Manski, 1993; Shalizi and Thomas, 2011; VanderWeele and An, 2013). This concern is pervasive mainly because it implies the absence of two well-known types of biases: contextual confounding and homophily bias. *Contextual confounding* – the primary focus of the spatial diffusion literature – can exist when units share some unobserved contextual factors. For example, in the motivating application of hate crime diffusion, the risk of having hate crimes is likely to be affected by some economic policies, which often affect multiple counties at the same time. In this case, researchers might observe spatial clusters of hate crimes even without diffusion.

Another well-known type of bias is *homophily bias* – the main concern in the network diffusion literature. This bias arises when units become connected due to their unobserved characteristics. For example, voters who are connected to each other can have similar political opinions without any diffusion or social influence because people who have similar political views might become friends in the first place (Fowler *et al.*, 2011). We discuss the causal DAG representation of these biases when we introduce our proposed methods in Section 4.

4 The Proposed Methodology

In this section, we examine identification of causal diffusion effects under an alternative assumption of structural stationarity. After introducing this assumption (Section 4.1), we first develop a statistical placebo test to detect a wide range of biases (Section 4.2) and then propose a bias-corrected estimator (Section 4.3).

4.1 Structural Stationarity

We use a causal directed acyclic graph (causal DAG) and its corresponding non-parametric structural equation model (NPEM) (Pearl, 2000) to explicitly examine potential violations of the no omitted confounders assumption. A causal DAG is a set of nodes $(V_1, \ldots, V_K)$, and directed edges among nodes such that the graph has no cycles. For each node V_k on the graph, the corresponding random variable is given by its non-parametric structural equation $V_k = f_k(\text{PA}(V_k), \epsilon_k)$ where $\text{PA}(V_k)$ are the parents of V_k on the graph, and the ϵ_k are mutually independent. In contrast to a linear structural equation model, non-parametric structural equations are entirely general — V_k may depend on any function of its parents and ϵ_k. The

non-parametric structural equations encode counterfactual relationships between the variables that are represented on the graph. We review basic terminologies for NPSEM and DAGs in Appendix B.

One key challenge of using NPSEMs in practice is that it is often difficult to specify one NPSEM that is valid and at the same time, general enough to accommodate various applied questions. This is especially difficult in the diffusion settings where units can be affected by other units over time. Thus, instead of specifying one particular NPSEM, we assume a class of NPSEMs that satisfy certain regularity conditions, what we call structural stationarity.

Intuitively, structural stationarity assumes that the existence of causal relationships between variables, not the effect or sign of such relationships, to be stable over time. This can be seen as an extension of dynamic causal DAGs (Dean and Kanazawa, 1989; Pearl and Russell, 2001) to the diffusion setting. We first formally define structural stationarity in general, and provide examples of NPSEMs below.

Definition 2 (Structural Stationarity)

Consider a NPSEM. Among random variables that have more than one child or have at least one parent, distinguish two types; the time-varying variable A_{it} and the time-invariant variable B_i. Then, a NPSEM is said to satisfy structural stationarity if random variables in the NPSEM satisfy the following conditions.

(2.1) $A_{it} \in \text{PA}(A_{i,t+1})$ for $i \in \{1, \ldots, n\}$ and $t = 0, \ldots, T-1$.

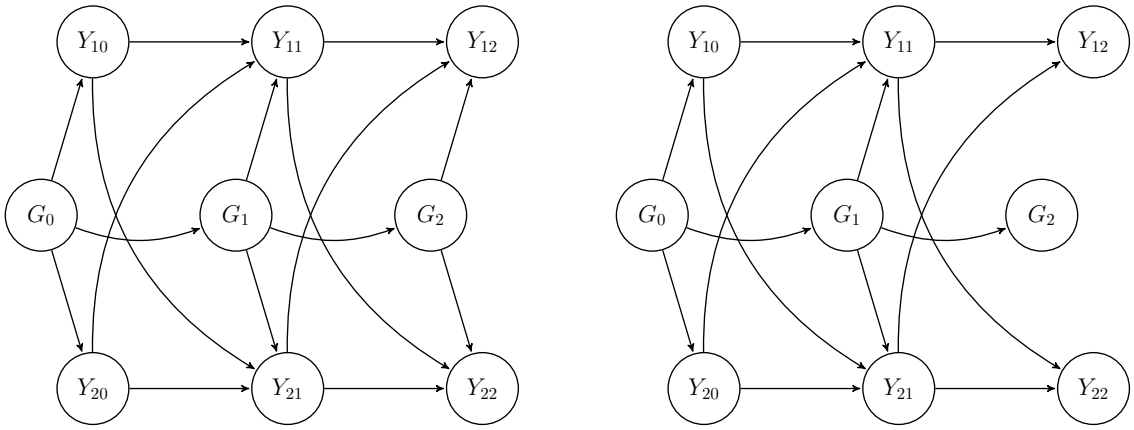
(2.2) For $i, i' \in \{1, \ldots, n\}$, if there exist two integers t and q such that $A_{it} \in \text{PA}(\widetilde{A}_{i',t+q})$, then $A_{it'} \in \text{PA}(\widetilde{A}_{i',t'+q})$ for all $t' = 0, \ldots, T-q$.

(2.3) For $i, i' \in \{1, \ldots, n\}$, if there exists integer t such that $B_i \in \text{PA}(A_{i't})$, then $B_i \in \text{PA}(A_{i't'})$ for all $t' = 0, \ldots, T$.

Example. We first consider a simple NPSEM that captures unmeasured contextual confounding. For $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, suppose data are generated by sequentially evaluating the following set of equations:

$$\begin{aligned}
&\text{(Outcome variable)} && Y_{it} = f_Y(\mathbf{Y}_{\mathcal{N}_i,t-1}, Y_{i,t-1}, \mathbf{L}_{it}, \widetilde{\mathbf{L}}_i, \mathbf{U}_{g[i],t}, \epsilon^Y_{it}) \\
&\text{(Time-varying Observed variables)} && \mathbf{L}_{it} = f_L(\mathbf{L}_{i,t-1}, Y_{i,t-1}, U_{g[i],t-1}, \boldsymbol{\epsilon}^L_{it}), \\
&\text{(Time-invariant Observed variables)} && \widetilde{\mathbf{L}}_i = f_{\widetilde{L}}(Y_{i,0}, U_{g[i],0}, \boldsymbol{\epsilon}^{\widetilde{L}}_i), \\
&\text{(Time-varying Unobserved variables)} && \mathbf{U}_{gt} = f_U(\mathbf{U}_{g,t-1}, \boldsymbol{\epsilon}^U_{gt}),
\end{aligned} \tag{5}$$

where $(\epsilon^Y_{it}, \boldsymbol{\epsilon}^L_{it}, \boldsymbol{\epsilon}^{\widetilde{L}}_i, \boldsymbol{\epsilon}^U_{gt})$ are unobserved exogenous errors. We use g to denote an unobserved context to which units belong, and use $g[i]$ to represent a context to which unit i belongs. Thus, $\mathbf{U}_{g[i],t}$ is an unobserved contextual variable for unit i, which induces non-causal associations



(a) Structural Stationarity (b) Violation of Structural Stationarity

Figure 2: Illustration of Structural Stationarity. Note: Six nodes Y_{it} represent outcome variables for two individuals $i \in \{1, 2\}$ over three time periods $t \in \{0, 1, 2\}$. Three nodes G_t are contextual variables for $t \in \{0, 1, 2\}$. In the first panel, the causal structure between variables Y and G are stable over time. In the second panel, variable G has no effect on Y at $t = 2$ and thus structural stationarity is violated.

between Y_{it} and $\mathbf{Y}_{\mathcal{N}_i,t-1}$ and violates the no omitted confounders assumption. The left panel of Figure 2 visualizes an instance of the NPSEM (5), while omitting observed variables for visual simplicity. Structural stationarity is violated in the right panel of Figure 2 because the causal relationships between outcomes and unmeasured context factors are different before and after time $t = 1$.

Condition 2.1 of Definition 2 requires that all time-varying variables that have at least one parent be affected by their own lagged variables. In NPSEM (5), outcomes Y_{it}, time-varying observed variables $\mathbf{L}_{it}$, and time-varying unobserved variables $\mathbf{U}_{gt}$ are all affected by their own lagged variables. This condition is more plausible when the time intervals are shorter. Condition 2.2 means that if two time-varying variables have a child-parent relationship at one time period, the same causal relationship should exist for all other time periods. For example, outcome Y_{it} is affected by unobserved contextual factor $\mathbf{U}_{g[i],t}$, and this child-parent relationship exists for all $t \in \{1, \ldots, T\}$. Finally, Condition 2.3 requires that if a time-invariant variable is a parent of a time-varying variable at one time period, the same child-parent relationship should exist at all other time periods. For example, outcome Y_{it} is affected by time-invariant variables $\widetilde{\mathbf{L}}_i$, and this child-parent relationship exists for all $t \in \{1, \ldots, T\}$.

The last two requirements are the core – the existence of causal relationships should be stable over time. Importantly, the effect of each variable can change over time; the only requirement is the time-invariant existence of the causal relationships. □

Remark. Structural stationarity is satisfied in a more general NPSEM as well. First, variables can be affected not only by one-time lag but also by longer-time lags. For example, outcome Y_{it} can be affected not only by the neighbors' outcomes at the last period $\mathbf{Y}_{\mathcal{N}_i,t-1}$ but also by the neighbors' outcomes at two periods before $\mathbf{Y}_{\mathcal{N}_i,t-2}$. Second, each variable can be not only affected by other variables within each unit but also by other variables of neighbors. For example, outcome Y_{it} can be affected by $\mathbf{L}_{\mathcal{N}_i,t-1}$ and $\mathbf{U}_{\mathcal{N}_i,t-1}$. We provide an additional example in Appendix C. □

Structural stationarity can accommodate many applied diffusion questions. Indeed, structural stationarity is often an implicit assumption researchers make in applied contexts. When analyzing panel data, analysts often adjust for the same set of confounders with only changing time indices (e.g., adjust for unemployment rates in 2015 when the outcome is the incidence of hate crimes in 2015, adjust for unemployment rates in 2016 when the outcome is the incidence of hate crimes in 2016, and so on). This implicitly assumes that the underlying NPSEM is stable and therefore, types of confounders they choose are also the same over time (only with the appropriate change in time indices).

Structural stationarity has also been a natural requirement for causal DAGs examined in causal diffusion analysis. In fact, causal DAGs in seminal papers about causal diffusion effects (Shalizi and Thomas, 2011; O'Malley *et al.*, 2014; Ogburn and VanderWeele, 2014) satisfy structural stationarity. Causal DAGs in the causal discovery literature often impose a similar but stronger condition (Danks and Plis, 2013; Hyttinen *et al.*, 2016). They often assume that variables are affected only by one-time lag (also known as the first-order Markov assumption) and this structure is time-invariant. In contrast, structural stationarity allows for any higher-order temporal dependence (see Condition 2.2 of Definition 2).

Structural stationarity is violated when the underlying causal structure changes at some time. For example, if a new time-varying confounder arises in the middle of the time periods we analyze, this will violate structural stationarity. If researchers know the time when the underlying structure changes, we can still make use of the structural stationarity assumption separately, before and after this time point. However, it is important to emphasize that structural stationarity is an untestable assumption as many other assumptions necessary for causal inference. Therefore, in general, structural stationarity is less plausible in applications where we expect the underlying diffusion structure is changing over time.

4.2 Placebo Test to Detect Bias

Under structural stationarity, we now propose a placebo test – using a lagged dependent variable as a general placebo outcome – that can detect a wide class of biases, including

contextual confounding and homophily bias. This placebo test can assess the validity of the confounder adjustment, thereby improving the credibility of identification of causal diffusion effects.

4.2.1 Equivalence Theorem

To formally prove a property of a placebo test, we first make the structural stationarity assumption.

Assumption 2

For $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, we assume that data – both observed and unobserved variables – are generated by sequentially evaluating a NPSEM that satisfies structural stationarity. We also assume the distribution of observed and unobserved variables is faithful[2] to this underlying NPSEM.

Two points are worth noting about Assumption 2. First, it requires that the underlying data is generated by a NPSEM that satisfies the structural stationarity. Importantly, however, it does not require researchers to specify a particular NPSEM. This can be important in practice where researchers have domain knowledge to justify that the existence of causal relationships is time-invariant but they lack precise knowledge necessary for justifying a particular NPSEM. Second, we also require faithfulness (Spirtes *et al.*, 2000) to an underlying NPSEM. This is important because, if the data distribution is not faithful to the underlying NPSEM, an unblocked backdoor path might induce no dependence, which we cannot detect from the data. This faithfulness assumption is commonly made in the causal discovery literature, and readers can find more details in Spirtes *et al.* (2000).

Under Assumption 2, we show the assumption of no omitted confounders is equivalent to the conditional independence of the simultaneous outcomes given *a placebo set* defined below.

Theorem 1 (Equivalence between No Omitted Confounders Assumption and Conditional Independence of Simultaneous Outcomes) Under Assumption 2, for covariates $\overline{\mathbf{C}}_{it}$ that are not affected by $\mathbf{Y}_{\mathcal{N}_i,t-1}$,

$$Y_{it}(d) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}_{it} \iff Y_{i,t-1} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}^P_{it}, \tag{6}$$

where a placebo set $\overline{\mathbf{C}}^P$ is defined as

$$\overline{\mathbf{C}}^P_{it} \equiv \{\overline{\mathbf{C}}_{it}, \overline{\mathbf{C}}^{(-1)}_{it}, \mathbf{Y}_{\mathcal{N}_i,t-2}\} \setminus \mathrm{Des}(Y_{i,t-1}), \tag{7}$$

where $\overline{\mathbf{C}}^{(-1)}_{it}$ is a lag of the time-varying variables in $\overline{\mathbf{C}}_{it}$, $\mathbf{Y}_{\mathcal{N}_i,t-2}$ is a lag of the treatment variable, and $\mathrm{Des}(Y_{i,t-1})$ is a descendant of $Y_{i,t-1}$, i.e., variables affected by $Y_{i,t-1}$.

[2]Faithfulness is defined as follows. If a distribution is faithful to a NPSEM, variables A and B are independent if and only if the variables are d-separated in the corresponding causal directed acyclic graph (Spirtes *et al.*, 2000).

The proof of Theorem 1 is in Appendix A.1.

In general, the assumption of no omitted confounders (the left-hand side of equation (6)) is not testable because it contains the potential outcomes $Y_{it}(d)$, which are inherently unobservables. This theorem shows that, under Assumptions 2, the assumption of no omitted confounders (the left-hand side) is equivalent to the conditional independence of the observed outcome of individual i and her neighbors' outcomes at the same time period given a placebo set (the right-hand side). Because this right-hand side is observable and testable, this theorem directly implies that we can statistically assess the assumption of no omitted confounders by the placebo test of the conditional independence of the simultaneous outcomes $Y_{i,t-1} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}_{it}^{P}$.

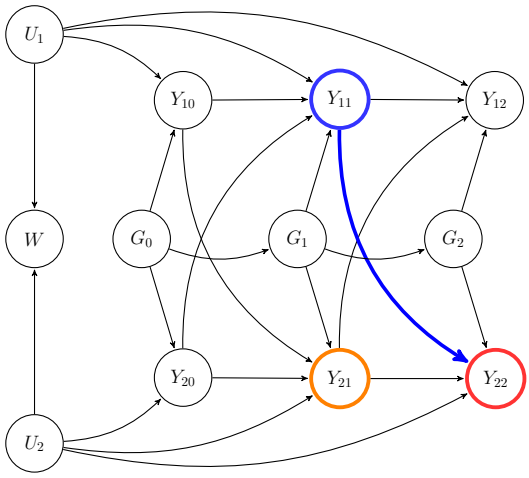
The basic idea behind the theorem is as follows: under the structural stationarity, back-door paths between the main outcome and the treatment are similar to those between the lagged dependent variable and the treatment. The difference between adjustment set $\overline{\mathbf{C}}$ and placebo set $\overline{\mathbf{C}}^{P}$ is to formally guarantee that unblocked back-door paths between the main outcome and the treatment are the same (from a causal graph perspective) to those between the placebo outcome and the treatment. To derive this placebo set, we only need to know which variables in the adjustment set are time-varying and which variables are affected by outcomes at time t. The former information is often readily available, and the latter one is the same as the information used to avoid post-treatment bias in the standard causal inference settings.

Every causal inference method requires some untestable assumption. Many existing approaches directly rely on the no omitted confounders assumption (Assumption 1), which is untestable and is also often untenable in practice. In contrast, Theorem 1 makes the no omitted confounders assumption testable under an alternative assumption of structural stationarity (Assumption 2), which is untestable and yet, can be more defensible in many applied settings.

4.2.2 Illustrations with Causal DAGs

Although the proposed placebo test is applicable to any NPSEMs and corresponding causal DAGs that satisfy structural stationarity, we consider a causal DAG in Figure 3 (a) as one concrete example. Suppose we are interested in the ACDE of Y_{11} on Y_{22} where Y_{11} is the treatment variable (blue), Y_{22} is the outcome variable (red), and the causal arrow of interest $Y_{11} \rightarrow Y_{22}$ is colored blue. The placebo outcome Y_{21} is colored orange.

Based on this causal DAG in Figure 3 (a), Table in Figure 3 (b) shows four different scenarios: no bias, contextual confounding, homophily bias, and both types of biases. For each set of control variables, the placebo test checks conditional independence, $Y_{11} \perp\!\!\!\perp Y_{21} \mid \overline{\mathbf{C}}^{P}$



(a) Example of Placebo Test

	$\overline{C}$	$\overline{C}^P$	Placebo Test
No Bias	Y_{21}, U_2, G_2	$Y_{20}, Y_{10}, U_2, G_2, G_1$	Accept
Contextual Confounding	Y_{21}, U_2	Y_{20}, Y_{10}, U_2	Reject
Homophily Bias	Y_{21}, G_2, G_1	$Y_{20}, Y_{10}, G_2, G_1, G_0$	Reject
Both	Y_{21}, Y_{20}	Y_{20}, Y_{10}	Reject

(b) Adjustment and Placebo Sets

Figure 3: Illustration of Placebo Test. Note: A DAG in (a) has nine variables in a DAG of Figure 2 in addition to two nodes U_i representing individual-level characteristics for $i \in \{1, 2\}$, and variable W indicating the connection of two individuals. We focus on the ACDE of Y_{11} on Y_{22} where Y_{11} is the treatment variable (blue), Y_{22} is the outcome variable (red), and the causal arrow of interest $Y_{11} \to Y_{22}$ is colored blue. The placebo outcome Y_{21} is colored orange.

where we derive a placebo set $\overline{\mathbf{C}}^P$ from a chosen control set $\overline{\mathbf{C}}$ using equation (7). These scenarios show how the placebo test detects biases by exploiting structural stationarity.

First, when we control for three variables $\{Y_{21}, U_2, G_2\}$, the ACDE of interest is identified ("No Bias"). Without knowledge of the entire causal DAG, we can assess the absence of bias by implementing the placebo test. Following equation (7), we derive a placebo set $\overline{\mathbf{C}}^P = \{Y_{20}, Y_{10}, U_2, G_2, G_1\}$ and then the placebo test checks $Y_{11} \perp\!\!\!\perp Y_{21} | \overline{\mathbf{C}}^P$. In Figure 3 (a), there is no unblocked back-door path between Y_{11} and Y_{21}, and the conditional independence holds as Theorem 1 implies.

Second, we consider a typical form of contextual confounding. When we control for two variables $\{Y_{21}, U_2\}$, the ACDE is not identified due to a back-door path ($Y_{11} \leftarrow G_1 \to G_2 \to Y_{22}$). We now verify that the placebo test correctly detects this bias. We first derive a placebo set as $\overline{\mathbf{C}}^P = \{Y_{20}, Y_{10}, U_2\}$ and then assess whether there is any unblocked back-door path between Y_{11} and Y_{21}. In fact, we correctly reject the placebo test; $Y_{11} \not\perp\!\!\!\perp Y_{21} | \overline{\mathbf{C}}^P$ due to a back-door path ($Y_{11} \leftarrow G_1 \to Y_{21}$). In Appendix, we also provide an illustration with homophily bias.

4.2.3 Parametric Placebo Test via Spatial Autoregressive Model

As Theorem 1 is nonparametric, researchers can employ a variety of non-, semi-parametric, or parametric conditional independence tests to implement the proposed placebo test. Among

many options, one practical approach is a parametric test based on the spatial autoregressive (SAR) model (e.g., Anselin, 2013). For example, when outcomes are continuous, we can implement the placebo test by the following linear spatial autoregressive model.

$$Y_{i,t-1} \quad = \quad \alpha_0 + \delta \mathbf{W}_i^\top \mathbf{Y}_{t-1} + \gamma_0^\top \overline{\mathbf{C}}_{it}^P + \epsilon_{i,t-1}, \tag{8}$$

where $\mathbf{W}_i^\top \mathbf{Y}_{t-1} \equiv D_{it}$ is the treatment variable, $\overline{\mathbf{C}}_{it}^P$ is a placebo set, and $\epsilon_{i,t-1}$ is an error term. In the motivating application (Section 5), we employ logistic spatial autoregressive model in a similar way. To account for spatial autocorrelation of errors, we rely on the spatial heteroskedasticity and autocorrelation consistent (spatial HAC) variance estimator by Conley (1999) to compute standard errors.

Theorem 1 implies that the placebo outcome $Y_{i,t-1}$ is conditionally independent of the treatment variable if the assumption of no omitted confounders holds. Therefore, the spatial autoregressive coefficient δ serves as a test statistic of the placebo test. By testing whether this spatial autoregressive coefficient is zero, researchers can assess the no omitted confounders assumption and thus detect biases, including contextual confounding and homophily bias. In Appendix D.1, we investigate the statistical power of the proposed placebo test through simulation studies and show that its power is comparable to a theoretical upper bound.

This use of the SAR model as a placebo test differs from existing approaches in the spatial econometrics literature that are designed to capture spatial correlations (e.g., Anselin, 2013). While researchers conventionally interpret the spatial autoregressive coefficient as the strength of the spatial correlation, the proposed placebo test uses the spatial autoregressive coefficient to detect biases rather than to estimate diffusion effects. For the estimation of the ACDE, we estimate the conditional expectation $\widehat{\mathbb{E}}[Y_{it} \mid D_{it}, \overline{\mathbf{C}}_{it}]$ and then use the identification formula in equation (4).

Remark. It is important to note that if the parametric assumptions of the model are violated, the spatial autoregressive coefficient in equation (8) can be zero even when unmeasured confounding remains. Like any other statistical tests, a specific parametric placebo test can fail if its underlying parametric assumptions do not hold. A key advantage of the proposed approach is that the equivalence theorem (Theorem 1) is nonparametric. The theorem implies that when there exist no omitted confounders, the placebo outcome and the treatment are conditionally independent in any parametric and nonparametric tests. Therefore, in practice, researchers can also verify the conditional independence of the placebo outcome and the treatment variable using additional non- or semiparametric conditional independence tests. □

4.3 Bias-Corrected Estimator

If the placebo test detects bias, one may want to collect more data and improve the selection of the adjustment set. This strategy might, however, be infeasible in many applied settings. To help researchers in such common situations, this section considers how to correct biases by introducing an additional assumption. We start with a simple example of linear models (Section 4.3.1) and then provide general results in Sections 4.3.2 and 4.3.3. We provide simulation evidence in Appendix D.2.

4.3.1 An Example with Linear Models

To develop an intuition for a bias-corrected estimator, we first consider a simple example with linear models. We assume here that a selected adjustment set is time-invariant and the same as its corresponding placebo set. A general result is provided in the following subsections.

Suppose we fit a linear model in which we regress the outcome at time t on the treatment variable and the selected adjustment set.

$$Y_{it} \quad = \quad \alpha + \beta D_{it} + \gamma^{\top} \overline{\mathbf{C}}_{it} + \tilde{\epsilon}_{it}, \tag{9}$$

where D_{it} is the treatment variable, $\overline{\mathbf{C}}_{it}$ is the selected adjustment set, and $\tilde{\epsilon}_{it}$ is an error term. If the assumption of no omitted confounders (Assumption 1) holds, $\hat{\beta} \times (d^H - d^L)$ is an unbiased estimator of the ACDE given that the linear model specification is correct. In contrast, when the no omitted confounders assumption is violated, this estimator is biased. We would like to assess whether the assumption of no omitted confounders holds and also correct biases, if any.

To assess the assumption of no omitted confounders, suppose we run a parametric placebo test using the following linear spatial autoregressive model as in equation (8).

$$Y_{i,t-1} \quad = \quad \alpha_0 + \delta D_{it} + \gamma_0^{\top} \overline{\mathbf{C}}_{it}^{P} + \epsilon_{i,t-1},$$

where $\overline{\mathbf{C}}_{it}^{P}$ is a placebo set and $\epsilon_{i,t-1}$ is an error term. If the assumption of no omitted confounders holds, the spatial autoregressive coefficient δ should be zero (Theorem 1). In contrast, if the assumption of no omitted confounders does not hold, an estimated coefficient $\hat{\delta}$ then serves as a bias-correction term.

In this simple example, a proposed bias-corrected estimator is given by subtracting the bias-correction term $\hat{\delta}$ from an original biased estimator $\hat{\beta}$.

$$\hat{\tau}_{BC}(d^H, d^L) \;\equiv\; (\hat{\beta} - \hat{\delta}) \times (d^H - d^L). \tag{10}$$

This bias-corrected estimator is unbiased for the ACDE for the treated under an additional causal assumption we discuss in detail in the next subsection (Assumption 3). Note that when

the assumption of no omitted confounders holds, the expected value of $\hat{\delta}$ is zero, meaning no bias correction.

4.3.2 Assumption

To describe a general bias-corrected estimator, we begin by defining the average causal diffusion effect for the treated (ACDT). We will show in Theorem 2 that the proposed bias-corrected estimator is unbiased for the ACDT. The formal definition is as follows.

$$\tau_t^H(d^H, d^L) \equiv \mathbb{E}[Y_{it}(d^H) - Y_{it}(d^L) \mid D_{it} = d^H]. \tag{11}$$

This is the average causal diffusion effect for units who received the higher level of the treatment. This quantity could represent the causal diffusion effect of hate crimes for counties in a higher risk neighborhood, i.e., $d^H\%$ of neighboring counties had hate crimes in month $t-1$.

To introduce necessary assumptions, we divide an adjustment set into three types of variables $\overline{\mathbf{C}}_{it} \equiv \{\overline{\mathbf{X}}_{it}, \overline{\mathbf{X}^*}_{it}, \widetilde{\mathbf{X}}_i\}$ where (1) $\overline{\mathbf{X}}_{it}$, the time-varying variables that are descendants of Y_{it}, (2) $\overline{\mathbf{X}^*}_{it}$, the time-varying variables that are not descendants of Y_{it}, and (3) $\widetilde{\mathbf{X}}_i$, the time-invariant variables.

Without loss of generality, first define an unobserved confounder U such that the no omitted confounder assumption holds conditional on U_{it} and the original adjustment set $\overline{\mathbf{C}}_{it}$, i.e., $Y_{it}(d^L) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid U_{it}, \overline{\mathbf{C}}_{it}$. For simpler illustrations, we assume here that this U_{it} is a descendant of Y_{it} (general results are in Appendix A.3). Theorem 1 then implies that observed simultaneous outcomes are independent conditional on $U_{i,t-1}$ and $\overline{\mathbf{C}}_{it}^P$, i.e., $Y_{i,t-1} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid U_{i,t-1}, \overline{\mathbf{C}}_{it}^P$.

With this setup, we introduce an assumption necessary for the bias correction; the effect and imbalance of unobserved confounders are constant over time. This is an extension of structural stationarity (Assumption 2): while structural stationarity only requires that the existence of causal relationships among outcomes and confounders be time-invariant, this additional causal assumption requires that some of such causal relationships should have the same effect size over time.

Assumption 3 (Time-Invariant Effect and Imbalance of Unobserved Confounder)

1. Time-invariant effect of unobserved confounder U: For all $u_1, u_0, \overline{\mathbf{x}}$ and $\overline{\mathbf{c}}$,

$$\begin{aligned} & \mathbb{E}[Y_{it}(d^L) \mid U_{it} = u_1, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}_{it}^B = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) \mid U_{it} = u_0, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}_{it}^B = \overline{\mathbf{c}}] \\ = \; & \mathbb{E}[Y_{i,t-1} \mid U_{i,t-1} = u_1, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}_{it}^B = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid U_{i,t-1} = u_0, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}_{it}^B = \overline{\mathbf{c}}]. \end{aligned}$$

2. Time-invariant imbalance of unobserved confounder U: For all $u, \overline{\mathbf{x}}$ and $\overline{\mathbf{c}}$,

$$\Pr(U_{it} \leq u \mid D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}_{it}^B = \overline{\mathbf{c}}) - \Pr(U_{it} \leq u \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}_{it}^B = \overline{\mathbf{c}})$$

$$= \Pr(U_{i,t-1} \leq u \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}) - \Pr(U_{i,t-1} \leq u \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}).$$

where $\overline{\mathbf{C}}^B_{it} \equiv \{\overline{\mathbf{X}^*}_{it}, \overline{\mathbf{X}^*}_{i,t-1}, \widetilde{\mathbf{X}}_i, \mathbf{Y}_{\mathcal{N}_i,t-1}\}$.

Assumption 3.1 requires that the effect of unobserved confounders on the potential outcomes be stable over time. This assumption is more plausible when we can control for a variety of observed time-varying confounders $\overline{\mathbf{X}}_{it}$ and $\overline{\mathbf{X}}_{i,t-1}$. However, this assumption might be violated when the change in the effect of U is quick and cannot be explained by observed covariates $\overline{\mathbf{X}}$. Suppose that the unemployment rate is the unobserved confounder in our motivating application. This assumption then implies that the effect of the unemployment rate on the incidence of hate crimes is the same over time. In the causal DAG in Figure 3, this means that the effect of G_2 on Y_{22} is the same as the effect of G_1 on Y_{21}.

Assumption 3.2 requires that the imbalance of unobserved confounders be stable over time. In other words, the strength of association between the treatment variable and unobserved confounders is the same at time t and $t-1$. Importantly, it does not require that the distribution of confounders is the same across different treatment groups. Instead, it requires that the difference between treatment groups be stable over time. For example, this means that an association between the incidence of hate crimes in neighborhoods (treatment) and the unemployment rate is stable over. In the causal DAG in Figure 3, this assumption implies that the association between G_2 and Y_{11} is the same as the one between G_1 and Y_{11}. This assumption substantively means the stability of omitted confounder G.

In practice, both conditions are more likely to hold when the interval between time t and $t-1$ is shorter because $U_{it} \approx U_{i,t-1}$ and $\overline{\mathbf{X}}_{it} \approx \overline{\mathbf{X}}_{i,t-1}$. In particular, when all confounders are time-invariant between time t and $t-1$, Assumption 3.2 holds exactly. Even when confounders are time-varying, we can make these assumptions more plausible by adjusting for observed time-varying confounders $\overline{\mathbf{X}}_{it}$ and $\overline{\mathbf{X}}_{i,t-1}$.

In a special case where there is no descendant of Y_{it} in the adjustment set, i.e., $\overline{\mathbf{X}}_{it} = \overline{\mathbf{X}}_{i,t-1} = \emptyset$, Assumption 3 is equivalent to the parallel trend assumption required for the standard difference-in-differences estimator (Angrist and Pischke, 2008). By allowing for time-varying confounders, Assumption 3 extends the parallel trend assumption. It is also closely connected to the change-in-change method (Athey and Imbens, 2006; Sofer *et al.*, 2016). Specifically, Assumption 3.2 (time-invariant imbalance) is a direct extension of Assumption 3.3 in Athey and Imbens (2006) to the diffusion setting.

4.3.3 Estimator and Identification

We introduce a general bias-corrected estimator under Assumption 3. Intuitively, it subtracts bias detected by the proposed placebo test from an estimator that we would use under the no

omitted confounders assumption.

Definition 3 (Bias-Corrected Estimator)

A bias-corrected estimator $\hat{\tau}_{\text{BC}}$ is the difference between two estimators $\hat{\tau}_{\text{Main}}$ and $\hat{\delta}_{\text{Placebo}}$.

$$\hat{\tau}_{\text{BC}} \equiv \hat{\tau}_{\text{Main}} - \hat{\delta}_{\text{Placebo}} \tag{12}$$

where

$$\hat{\tau}_{\text{Main}} \equiv \int \{\widehat{\mathbb{E}}[Y_{it} \mid D_{it} = d^H, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}] - \widehat{\mathbb{E}}[Y_{it} \mid D_{it} = d^L, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}]\} dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} \mid D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}),$$

$$\hat{\delta}_{\text{Placebo}} \equiv \int \{\widehat{\mathbb{E}}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1}, \overline{\mathbf{C}}^B_{it}] - \widehat{\mathbb{E}}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1}, \overline{\mathbf{C}}^B_{it}]\} dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} \mid D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}),$$

where $\widehat{\mathbb{E}}[\cdot]$ is any unbiased estimator of $\mathbb{E}[\cdot]$, and researchers can use regression, weighting, matching or other techniques to obtain such an unbiased estimator. Note that both estimators are marginalized over the same conditional distribution $F_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} \mid D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}})$.

This bias-corrected estimator consists of two parts, $\hat{\tau}_{\text{Main}}$ and $\hat{\delta}_{\text{Placebo}}$. The first part is an estimator unbiased for the ACDT under the no omitted confounders assumption. However, $\hat{\tau}_{\text{Main}}$ suffers from bias when this identification assumption is violated. The purpose of the second part $\hat{\delta}_{\text{Placebo}}$ is to correct this bias. It is closely connected to the proposed placebo test; when the assumption of no omitted confounders holds, $\mathbb{E}[\hat{\delta}_{\text{Placebo}}] = 0$ and there is no bias correction. When the assumption is instead violated, $\hat{\delta}_{\text{Placebo}}$ serves as an estimator of the bias. We rely on $\widehat{\text{Var}}(\hat{\tau}_{\text{Main}}) + \widehat{\text{Var}}(\hat{\delta}_{\text{Placebo}})$ as a conservative variance estimator of the bias-corrected estimator given that $\hat{\tau}_{\text{Main}}$ and $\hat{\delta}_{\text{Placebo}}$ are often positively correlated. In our motivating application, we rely on the spacial heteroskedasticity and autocorrelation consistent (spatial HAC) variance estimator by Conley (1999) to compute each variance, while accounting for spatial autocorrelation of errors.

The theorem below shows that under Assumption 3, the bias-corrected estimator is unbiased for the ACDT.

Theorem 2 (Identification with A Bias-Corrected Estimator) Under Assumption 3, the proposed bias-corrected estimator is unbiased for the ACDT.

$$\mathbb{E}[\hat{\tau}_{\text{BC}}] \quad = \quad \tau_t^H(d^H, d^L).$$

The proof is in Appendix A.3. It is also true that this estimator is unbiased for the ACDT when the no omitted confounders assumption holds. Through a simulation study calibrated to the hate crime data, we show that the proposed bias-corrected estimator can reduce the bias and root mean squared error even when the required time-invariance assumption (Assumption 3) is slightly violated (Appendix D.2). We also introduce a sensitivity analysis method in Appendix A.4 to investigate the robustness of the bias-corrected estimates to the potential violation of the time-invariance assumption (Assumption 3).

5 Empirical Analysis

Applying the proposed methods, we estimate the ACDE of hate crimes against refugees in Germany. We begin with the setup of data analysis (Section 5.1) and then turn to estimation of the ACDE (Section 5.2) and heterogeneous effects (Section 5.3).

5.1 Setup

As one of the most well-studied outcomes, we focus on physical attacks against refugees as the main dependent variable. Formally, we define the outcome variable Y_{it} to be binary, taking the value 1 if there exists any physical attack against refugees at county i in month t, and taking the value 0 otherwise. The outcomes are defined for 402 counties in Germany every month from the beginning of 2015 to the end of 2016. Averaging over all counties in Germany during this period, the sample mean of the outcome variable is 6.4%. This means that 6.4% of counties experienced at least one physical attack in a typical month. In Saxony, a state with the largest number of hate crimes, the sample mean of the outcome variable is 34%.

We use a distance matrix to encode the physical proximity between counties. In particular, we construct an initial distance matrix $\widetilde{\mathbf{W}}$ using an inverse of the straight distance between counties i and j as $\widetilde{W}_{ij}$. We then row-standardize the initial matrix $\widetilde{\mathbf{W}}$ and obtain a final distance matrix $\mathbf{W}$. For the outcome variable in month t, the treatment variable is defined to be $D_{it} \equiv \mathbf{W}_i^\top \mathbf{Y}_{t-1}$, the weighted proportion of neighboring counties that experience the incidence of physical attacks in month $t-1$. The first causal quantity of interest is the ACDE, which quantifies how much the probability of having hate crimes changes due to the increase in the proportion of neighboring counties that have experienced hate crimes last month.

To investigate how the proposed methods detect and correct biases, we consider five different adjustment sets in order (summarized in Table 1). As the first adjustment set, we include one-month lagged dependent and treatment variables. We also adjust for basic summary statistics of $\mathbf{W}_i$, i.e., the number of neighbors and variance of $\mathbf{W}_i$, in order to compare observations with similar spatial characteristics. These lagged variables and basic summary statistics of the spatial distance are sufficient for identification if the spatial diffusion is the only mechanism through which neighboring counties exhibit similar outcomes. Then, as the second adjustment set, we add two-month lagged dependent variables to see whether adjusting for a longer history of past outcomes can reduce bias (e.g., Christakis and Fowler, 2013; Eckles and Bakshy, 2017). The third adjustment set adds state fixed effects. Although the state fixed effects are often excluded from existing studies (e.g., Jäckle and König, 2016), we show how much these fixed effects help remove biases. Then, the fourth set adds a list of contextual

C1	$Y_{i,t-1}, D_{i,t-1}$, summary statistics of $\mathbf{W}_i(\lvert\mathcal{N}_i\rvert, \text{Var}(\mathbf{W}_i))$
C2	C1 + $Y_{i,t-2}$
C3	C2 + state fixed-effects
C4	C3 + contextual variables studied in the literature
C5	C4 + time trend (third-order polynomials)

Table 1: Five Different Adjustment Sets.

variables related to the number of refugees, demographics, education, general crimes, economic indicators, and politics. Finally, the fifth set adjusts for the time trend using third-order polynomials. We provide details of the five adjustment sets and the corresponding placebo sets in Appendix E.

For the proposed placebo test, we rely on the structural stationarity assumption (Assumption 2). For example, if discussions of the refugee crisis in newspapers, which we do not measure, are confounders, structural stationarity requires that such discussions in newspapers remain confounders throughout 2015 and 2016. Importantly, the placebo test is valid even when the tone of discussions is changing over time (unmeasured time-varying confounders) and the effect of discussions changes over time. For the bias-corrected estimator, the time-invariance assumption (Assumption 3) requires a stronger assumption, similar to the difference-in-differences literature (Athey and Imbens, 2006; Angrist and Pischke, 2008; Sofer *et al.*, 2016), that the effect of newspapers is stable over time and the imbalance of unobserved discussions in newspapers is stable over time after adjusting for observed time-varying confounders.

5.2 Estimation of Average Causal Diffusion Effect

To estimate the ACDE, we use the following logistic regression to model the main outcome variable Y_{it} with the treatment variable and each of the five adjustment sets.

$$\text{logit}(\Pr(Y_{it} = 1 \mid D_{it}, \overline{\mathbf{C}}_{it})) \;= \alpha + \beta D_{it} + \gamma^\top \overline{\mathbf{C}}_{it}, \tag{13}$$

where D_{it} is the treatment variable and $\overline{\mathbf{C}}_{it}$ is a specified adjustment set. Under the assumption of no omitted confounders, the difference in the estimated probabilities of Y_{it} under $D_{it} = d^H$ and $D_{it} = d^L$ serves as an estimator for the ACDE. In particular, we estimate the ACDE that compares the following two treatment values; $d^H = 27\%$, the treatment received by the average counties in Saxony (a state with the largest number of hate crimes) and $d^L = 0\%$, none of the neighbors experiencing hate crimes (common for safe areas in West Germany). Formally, $\hat{\tau} \equiv \int \{\widehat{\Pr}(Y_{it} = 1 \mid D_{it} = 0.27, \overline{\mathbf{C}}_{it}) - \widehat{\Pr}(Y_{it} = 1 \mid D_{it} = 0, \overline{\mathbf{C}}_{it})\} dF_{\overline{\mathbf{C}}_{it}}(\overline{\mathbf{c}})$.

To assess the no omitted confounders assumption, we also estimate the following placebo

logistic regression.

$$\text{logit}(\Pr(Y_{i,t-1} = 1 \mid D_{it}, \overline{\mathbf{C}}_{it}^P)) \ = \alpha_0 + \rho D_{it} + \gamma_0^\top \overline{\mathbf{C}}_{it}^P, \tag{14}$$

where $Y_{i,t-1}$ is the placebo outcome and $\overline{\mathbf{C}}_{it}^P$ is a placebo set corresponding to the adjustment set $\overline{\mathbf{C}}_{it}$. When the no omitted confounders assumption holds, Theorem 1 implies that $\rho = 0$. We use the difference in the estimated probabilities of $Y_{i,t-1}$ under $D_{it} = d^H$ and $D_{it} = d^L$ as a test statistic of the placebo test. Formally, $\hat{\delta} \equiv \int \{\widehat{\Pr}(Y_{i,t-1} = 1 \mid D_{it} = 0.27, \overline{\mathbf{C}}_{it}^P) - \widehat{\Pr}(Y_{i,t-1} = 1 \mid D_{it} = 0, \overline{\mathbf{C}}_{it}^P)\} dF_{\overline{\mathbf{C}}_{it}^P}(\overline{\mathbf{c}}^P)$.

To account for spatial and temporal autocorrelation of errors, we use the spatial HAC variance estimator by Conley (1999) to compute standard errors by allowing for arbitrary spatial dependence between units within 100 km and temporal dependence within units over six months. As a robustness check, we also compute standard errors clustered at the state level, which can allow for any spatial and temporal dependence between units within the same state. The results are similar to those based on the spatial HAC variance estimator we report below.

Figures 4 (a) and (b) present results from the placebo tests (equation (14)) and estimates from the main model (equation (13)) with 95% confidence intervals, respectively. C1, C2, C3, C4, and C5 refer to the five different adjustment sets we introduced before. When a given adjustment set satisfies the no omitted confounders assumption, estimates from the placebo tests should be close to zero. Figure 4 (a) shows that while the first four adjustment sets are not sufficient, the fifth set (C5) successfully adjusts for confounders; a placebo estimate is close to zero and its 95% confidence interval covers zero. It is not enough to adjust for lagged dependent variables and contextual variables and it is critical to adjust for the time trend flexibly.

On the basis of these results from the placebo tests, we can now investigate estimates of the ACDE from the main model (equation (13)) in Figure 4 (b). For the first two cases (C1 and C2), estimates are as large as 5 percentage points, but the placebo tests suggest that these estimates are heavily biased. Similarly, while the next two cases show point estimates of around 2 percentage points, they are also likely to be biased. When we focus on the fifth adjustment set, which produces a placebo estimate close to zero, a point estimate of the ACDE is smaller than 1 percentage point, and its 95% confidence interval covers zero. The comparison between this more credible estimate and the one from the fourth set shows that an estimate of the ACDE can suffer from 100% bias by missing just one variable. This demonstrates the importance of bias detection in causal diffusion analysis.

Although the proposed placebo tests suggest that the fifth set successfully adjusts for

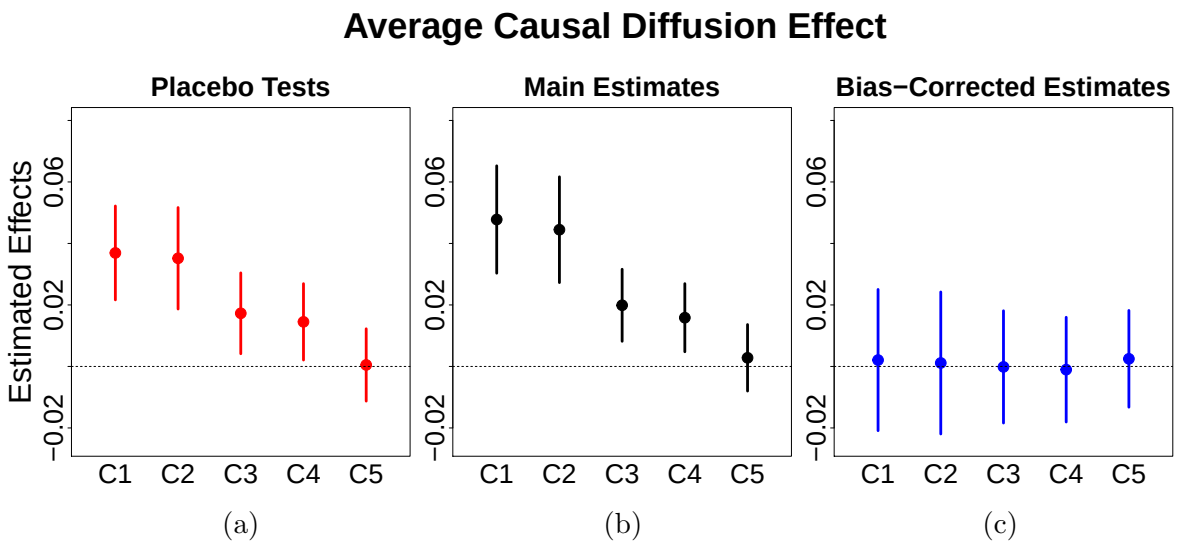


Figure 4: Placebo Tests, Main Estimates, and Bias-Corrected Estimates of the ACDE.
Note: Figures (a), (b) and (c) present results from the placebo tests, estimates of the ACDE under the no omitted confounders assumption, and estimates from bias-corrected estimators with 95% confidence intervals, respectively.

relevant confounders in this analysis, it is often infeasible to find such adjustment sets in many other applications. To address these common scenarios, we now examine whether researchers could obtain similar results using a bias-corrected estimator even with adjustment sets that reject the null hypothesis of the placebo test.

Figure 4 (c) shows that bias-corrected estimates are similar regardless of the selection of adjustment sets and they all cover the most credible point estimate from the fifth control set. Even though the proposed placebo test detected a large amount of bias, researchers can obtain credible estimates by correcting the biases in this example.

These results suggest that, in contrast to existing studies (Braun, 2011; Jäckle and König, 2016), the ACDE on the incidence of hate crimes is small when averaging over all counties in Germany. In the next subsection, we show that the spatial diffusion of hate crimes is concentrated among a small subset of counties that have a higher proportion of school dropouts.

5.3 Heterogeneous Diffusion Effects by Education

Now, we extend the previous analysis by considering the types of counties that are more susceptible to the diffusion of hate crimes. In particular, we examine the role of education. Given rich qualitative and quantitative evidence that hate crime is often a problem of young people, it is critical to take into account one of the most important institutional contexts around them, i.e., schooling. The literature has discussed at least three mechanisms through which education can reduce the risk of hate crimes. First, education increases economic returns to current and

future legitimate work, thereby raising the opportunity cost of committing hate crimes (e.g., Lochner and Moretti, 2004). Second, education may change the psychological costs associated with hate crimes. More educated people tend to have lower levels of ethnocentrism and place more emphasis on cultural diversity (Hainmueller and Hiscox, 2007). Finally, schooling has incapacitation effects – keeping adolescents busy and off the street, thereby directly reducing the chances of committing crimes (Jacob and Lefgren, 2003).

Building on the literature above, we investigate whether local educational contexts condition the spatial diffusion dynamics of hate crimes. We use a proportion of school dropouts without a secondary school diploma as a measure of local educational performance. To better disentangle the education explanation, we analyze East Germany and West Germany separately because they have substantially different distributions of proportions of school dropouts (counties in East Germany have much higher proportions of school dropouts). Here we report results from East Germany and provide those for West Germany in Appendix E. In particular, we estimate the conditional average causal diffusion effects (conditional ACDEs) for counties that have high and low proportions of school dropouts without a secondary school diploma. We use 9% as a cutoff for high and low proportions of school dropouts, which is approximately the median value in East Germany. We add an interaction term between the treatment variable and this indicator variable to the original model in equation (13) and to the original placebo model in equation (14).

Figure 5 presents results for the conditional ACDE for counties that have a higher proportion of school dropouts. Similar to the case of the ACDE estimation, Figure 5 (a) shows strong concerns of biases in the first four adjustment sets. Even though a 95% confidence interval of the fourth estimate covers zero, its point estimate is far from zero (around 4 percentage points). In contrast, the placebo test suggests that the fifth set adjusts for relevant confounders where a placebo estimate is close to zero.

Based on results from the placebo tests, we examine estimates from the main model in Figure 5 (b). The first four sets, likely to be biased, exhibit large point estimates, larger than 10 percentage points. More interestingly, even with the most credible fifth adjustment set, a point estimate is as large as 6 percentage points and is statistically significant. This effect size is substantively important given that it is about one-fourth of the sample average outcome in this subset (26%). Bias-corrected estimates in Figure 5 (c) confirm that the conditional ACDE for counties with a higher proportion of school dropouts is large and similar regardless of the selection of adjustment sets, while it is not statistically significant.

When we estimate the conditional ACDE for counties that have a lower proportion of

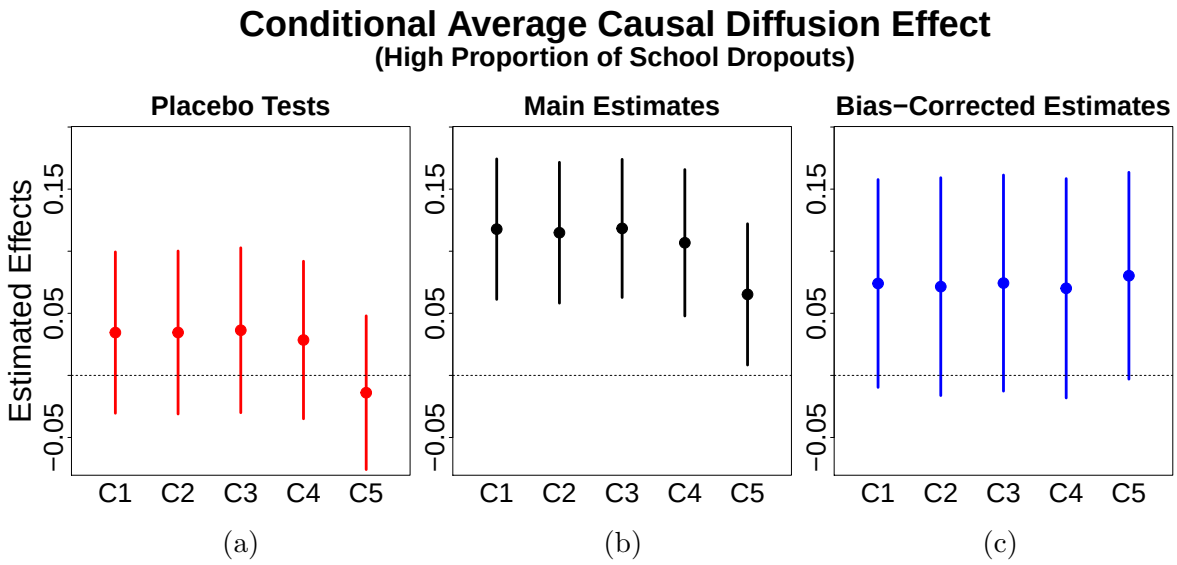


Figure 5: Placebo Tests, Main Estimates, and Bias-Corrected Estimates of the conditional ACDE for counties with a high proportion of school dropouts. Note: Figures (a), (b) and (c) present results from the placebo tests, estimates of the conditional ACDE under the no omitted confounders assumption, and estimates from bias-corrected estimators with 95% confidence intervals, respectively.

school dropouts, effects are close to zero and their 95% confidence intervals cover zero, as the education hypothesis expects (see Appendix E). Causal diffusion effects are also precisely estimated to be zero in West Germany, where the proportions of school dropouts are much lower than in East Germany. This additional analysis suggests that the spatial diffusion dynamics of hate crimes operate, if any, only in places with low educational performance and thus, prevention policies can have positive multiplier effects only when targeting areas with low educational performance.

6 Concluding Remarks

Causal diffusion dynamics have been an integral part of many social and biomedical science theories. Given that spatial and network panel data have become increasingly common, it is essential to develop methodologies to draw causal inference for diffusion effects. However, causal diffusion analysis has been challenging due to two well-known types of biases, i.e., contextual confounding and homophily bias. Recognizing that causal inference for diffusion effects is generally impossible without further assumptions (Shalizi and Thomas, 2011; VanderWeele and An, 2013; Ogburn, 2018), this paper examines identification of causal diffusion effects using placebo outcomes under a new assumption of structural stationarity. This structural stationarity requires the existence of causal relationships among variables — not the effect or sign of such relationships — to be stable over time. Instead of directly assuming the validity

of placebo outcomes, we show that we can transparently choose and justify placebo outcomes for identifying causal diffusion effects under the structural stationarity assumption.

Under structural stationarity, we first propose a statistical placebo test that can detect a wide class of biases, including contextual confounding and homophily bias. Then, we develop a difference-in-differences style estimator that can directly correct biases under an additional causal assumption. Applying the proposed methods to geo-coded hate crime data, we examined the spatial diffusion of hate crimes in Germany. After removing upward bias in previous studies, we found that the average effect of spatial diffusion is small, in contrast to recent quantitative analyses (Braun, 2011; Jäckle and König, 2016). This empirical analysis demonstrates the large differences in substantive conclusions that can result from contextual confounding.

References

An, W. (2015). Instrumental Variables Estimates of Peer Effects in Social Networks. *Social Science Research*, **50**, 382–394.

Angrist, J. D. (2014). The Perils of Peer Effects. *Labour Economics*, **30**, 98–108.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

Anselin, L. (2013). *Spatial Econometrics: Methods and Models*. Springer.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing Influence-based Contagion from Homophily-driven Diffusion in Dynamic Networks. *Proceedings of the National Academy of Sciences*, **106**(51), 21544–21549.

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*.

Athey, S. and Imbens, G. W. (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica*, **74**(2), 431–497.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of Peer Effects through Social Networks. *Journal of Econometrics*, **150**(1), 41–55.

Braun, R. (2011). The Diffusion of Racist Violence in the Netherlands: Discourse and Distance. *Journal of Peace Research*, **48**(6), 753–766.

Cai, X., Loh, W. W., and Crawford, F. W. (2019). Identification of Causal Intervention Effects Under Contagion. *arXiv preprint arXiv:1912.04151*.

Christakis, N. A. and Fowler, J. H. (2013). Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine*, **32**(4), 556–577.

Cohen-Cole, E. and Fletcher, J. M. (2008). Is Obesity Contagious? Social Networks vs. Environmental Factors in the Obesity Epidemic. *Journal of health economics*, **27**(5), 1382–1387.

Conley, T. G. (1999). GMM Estimation with Cross Sectional Dependence. *Journal of Econometrics*, **92**(1), 1–45.

Dancygier, R. M., Egami, N., Jamal, A. A., and Rischke, R. (2020). Hate Crimes and Gender Imbalances: Fears over Mate Competition and Violence against Refugees. *American Journal of Political Science*.

Danks, D. and Plis, S. (2013). Learning Causal Structure from Undersampled Time Series. In *NIPS 2013 Workshop on Causality*.

Dean, T. and Kanazawa, K. (1989). A Model for Reasoning About Persistence and Causation. *Computational intelligence*, **5**(2), 142–150.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and The Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, **101**(5), 1739–74.

Eckles, D. and Bakshy, E. (2017). Bias and High-Dimensional Adjustment in Observational Studies of Peer Effects. *arXiv preprint arXiv:1706.04692*.

Egami, N. and Tchetgen Tchetgen, E. J. (2021). Identification and Estimation of Causal Peer Effects Using Double Negative Controls for Unmeasured Network Confounding. *arXiv preprint arXiv:2109.01933*.

Fowler, J. H., Heaney, M. T., Nickerson, D. W., Padgett, J. F., and Sinclair, B. (2011). Causality in Political Networks. *American Politics Research*, **39**(2), 437–480.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and Social interactions. *The Quarterly Journal of Economics*, **111**(2), 507–548.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social Networks and the Identification of Peer Effects. *Journal of Business & Economic Statistics*, **31**(3), 253–264.

Graham, E. R., Shipan, C. R., and Volden, C. (2013). The Diffusion of Policy Diffusion Research in Political Science. *British Journal of Political Science*, **43**(03), 673–701.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, **78**(6), 1360–1380.

Hainmueller, J. and Hiscox, M. J. (2007). Educated Preferences: Explaining Attitudes toward Immigration in Europe. *International Organization*, **61**(2), 399–442.

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

Halloran, M. E. and Struchiner, C. J. (1995). Causal Inference in Infectious Diseases. *Epidemiology*, **6**(2), 142–151.

Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., and Danks, D. (2016). Causal Discovery from Subsampled Time Series Data by Constraint Optimization. In *Proceedings of the 8th International Conference on Probabilistic Graphical Models (PGM)*, pages 216–227.

Jäckle, S. and König, P. D. (2016). The Dark Side of the German 'Welcome Culture': Investigating the Causes behind Attacks on Refugees in 2015. *West European Politics*, **40**(2), 223–251.

Jacob, B. A. and Lefgren, L. (2003). Are Idle Hands the Devil's Workshop? Incapacitation, Concentration, and Juvenile Crime. *American Economic Review*, **93**(5), 1560–1577.

Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., and Fowler, J. H. (2017). Social Influence and Political Mobilization: Further Evidence from a Randomized Experiment in the 2012 US Presidential Election. *PloS one*, **12**(4), e0173851.

Lipsitch, M., Tchetgen Tchetgen, E. J., and Cohen, T. (2010). Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*, **21**(3), 383.

Liu, L. and Tchetgen Tchetgen, E. (2020). Regression-based Negative Control of Homophily in Dyadic Peer Effect Analysis. *arXiv preprint arXiv:2002.06521*.

Lochner, L. and Moretti, E. (2004). The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *American Economic Review*, **94**(1), 155–189.

Lyons, R. (2011). The Spread of Evidence-Poor Medicine via Flawed Social-Network Analysis. *Statistics, Politics, and Policy*, **2**(1).

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *The Review of Economic Studies*, **60**(3), 531–542.

McFowland III, E. and Shalizi, C. R. (2021). Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations. *Journal of the American Statistical Association*, pages 1–27.

Miao, W. and Tchetgen Tchetgen, E. J. (2017). Invited Commentary: Bias Attenuation and Identification of Causal Effects with Multiple Negative Controls. *American Journal of Epidemiology*, **185**(10), 950–953.

Morozova, O., Cohen, T., and Crawford, F. W. (2018). Risk Ratios for Contagious Outcomes. *Journal of The Royal Society Interface*, **15**(138), 20170696.

Myers, D. J. (2000). The Diffusion of Collective Violence: Infectiousness, Susceptibility, and Mass Media Networks. *American Journal of Sociology*, **106**(1), 173–208.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles (with discussion). Section 9 (translated). *Statistical Science*, **5**(4), 465–472.

Ogburn, E. L. (2018). Challenges to Estimating Contagion Effects from Observational Data. In *Complex Spreading Phenomena in Social Systems*, pages 47–64. Springer.

Ogburn, E. L. and VanderWeele, T. J. (2014). Causal Diagrams for Interference. *Statistical Science*, **29**(4), 559–578.

Ogburn, E. L., Shpitser, I., and Lee, Y. (2020). Causal Inference, Social networks, and Chain Graphs. *The Journal of the Royal Statistical Society, Series A*.

O'Malley, A. J., Elwert, F., Rosenquist, J. N., Zaslavsky, A. M., and Christakis, N. A. (2014). Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables. *Biometrics*, **70**(3), 506–515.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge.

Pearl, J. and Russell, S. (2001). Bayesian Networks. In *Handbook of Brain Theory and Neural Networks*. MIT Press.

Robins, J. (1986). A New Approach To Causal Inference In Mortality Studies With A Sustained Exposure Period — Application to Control of the Healthy Worker Survivor Effect. *Mathematical modelling*, **7**(9-12), 1393–1512.

Rogers, E. M. (1962). *Diffusion of Innovations*. Simon and Schuster.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**(5), 688.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *The Quarterly Journal of Economics*, **116**(2), 681–704.

Shalizi, C. R. and Thomas, A. C. (2011). Homophily and Contagion are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research*, **40**(2), 211–239.

Sinclair, B. (2012). *The Social Citizen: Peer Networks and Political Behavior*. University of Chicago Press.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen Tchetgen, E. J. (2016). On Negative Outcome Control of Unobserved Confounding as a Generalization of Difference-in-Differences. *Statistical Science*, **31**(3), 348.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press.

Tchetgen Tchetgen, E. J., Fulcher, I., and Shpitser, I. (2020). Auto-G-Computation of Causal Effects on a Network. *Journal of American Statistical Association*.

United Nations High Commissioner for Refugees. (2017). Global Trends: Forced Displacement in 2017.

VanderWeele, T. J. and An, W. (2013). Social Networks and Causal Inference. In *Handbook of Causal Analysis for Social Research*, pages 353–374. Springer.

VanderWeele, T. J., Ogburn, E. L., and Tchetgen Tchetgen, E. J. (2012). Why and When "Flawed" Social Network Analyses Still Yield Valid Tests of No Contagion. *Statistics, Politics and Policy*, **3**(1).

Wilson, J. Q. and Kelling, G. L. (1982). Broken Windows. *Atlantic Monthly*, **249**(3), 29–38.

Supplementary Appendix

A Proofs

A.1 Proof of Theorem 1

In this proof, we use $\overline{\mathbf{C}}$ and $\overline{\mathbf{C}}^P$ to denote $\overline{\mathbf{C}}_{it}$ and $\overline{\mathbf{C}}_{it}^P$ for notational simplicity.

A.1.1 Setup

Given that adjustment set $\overline{\mathbf{C}}$ are defined to be pre-treatment (i.e., variables not affected by the treatment), theoretical results on causal DAGs (Pearl, 1995; Shpitser *et al.*, 2012) imply that $Y_{it}(d) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}$ is equivalent to no unblocked back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it} with respect to $\overline{\mathbf{C}}$ in causal DAG $\mathcal{G}$ (see Lemma 1). Additionally, $Y_{i,t-1} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}^P$ is equivalent to no unblocked back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to $Y_{i,t-1}$ with respect to $\overline{\mathbf{C}}^P$ in causal DAG $\mathcal{G}$.

The theorem requires one regularity condition – the violation of the no omitted confounders assumption, if any, is *proper*. Intuitively, it means that bias (i.e., the violation of the no omitted confounders assumption) is in fact driven by omitted variables. Bias is not proper when the only source of bias is the misadjustment of the lag structure of observed covariates. Importantly, contextual confounding and homophily bias are proper, and hence within the scope of this theorem.

Definition 1 (Proper Bias)

Suppose adjustment set $\overline{\mathbf{C}}$ does not satisfy Assumption 1. This violation (bias) is defined to be proper when it satisfies the following condition: If control set $\overline{\mathbf{C}}_{it}$ cannot block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it}, there is at least one back-door path that any subset of the following set cannot block.

$$\{\overline{\mathbf{C}}_{it}, \overline{\mathbf{C}}_{it}^{(-1)}, \overline{\mathbf{C}}_{it}^{(+1)}, \mathbf{Y}_{\mathcal{N}_i,t-2}\},$$

where $\overline{\mathbf{C}}_{it}^{(-1)}$ and $\overline{\mathbf{C}}_{it}^{(+1)}$ are a lag and a lead of the time-dependent variables in $\overline{\mathbf{C}}_{it}$.

A.1.2 Bias $\rightarrow$ Dependence in Placebo Test

Here, we show that when set $\overline{\mathbf{C}}$ cannot block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it}, set $\overline{\mathbf{C}}^P$ cannot block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to $Y_{i,t-1}$.

Step 1 (Proper Bias): Given the assumption that the set $\overline{\mathbf{C}}$ is proper, set $\overline{\mathbf{C}}^P$ cannot block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it} because $\overline{\mathbf{C}}^P$ is a subset of $\{\overline{\mathbf{C}}, \overline{\mathbf{C}}^{(-1)}, \overline{\mathbf{C}}^{(+1)}, \mathbf{Y}_{\mathcal{N}_i,t-2}\}$.

Step 2 (Set up the main unblocked back-door path to investigate): Let π be a back-door path from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it} that both $\overline{\mathbf{C}}$ and $\overline{\mathbf{C}}^P$ and any subset of $\{\overline{\mathbf{C}}, \overline{\mathbf{C}}^{(-1)}, \overline{\mathbf{C}}^{(+1)}, \mathbf{Y}_{\mathcal{N}_i,t-2}\}$ cannot block. Without loss of generality, we assume that this unblocked back-door path starts with an arrow pointing to $Y_{k,t-1}$ where $k \in \mathcal{N}_i$ and it ends with an arrow pointing to Y_{it}.

Step 3 (Case I. the last node of the unblocked back-door path is time-independent): First, consider a case in which the last variable in an unblocked back-door path has a directed arrow pointing to Y_{it} and time-independent. Let (Z, Y_{it}) denote the last two node path segment on π where Z is a time-independent variable and there exists a directed arrow from Z to Y_{it}. Note that we do not put any individual index to Z because the proof holds for any index. Since this is an unblocked path, Z is not in $\overline{\mathbf{C}}^P$ and there is an unblocked back-door path from $Y_{k,t-1}$ to Z. Since Z is time-independent, there is a directed arrow from Z to $Y_{i,t-1}$ by the structural stationarity (Assumption 2). Therefore, set $\overline{\mathbf{C}}^P$ cannot block this back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$.

Step 4 (Case II. the last node of the unblocked back-door path is time-dependent): Next, consider the case in which the last variable in an unblocked back-door path points to Y_{it} and time-dependent. Let (B, X_t, Y_{it}) denote the last three node path segment on π where X_t is a time-dependent direct cause of Y_{it}. Note that we do not put any individual index to X_t because the proof holds for any index. $X_{t-1}, X_t \notin \overline{\mathbf{C}}^P$ because $X_t \notin \overline{\mathbf{C}}$ (see Lemma 2 in Section A.2).

Step 4.1 (sub-Case: the second last node is time-independent): First, assume B is time-independent. Then, because a causal DAG satisfies structural stationarity (Assumption 2), X_{t-1} and B have the same relationship as the one between X_t and B. In addition, since there is an unblocked path from $Y_{k,t-1}$ to X_t to through B, there exists an unblocked path from $Y_{k,t-1}$ to X_{t-1} through B. Given that there exists a directed arrow from X_t to Y_{it}, there exists a directed arrow from X_{t-1} to $Y_{i,t-1}$. Therefore, there is an unblocked back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$.

Step 4.2 (sub-Case: the second last node is time-dependent): Next, assume B is time-dependent and therefore we use B_t. First, we show that whenever B is time-dependent, then the directed arrow is always from X_t to B_t. Suppose there is a directed arrow from B_t

to X_t. If B_t in $\overline{\mathbf{C}}^P$, then this back-door is blocked (therefore, choose another π). So, B_t is not in $\overline{\mathbf{C}}^P$. Therefore, we can collapse B_t into X_t, meaning that if B is time dependent, then the directed arrow is always from X_t to B_t.

Now, suppose there is a directed arrow from X_t to B_t. We know there exists an unblocked path from $Y_{k,t-1}$ to X_t through B_t. Now, because $Y_{i,t-1} \leftarrow X_{t-1} \rightarrow X_t \rightarrow B_t$, there is an unblocked back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$ because the underlying causal DAG satisfies structural stationarity. □

A.1.3 No Bias → Independence in Placebo Test

Next, we prove that when set $\overline{\mathbf{C}}$ can block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it}, set $\overline{\mathbf{C}}^P$ can block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to $Y_{i,t-1}$. We show the contraposition: when there is a back-door path from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to $Y_{i,t-1}$ that set $\overline{\mathbf{C}}^P$ cannot block, set $\overline{\mathbf{C}}$ cannot block all back-door paths from $\mathbf{Y}_{\mathcal{N}_i,t-1}$ to Y_{it}. Since $\overline{\mathbf{C}}$ does not include any $\text{Des}(Y_{k,t-1})$, we know $\overline{\mathbf{C}}^P$ also does not include any $\text{Des}(Y_{k,t-1})$. Also, by definition, $\overline{\mathbf{C}}^P$ does not include any $\text{Des}(Y_{i,t-1})$. Therefore, without loss of generality, we can focus on unblocked back-door paths that start with an arrow pointing to $Y_{k,t-1}$ where $k \in \mathcal{N}_i$ and end with an arrow pointing to $Y_{i,t-1}$.

Step 1 (Control Set cannot block all back-door paths to the Placebo outcome): First, we show that when there is a back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$ that set $\overline{\mathbf{C}}^P$ cannot block, set $\overline{\mathbf{C}}$ cannot block all back-door paths from $Y_{k,t-1}$ to $Y_{i,t-1}$. From set $\overline{\mathbf{C}}^P$ to set $\overline{\mathbf{C}}$, we need to (1) add $\text{Des}(Y_{i,t-1})$ and (2) remove $\overline{\mathbf{C}}^{(-1)}$ and $\mathbf{Y}_{\mathcal{N}_i,t-2}$. We show here that this process cannot block a back-door path that set $\overline{\mathbf{C}}^P$ cannot block. The step (1) cannot block the back-door path because adding $\text{Des}(Y_{i,t-1})$ cannot block a back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$ unblocked by set $\overline{\mathbf{C}}^P$ (see Lemma 3 in Section A.2). For (2), we first check whether removing $X_{t-1} \in \overline{\mathbf{C}}^{(-1)}$ can block a back-door path that set $\overline{\mathbf{C}}^P$ cannot block. To begin with, we can remove X_{t-1} because $X_t \in \overline{\mathbf{C}}$. Removing variables X_{t-1} can be helpful if X_{t-1} is a collider or a descendant of a collider for a back-door path. However, if so, X_t is a descendant of a collider and it is in set $\overline{\mathbf{C}}$ and therefore, removing X_{t-1} cannot block any additional paths. Next, we need to check whether removing a variable $B \in \mathbf{Y}_{\mathcal{N}_i,t-2}$ can block the back-door path that the set $\overline{\mathbf{C}}^P$ cannot block. Removing variable B can be helpful if B is a collider or a descendant of a collider for a back-door path. If so, there is an unblocked back-door path (with respect to $\overline{\mathbf{C}}^P$) that starts with an arrow pointing to B and ends with an arrow pointing

to $Y_{i,t-1}$, i.e., $B \leftarrow \ldots \rightarrow Y_{i,t-1}$. Since B has a directed arrow pointing to $Y_{k,t-1}$, removing B unblock a new back-door path from $Y_{k,t-1}$ through B, which points to $Y_{i,t-1}$. Although this unblocked back-door path with respect to $\overline{\mathbf{C}}$ is different from the unblocked back-door path with respect to $\overline{\mathbf{C}}^P$, the paths are the same after node B and therefore at least the last three nodes are the same. Therefore, we can use π to be a back-door from $Y_{k,t-1}$ to $Y_{i,t-1}$ that both sets $\overline{\mathbf{C}}$ and $\overline{\mathbf{C}}^P$ cannot block.

Step 2 (Case I: the last node of the unblocked back-door path is time-independent): Consider the case in which the last two nodes are $(Z \rightarrow Y_{i,t-1})$ and Z is time-independent. Then, since $Z \rightarrow Y_{it}$ from structural stationarity (Assumption 2), set $\overline{\mathbf{C}}$ cannot block this back-door.

Step 3 (Case II: the last node of the unblocked back-door path is time-dependent): Next, consider the case in which the last two nodes are $(X_{t-1} \rightarrow Y_{i,t-1})$. Since $X_{t-1} \notin \overline{\mathbf{C}}^P$ and $X_{t-1} \notin \text{Des}(Y_{i,t-1})$, $X_{t-1}, X_t \notin \overline{\mathbf{C}}$. Therefore, set $\overline{\mathbf{C}}$ cannot block $Y_{k,t-1} \leftarrow \cdots X_{t-1} \rightarrow X_t \rightarrow Y_{it}$. □

A.2 Proof of Lemmas used for Theorem 1

Here, we prove all the lemmas used to prove Theorem 1.

Lemma 1 (Equivalence between Back-Door Criteria and No Omitted Confounder Assumption (Shpitser *et al.*, 2012)) For a pretreatment adjustment set $\overline{\mathbf{C}}$ (i.e., variables not affected by the treatment), the following two statements hold.

1. If a set $\overline{\mathbf{C}}$ satisfies the back-door criterion with respect to $(Y_{it}, \mathbf{Y}_{\mathcal{N}_i,t-1})$ in causal DAG $\mathcal{G}$, then $Y_{it}(d) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}$ holds in every causal model inducing causal DAG $\mathcal{G}$ (Pearl, 1995).

2. If $Y_{it}(d) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}$ holds in every causal model inducing causal DAG $\mathcal{G}$, then a set $\overline{\mathbf{C}}$ satisfies the back-door criterion with respect to $(Y_{it}, \mathbf{Y}_{\mathcal{N}_i,t-1})$ in causal DAG $\mathcal{G}$ (Shpitser *et al.*, 2012).

Lemma 2 $X_t \notin \overline{\mathbf{C}} \rightarrow X_{t-1}, X_t \notin \overline{\mathbf{C}}^P$.

Proof First, we show that $X_{t-1}, X_t, X_{t+1} \notin \overline{\mathbf{C}}$ because set $\overline{\mathbf{C}}$ is proper. It is because if X_{t-1} or X_t are in $\overline{\mathbf{C}}$, then the lag adjustment of the control set $\overline{\mathbf{C}}$ can block this path. If this path

is the only back-door path, then $\overline{\mathbf{C}}$ is not proper. If there is another back-door path that any subset of $\{\overline{\mathbf{C}}, \overline{\mathbf{C}}^{(-1)}, \overline{\mathbf{C}}^{(+1)}, \mathbf{Y}_{\mathcal{N}_i,t-2}\}$ cannot block, choose it as π.

Next, we show that $X_{t-1}, X_t \notin \overline{\mathbf{C}}^P$. There are three ways for a variable to be in the placebo set $\overline{\mathbf{C}}^P$. We discuss them in order. First, a variable can be in the placebo set because it was already in the control set. We know $X_{t-1}, X_t \notin \overline{\mathbf{C}}$, so this option is not feasible. Second, a variable can be in the placebo set because it is a lag of the original control variables. Given that X_t, X_{t+1} are not in the control set, this option is also not feasible. Finally, a variable can be in the placebo set because it is a lag of the treatment variable. (a) It is important to notice that $X_{t-1} \notin \mathbf{Y}_{\mathcal{N}_i,t-2}$ because $X_t \notin \mathbf{Y}_{\mathcal{N}_i,t-1}$ (i.e., the treatment cannot be the last node of the unblocked back-door path). (b) Now, we verify $X_t \notin \mathbf{Y}_{\mathcal{N}_i,t-2}$. First, this back-door path can be blocked by a subset of $\{\overline{\mathbf{C}}, \overline{\mathbf{C}}^{(-1)}, \overline{\mathbf{C}}^{(+1)}, \mathbf{Y}_{\mathcal{N}_i,t-2}\}$. If this back-door is the only unblocked back-door, set $\overline{\mathbf{C}}$ is not proper, therefore this is contradictory. If there is another back-door path that both $\overline{\mathbf{C}}$ and $\overline{\mathbf{C}}^P$ cannot block, choose it as π. □

Lemma 3 Adding $\text{Des}(Y_{i,t-1})$ cannot block a back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$ unblocked by set $\overline{\mathbf{C}}^P$.

Proof Suppose controlling for $\text{Des}(Y_{i,t-1})$ can block a back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$ that the original set $\overline{\mathbf{C}}^P$ cannot block. Since $\overline{\mathbf{C}}^P$ does not include any $\text{Des}(Y_{k,t-1})$ or $\text{Des}(Y_{i,t-1})$, this unblocked back-door path contains an arrow pointing to $Y_{i,t-1}$.

Step 1 (Set up the main node B): At least one of $\text{Des}(Y_{i,t-1})$ is a non-collider on this path given that controlling for $\text{Des}(Y_{i,t-1})$ can block this path. Let B be such a variable and focus on one arrow pointing out from the node B.

Step 2 (Case I. Consider one side of the main node B): First, suppose this direction leads to $Y_{i,t-1}$. Then, since B is a $\text{Des}(Y_{i,t-1})$, a directed path from node B to $Y_{i,t-1}$ cannot exist and therefore, there must be a collider on this direction of the path. Since this collider is also in $\text{Des}(Y_{i,t-1})$ and therefore not controlled in the original $\overline{\mathbf{C}}^P$, this back-door is blocked by set $\overline{\mathbf{C}}^P$.

Step 3 (Case II. Consider the other side of the main node B): Next, consider the direction that leads to $Y_{k,t-1}$. Then, since $Y_{i,t-1}$ is not a cause of $Y_{k,t-1}$, a directed path from node B to $Y_{k,t-1}$ cannot exist and therefore, there must be a collider on this direction of the

path. Since this collider is also in $\mathrm{Des}(Y_{i,t-1})$ and therefore not controlled in the original $\overline{\mathbf{C}}^P$, this back-door is blocked by set $\overline{\mathbf{C}}^P$. Hence, this is contradiction. This proves that controlling for $\mathrm{Des}(Y_{i,t-1})$ cannot block a back-door path from $Y_{k,t-1}$ to $Y_{i,t-1}$ that set $\overline{\mathbf{C}}^P$ cannot block. □

A.3 Proof of Theorem 2

Below, we describe two lemmas useful for proving Theorem 2. For completeness, their proofs follow.

Lemma 4

$$Y_{it}(d^L) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid U_{it}, \overline{\mathbf{C}}_{it} \Longrightarrow Y_{it}(d^L) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid U_{it}, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}$$

Lemma 5 Under Assumption 3,

$$\begin{aligned} & \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ = \; & \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]. \end{aligned}$$

Proof of the theorem Based on Lemma 5 and Assumption 3,

$$\begin{aligned} & \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ = \; & \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ & + \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ = \; & \mathbb{E}[Y_{it} \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ & + \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]. \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{E}[Y_{it}(d^H) - Y_{it}(d^L) \mid D_{it} = d^H] \\ = \; & \int \{\mathbb{E}[Y_{it}(d^H) \mid D_{it} = d^H, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}] \\ & \quad - \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^H, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}]\} dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} \mid D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}) \\ = \; & \int \mathbb{E}[Y_{it} \mid D_{it} = d^H, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}] dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} \mid D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}) \\ & - \big\{\mathbb{E}[Y_{it} \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] + \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ & \quad - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]\big\} dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} \mid D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}) \end{aligned}$$

$$
\begin{aligned}
= & \int \{\mathbb{E}[Y_{it} \mid D_{it} = d^H, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}] - \mathbb{E}[Y_{it} \mid D_{it} = d^L, \overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it}]\} dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} | D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}) \\
& - \int \{\mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1}, \overline{\mathbf{C}}^B_{it}] - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1}, \overline{\mathbf{C}}^B_{it}]\} dF_{\overline{\mathbf{X}}_{it}, \overline{\mathbf{C}}^B_{it} | D_{it} = d^H}(\overline{\mathbf{x}}, \overline{\mathbf{c}}).
\end{aligned}
$$

This completes the proof of Theorem 2 in cases where U_{it} is time-dependent and affected by the outcome at time t. In Section A.3.3, we extend results to two other cases (1) when U_{it} is time-dependent but is not affected by the outcome at time t and (2) when unobserved confounder is time-independent Z_i. □

A.3.1 Proof of Lemma 4

If we write out control set $\overline{\mathbf{C}}$, the lemma can be rewritten as

$$
\begin{aligned}
& Y_{it}(d^L) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i, t-1} \mid U_{it}, \overline{\mathbf{X}}_{it}, \overline{\mathbf{X}^*}_{it}, \widetilde{\mathbf{X}}_i \\
\implies & Y_{it}(d^L) \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i, t-1} \mid U_{it}, \overline{\mathbf{X}}_{it}, \overline{\mathbf{X}^*}_{it}, \overline{\mathbf{X}^*}_{i,t-1}, \widetilde{\mathbf{X}}_i, \mathbf{Y}_{\mathcal{N}_i, t-2}.
\end{aligned}
$$

First, note that all variables in set $\{U_{it}, \overline{\mathbf{X}}_{it}, \overline{\mathbf{X}^*}_{it}, \overline{\mathbf{X}^*}_{i,t-1}, \widetilde{\mathbf{X}}_i, \mathbf{Y}_{\mathcal{N}_i, t-2}\}$ are neither affected by the potential outcome, $Y_{it}(d^L)$, nor affected by the treatment $\mathbf{Y}_{\mathcal{N}_i, t-1}$. The difference between the conditioning sets in the right- and left-hand sides is $\overline{\mathbf{X}^*}_{i,t-1}$ and $\mathbf{Y}_{\mathcal{N}_i, t-2}$. Including these variables can open back-door paths only when these variables are colliders for these new back-door paths. However, because a descendant of $\overline{\mathbf{X}^*}_{i,t-1}$, $\overline{\mathbf{X}^*}_{it}$, is in the conditioning set, it is contradictory if conditioning on $\overline{\mathbf{X}^*}_{i,t-1}$ can open a new back-door path. Additionally, because $\mathbf{Y}_{\mathcal{N}_i, t-2}$ is a parent of the treatment $\mathbf{Y}_{\mathcal{N}_i, t-1}$, it is contradictory if conditioning on $\mathbf{Y}_{\mathcal{N}_i, t-2}$ can open a new back-door path. Therefore, including $\overline{\mathbf{X}^*}_{i,t-1}$ and $\mathbf{Y}_{\mathcal{N}_i, t-2}$ don't open any back-door path, which completes the proof. □

A.3.2 Proof of Lemma 5

Under Assumption 3,

$$
\begin{aligned}
& \int_{\mathcal{C}} \{\mathbb{E}[Y_{it}(d^L) | U_{it} = u_1, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) | U_{it} = u_0, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]\} \\
& \qquad \times \{dF_{U_{it} | D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}}(u_1) - dF_{U_{it} | D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}}(u_1)\} \\
= & \int_{\mathcal{C}} \{\mathbb{E}[Y_{i,t-1} | U_{i,t-1} = u_1, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} | U_{i,t-1} = u_0, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]\} \\
& \qquad \times \{dF_{U_{i,t-1} | D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}}(u_1) - dF_{U_{i,t-1} | D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}}(u_1)\}.
\end{aligned}
$$

Now we analyze each side of the equation.

$$
\int_{\mathcal{C}} \{\mathbb{E}[Y_{it}(d^L) | U_{it} = u_1, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) | U_{it} = u_0, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]\}
$$

$$
\begin{aligned}
&\times\{dF_{U_{it}|D_{it}=d^H,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1) - dF_{U_{it}|D_{it}=d^L,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1)\} \\
= &\int_{\mathcal{C}} \mathbb{E}[Y_{it}(d^L)|U_{it}=u_1,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] \\
&\times\{dF_{U_{it}|D_{it}=d^H,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1) - dF_{U_{it}|D_{it}=d^L,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1)\} \\
= &\int_{\mathcal{C}} \mathbb{E}[Y_{it}(d^L)|D_{it}=d^H,U_{it}=u_1,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}]dF_{U_{it}|D_{it}=d^H,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1) \\
&- \int_{\mathcal{C}} \mathbb{E}[Y_{it}(d^L)|D_{it}=d^L,U_{it}=u_1,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}]dF_{U_{it}|D_{it}=d^L,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1) \\
= &\mathbb{E}[Y_{it}(d^L)|D_{it}=d^H,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L)|D_{it}=d^L,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}],
\end{aligned}
$$

where the first equality follows from the fact that $\mathbb{E}[Y_{it}(d^L)|U_{it}=u_0,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}]$ does not include u_1, the second equality comes from Lemma 4, and the final from the rule of conditional expectations. Similarly,

$$
\begin{aligned}
&\int_{\mathcal{C}}\{\mathbb{E}[Y_{i,t-1}|U_{i,t-1}=u_1,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1}|U_{i,t-1}=u_0,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}]\} \\
&\times\{dF_{U_{i,t-1}|D_{it}=d^H,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1) - dF_{U_{i,t-1}|D_{it}=d^L,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}}(u_1)\} \\
= &\mathbb{E}[Y_{i,t-1} \mid D_{it}=d^H,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it}=d^L,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}].
\end{aligned}
$$

Taken together,

$$
\begin{aligned}
&\mathbb{E}[Y_{it}(d^L) \mid D_{it}=d^H,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) \mid D_{it}=d^L,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] \\
= &\mathbb{E}[Y_{i,t-1} \mid D_{it}=d^H,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it}=d^L,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}].
\end{aligned}
$$

□

A.3.3 Other cases

In Theorem 2, we consider cases in which U_{it} is time-dependent and affected by the outcome at time t. Now we study two other cases (1) when U_{it} is time-dependent but is not affected by the outcome at time t and (2) when unobserved confounder is time-independent Z_i. For both cases, Assumption 3 needs to be modified accordingly, although their substantive meanings stay the same. The definition of the bias-corrected estimator is also the same. For case (1), define $\widetilde{U}_i \equiv (U_{it}, U_{i,t-1})$ and for case (2), define $\widetilde{U}_i \equiv Z_i$. Then, Assumption 3 is modified as follows.

1. Time-invariant effect of unobserved confounder $\widetilde{U}$: For all $u_1, u_0, \overline{\mathbf{x}}$ and $\overline{\mathbf{c}}$,

$$
\begin{aligned}
&\mathbb{E}[Y_{it}(d^L) \mid \widetilde{U}_i=u_1,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) \mid \widetilde{U}_i=u_0,\overline{\mathbf{X}}_{it}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] \\
= &\mathbb{E}[Y_{i,t-1} \mid \widetilde{U}_i=u_1,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid \widetilde{U}_i=u_0,\overline{\mathbf{X}}_{i,t-1}=\overline{\mathbf{x}},\overline{\mathbf{C}}^B_{it}=\overline{\mathbf{c}}].
\end{aligned}
$$

2. Time-invariant imbalance of unobserved confounder $\widetilde{U}$: For all $u, \overline{\mathbf{x}}$ and $\overline{\mathbf{c}}$,

$$\begin{aligned} & \Pr(\widetilde{U}_i \leq u \mid D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}) - \Pr(\widetilde{U}_i \leq u \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}) \\ = \; & \Pr(\widetilde{U}_i \leq u \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}) - \Pr(\widetilde{U}_i \leq u \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}). \end{aligned}$$

A.4 Extensions

A.4.1 Sensitivity Analysis

As Lemma 5 shows, Assumption 3 is equivalent to the following equality.

$$\begin{aligned} & \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] \\ = \; & \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}], \end{aligned}$$

which substantively means the time-invariant bias. However, this assumption might hold only approximately in applied settings. To assess the robustness of the bias-corrected estimates, we consider a sensitivity analysis. In particular, we introduce sensitivity parameter λ as follows.

$$\frac{\mathrm{B}_t(\overline{\mathbf{x}}, \overline{\mathbf{c}})}{\mathrm{B}_{t-1}(\overline{\mathbf{x}}, \overline{\mathbf{c}})} = \lambda$$

where

$$\begin{aligned} \mathrm{B}_t(\overline{\mathbf{x}}, \overline{\mathbf{c}}) &= \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^H, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it}(d^L) \mid D_{it} = d^L, \overline{\mathbf{X}}_{it} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}], \\ \mathrm{B}_{t-1}(\overline{\mathbf{x}}, \overline{\mathbf{c}}) &= \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^H, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{i,t-1} \mid D_{it} = d^L, \overline{\mathbf{X}}_{i,t-1} = \overline{\mathbf{x}}, \overline{\mathbf{C}}^B_{it} = \overline{\mathbf{c}}]. \end{aligned}$$

The time-invariance assumption (Assumption 3) corresponds to $\lambda = 1$. Using this sensitivity parameter, we can re-define the bias-corrected estimator as follows.

$$\hat{\tau}_{\text{Main}} - \lambda \times \hat{\delta}_{\text{Placebo}}$$

Therefore, a sensitivity analysis is to compute the bias-corrected estimator for a range of plausible values of λ and investigate whether substantive conclusions vary according to the choice of the sensitivity parameter.

B Causal Directed Acyclic Graphs: Review

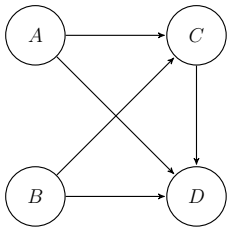
In the paper, we use a causal directed acyclic graph and nonparametric structural equations to represent causal relationships. Here, we review basic definitions and results. See Pearl (2000)

for a comprehensive review. Following Pearl (1995), we define a causal directed acyclic graph (causal DAG) to be a set of nodes and directed edges among nodes such that the graph has no cycles and each node corresponds to a univariate random variable. Each random variable is given by its nonparametric structural equation. When there is a directed edge from one variable to another variable, the latter variable is a function of the former variable. For example, in a causal DAG in Figure A1 (a), four random variables (A, B, C, D) are given by nonparametric structural equations in Figure A1 (b); $A = f_A(\epsilon_A), B = f_B(\epsilon_B), C = f_C(A, B, \epsilon_C)$, and $D = f_D(A, B, C, \epsilon_D)$, where f_A, f_B, f_C and f_D are unknown nonparametric structural equations and $(\epsilon_A, \epsilon_B, \epsilon_C, \epsilon_D)$ are mutually independent errors. The node that a directed edge starts from is called the *parent* of the node that the edge goes into. The node that the edge goes into is the *child* of the node it comes from. If two nodes are connected by a directed path, the first node is the *ancestor* of every node on the path, and every node on the path is the *descendant* of the first node (Pearl, 2000). For example, node A is a parent of node C, and nodes C and D are descendants of node B. The requirement that the errors be mutually independent essentially means that there is no variable absent from the graph which, if included on the graph, would be a parent of two or more variables.

The nonparametric structural equations are general – random variables may depend on any function of their parents and variable-specific errors. They encode counterfactual relationships between the variables on the graph by recursively representing one-step-ahead counterfactuals. Under a hypothetical intervention setting A to a, the distribution of the variables B, C, and D are then recursively given by the nonparametric structural equations with $A = f_A(\epsilon_A)$ replaced by $A = a$. Specifically, $B = f_B(\epsilon_B),\ C = C(a) = f_C(A = a, B, \epsilon_C)$, and $D = D(a) = f_D(A = a, B, C = C(a), \epsilon_D)$ where $C(a), D(a)$ are the counterfactual values of C and D when A is set to a.

C Example of Structural Stationarity

Structural stationarity is satisified in a more general NPSEM than the example in the main text. First, variables can be affected not only by one-time lag but also by longer-time lags. For example, outcome Y_{it} can be affected not only by the neighbors' outcomes at the last period $\mathbf{Y}_{\mathcal{N}_i,t-1}$ but also by the neighbors' outcomes at two periods before $\mathbf{Y}_{\mathcal{N}_i,t-2}$. Second,



(a) A causal directed acyclic graph

$$
\begin{aligned}
A &= f_A(\epsilon_A) \\
B &= f_B(\epsilon_B) \\
C &= f_C(A, B, \epsilon_C) \\
D &= f_D(A, B, C, \epsilon_D)
\end{aligned}
$$

(b) A structural equation model

Figure A1: An Example of Causal DAGs and SEMs

each variable can be not only affected by other variables within each unit but also by other variables of neighbors. For example, outcome Y_{it} can be affected by $\mathbf{L}_{\mathcal{N}_i,t-1}$ and $\mathbf{U}_{\mathcal{N}_i,t-1}$.

We now consider an example that incorporates more complex feedback between variables across time and neighbors. For $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, suppose the data are generated by sequentially evaluating the following set of equations:

$$
\begin{aligned}
&\text{(Outcome variable)} \\
&\quad Y_{it} = f_Y(\mathbf{Y}_{\mathcal{N}_i,t-1}, \mathbf{Y}_{\mathcal{N}_i,t-2}, Y_{i,t-1}, \mathbf{L}_{it}, \mathbf{L}_{\mathcal{N}_i,t-1}, \widetilde{\mathbf{L}}_i, \mathbf{U}_{it}, \mathbf{U}_{\mathcal{N}_i,t-1}, \epsilon_{it}^Y), \\
&\text{(Time-varying Observed variables)} \\
&\quad \mathbf{L}_{it} = f_L(\mathbf{L}_{i,t-1}, \mathbf{L}_{\mathcal{N}_i,t-1}, \widetilde{\mathbf{L}}_i, Y_{i,t-1}, \mathbf{Y}_{\mathcal{N}_i,t-2}, \mathbf{U}_{i,t-1}, \mathbf{U}_{\mathcal{N}_i,t-2}, \boldsymbol{\epsilon}_{it}^L), \\
&\text{(Time-invariant Observed variables)} \\
&\quad \widetilde{\mathbf{L}}_i = f_{\widetilde{L}}(\mathbf{L}_{i,0}, \mathbf{L}_{\mathcal{N}_i,0}, Y_{i,0}, \mathbf{Y}_{\mathcal{N}_i,0}, \mathbf{U}_{i,0}, \mathbf{U}_{\mathcal{N}_i,0}, \boldsymbol{\epsilon}_i^{\widetilde{L}}), \\
&\text{(Time-varying Unobserved variables)} \\
&\quad \mathbf{U}_{it} = f_U(\mathbf{U}_{i,t-1}, \mathbf{U}_{\mathcal{N}_i,t-1}, Y_{i,t-1}, \mathbf{Y}_{\mathcal{N}_i,t-2}, \mathbf{L}_{i,t-1}, \mathbf{L}_{\mathcal{N}_i,t-2}, \widetilde{\mathbf{L}}_i, \boldsymbol{\epsilon}_{it}^U).
\end{aligned}
\tag{A1}
$$

Several points are worth noting. First, variables can be affected not only by one-time lag but also by longer-time lags. For example, outcome Y_{it} is affected not only by the neighbors' outcomes at the last period $\mathbf{Y}_{\mathcal{N}_i,t-1}$ but also by the neighbors' outcomes at two periods before $\mathbf{Y}_{\mathcal{N}_i,t-2}$. While we do not restrict the number of time-lags and allow for higher-order temporal dependence, we keep our focus on the ACDE defined in equation (1) as our causal estimand. Second, each variable is not only affected by other variables within each unit but also by other variables of neighbors. For example, outcome Y_{it} is affected by $\mathbf{L}_{\mathcal{N}_i,t-1}$ and $\mathbf{U}_{\mathcal{N}_i,t-1}$. Time-vaying unmeasured variables $\mathbf{U}_{it}$ is affected by $\mathbf{U}_{\mathcal{N}_i,t-1}$, $\mathbf{Y}_{\mathcal{N}_i,t-2}$, and $\mathbf{L}_{\mathcal{N}_i,t-2}$. Even though the

complexity of the NPSEMs are different in equations (5) and (A1), they both satisfy structural stationarity.

D Simulation Study

In this section, we consider the performance of the proposed placebo test and bias-corrected estimator in a simulation study calibrated to the real hate crime data. In Section D.1, we show that (1) a placebo estimator is consistent for zero under the no omitted confounders assumption as Theorem 1 implies and (2) the statistical power of the proposed placebo test is comparable to an "oracle" test — test whether an estimated ACDE is statistically distinguishable from the true ACDE, which is available only in simulations. In Section D.2, we demonstrate that the bias-corrected estimator reduces bias and root mean squared error (RMSE) even under a slight violation of the time-invariance assumption (Assumption 3).

Setup. To approximate realistic data generating processes, we use the same hate crime data as in the main application but focus on another important outcome, the number of attacks against refugee housing, which is also an important aspect of hate crimes studied in the literature. As for observed covariates, we include five major contextual variables; the number of refugees, the number of crimes per 100,000 inhabitants, per capita income, the unemployment rate, and the share of school leavers without lower secondary education graduation. We fit a linear regression with these five covariates, as in equation (9), to estimate the basic parameters of the data generating process.

We simulate a distance matrix $\mathbf{W}$ based on the stochastic block model (Holland *et al.*, 1983) for each of the sample size $n \in \{100, 500, 1000, 2000\}$. Each group consists of ten units and there exist $K = n/10$ groups. K groups are divided into $L = K/5$ blocks. If units i and j are within the same group, $\Pr(W_{ij} = 1) = 0.8$. If units i and j are within the same block but not in the same group, $\Pr(W_{ij} = 1) = 0.2$. If units i and j are in different blocks, $\Pr(W_{ij} = 1) = 0$. This setup is designed to ensure that the network dependency does not keep growing as the sample size grows. See Sävje *et al.* (2017) and Ogburn *et al.* (2017) for general discussions on network asymptotics.

We then simulate an unobserved contextual variable U_{it}. In particular, we consider two scenarios; (1) time-invariant confounding where assumptions for both the placebo test and the

bias-corrected estimator hold, and (2) structural stationarity where assumptions hold for the placebo test but the time-invariance assumption required for the bias-correction is violated. For the first scenario, we set unobserved contextual variable U to be time-invariant where $U_i = \widetilde{U}_{k[i]}$ where $\tilde{U}_k \sim \mathcal{N}(0, 0.5)$ and $k[i]$ is a group indicator for unit i. For the second scenario, we draw unobserved contextual variable U as follows. $U_{it} = \widetilde{U}_{k[i],t}$ where $U_{k,t} = 0.9U_{k,t-1} + \mathcal{N}(0, 0.1)$ where $U_{k0} \sim \mathcal{N}(0, 0.5)$.

Given this setup, we sample potential outcomes using the following data generating process.

$$Y_{i,t+1}(D_{it}) = \alpha + \tau D_{it} + \overline{\mathbf{X}}_{i,t+1}^{\top}\beta + \gamma U_{i,t+1} + \epsilon_{i,t+1}, \tag{A2}$$

for sample size in each time period $n \in \{100, 500, 1000, 2000\}$ and the total number of time periods $T = 20$. $D_{it} \equiv \mathbf{W}_i^{\top}\mathbf{Y}_t$ indicates the treatment variable, five-dimensional vector $\overline{\mathbf{X}}_{i,t+1}$ represents five observed covariates from the real hate crime data, $U_{i,t+1}$ is the unobserved contextual confounder affecting multiple units, and the error term $\epsilon_{i,t+1}$ follows the normal distribution, $\epsilon_{i,t+1} \sim \mathcal{N}(0, 0.1)$. Coefficients $\{\alpha = 0.59, \tau = 0.74, \beta = (0.75, -0.11, -0.28, -3.38, 3.90)\}$ are based on estimated parameters from the real hate crime data. The effect of unobserved contextual confounder U is set to $\gamma = 0.1$. Based on this data generating process, we conduct 5000 independent Monte Carlo simulations.

D.1 Placebo Test

First, we consider the consistency of the proposed placebo test under the no omitted confounders assumption. Theorem 1 implies that when the no omitted confounders assumption holds, the treatment variable and the lagged dependent variable are conditionally independent. In particular, we fit a placebo regression:

$$Y_{it} = \alpha_0 + \delta D_{it} + \tau_0 D_{i,t-1} + \overline{\mathbf{X}}_{it}^{\top}\beta_0 + \gamma_0 U_{it} + \epsilon_{it}. \tag{A3}$$

We expect that a test statistic $\widehat{\delta}$ is consistent for zero under the no omitted confounders assumption. The first row in Figure A2 presents the results. As Theorem 1 shows, under the no omitted confounders assumption, the placebo estimator $\widehat{\delta}$ converges to zero as the sample size grows. Because Theorem 1 only requires structural stationarity, the placebo test is consistent under both scenarios.

We also investigate the statistical power of the proposed placebo test when the no omitted

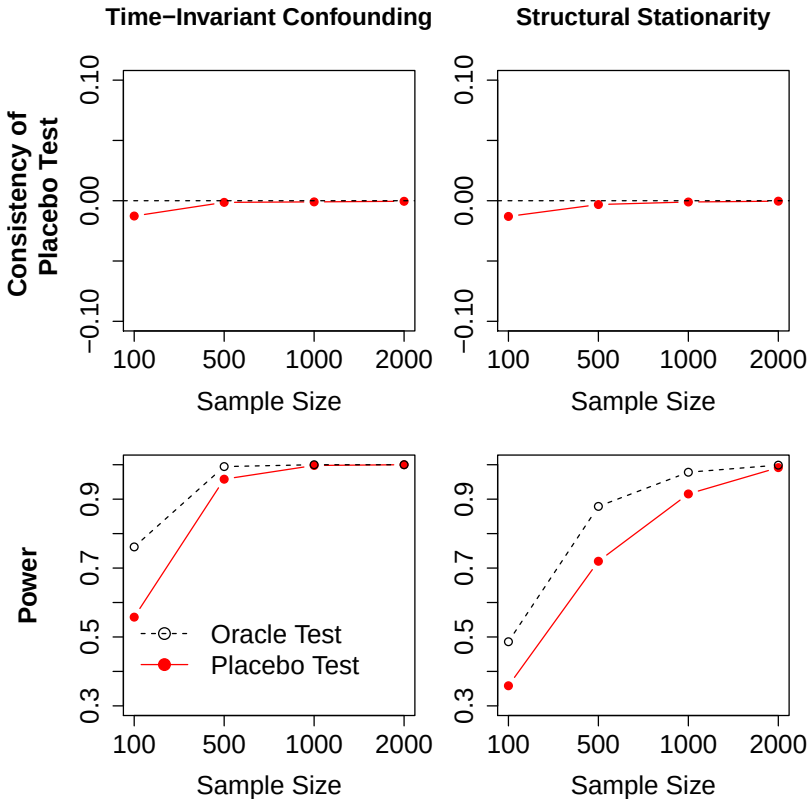


Figure A2: Simulation Results on Placebo Test. *Note:* The first row considers the consistency of the placebo test under the no omitted confounders assumption. The second row compares the statistical power of the proposed placebo test (solid red line) and the oracle test (dotted black line). The first and second columns correspond to the time-invariant confounding and structural stationarity, respectively. Results are based on 5000 Monte Carlo draws using four sample sizes.

confounders assumption is violated. We fit a placebo regression:

$$Y_{it} = \tilde{\alpha}_0 + \tilde{\delta} D_{it} + \tilde{\tau}_0 D_{i,t-1} + \overline{\mathbf{X}}_{it}^{\top} \tilde{\beta}_0 + \tilde{\epsilon}_{it}. \tag{A4}$$

The key difference is that this regression now ignores contextual confounder U_{it}. Here, $\widehat{\tilde{\delta}}$ serves as a test statistic for the placebo test. We compare this to an oracle test where we fit the following main linear regression,

$$Y_{i,t+1} = \alpha_m + \tau_m D_{it} + \overline{\mathbf{X}}_{i,t+1}^{\top} \beta_m + \xi_{i,t+1}, \tag{A5}$$

and test $H_0 : \tau_m = \tau$. This test is an "oracle" test because it is available only in the simulation where we know the true ACDE τ. The second row in Figure A2 presents the results. Even when the sample size is small, the proposed placebo test achieves more than 70% of the oracle

test's power. As the sample size grows, the proposed placebo test attains the statistical power as high as that of the oracle test. Given that the oracle test is available only in simulations where the true ACDE is known, these results suggest that the placebo test can serve as a powerful practical tool to detect biases in applied settings.

D.2 Bias-Corrected Estimator

In Section 4.3, we show that the proposed bias-corrected estimator can identify the ACDE for the treated under Assumption 3. Here, we investigate how much the bias-corrected estimator can reduce bias and RMSE even in settings where this required time-invariance assumption is slightly violated.

In particular, we compare an uncorrected estimator, which ignores unobserved contextual confounder U, and the proposed bias-corrected estimator under two scenarios; (1) time-invariant confounding and (2) structural stationarity. The time-invariance assumption required for the bias correction (Assumption 3) holds in the first but not in the second scenario.

Figure A3 presents the simulation results. In the time-invariant confounding case (the first column), whereas the bias in the conventional uncorrected estimator is about 0.12, the bias in the proposed bias-corrected estimator is essentially 0. The bias is corrected as Theorem 2 implies. The RMSE also significantly improves upon the uncorrected conventional estimator. The 95% confidence interval is close to its nominal coverage rate in contrast to that of the uncorrected estimator.

More importantly, even in structural stationarity case (the second column in Figure A3) where the required assumption for the bias correction is slightly violated, the bias-corrected estimator shows reasonable performance. While the bias in the conventional uncorrected estimator is about 0.04, the bias in the proposed bias-corrected estimator is less than 0.01. Although the bias does not vanish, it reduces by about 80%. This benefit is also clear in the results of RMSE. Because the bias-corrected estimator tends to have a larger standard error, the RMSE of the bias-corrected estimator is bigger than the one of the uncorrected estimator when the sample size is small. However, as the sample size grows, the bias-corrected estimator outperforms the uncorrected estimator. Finally, as the required time-invariance assumption is violated, the coverage of the 95% confidence interval for the bias-corrected estimator is slightly smaller than its nominal coverage rate, but it attains more than 90% in contrast to the performance of the uncorrected estimator. These results suggest that the proposed bias-corrected estimator can reduce bias and RMSE in applied settings where the necessary assumption might hold only approximately.

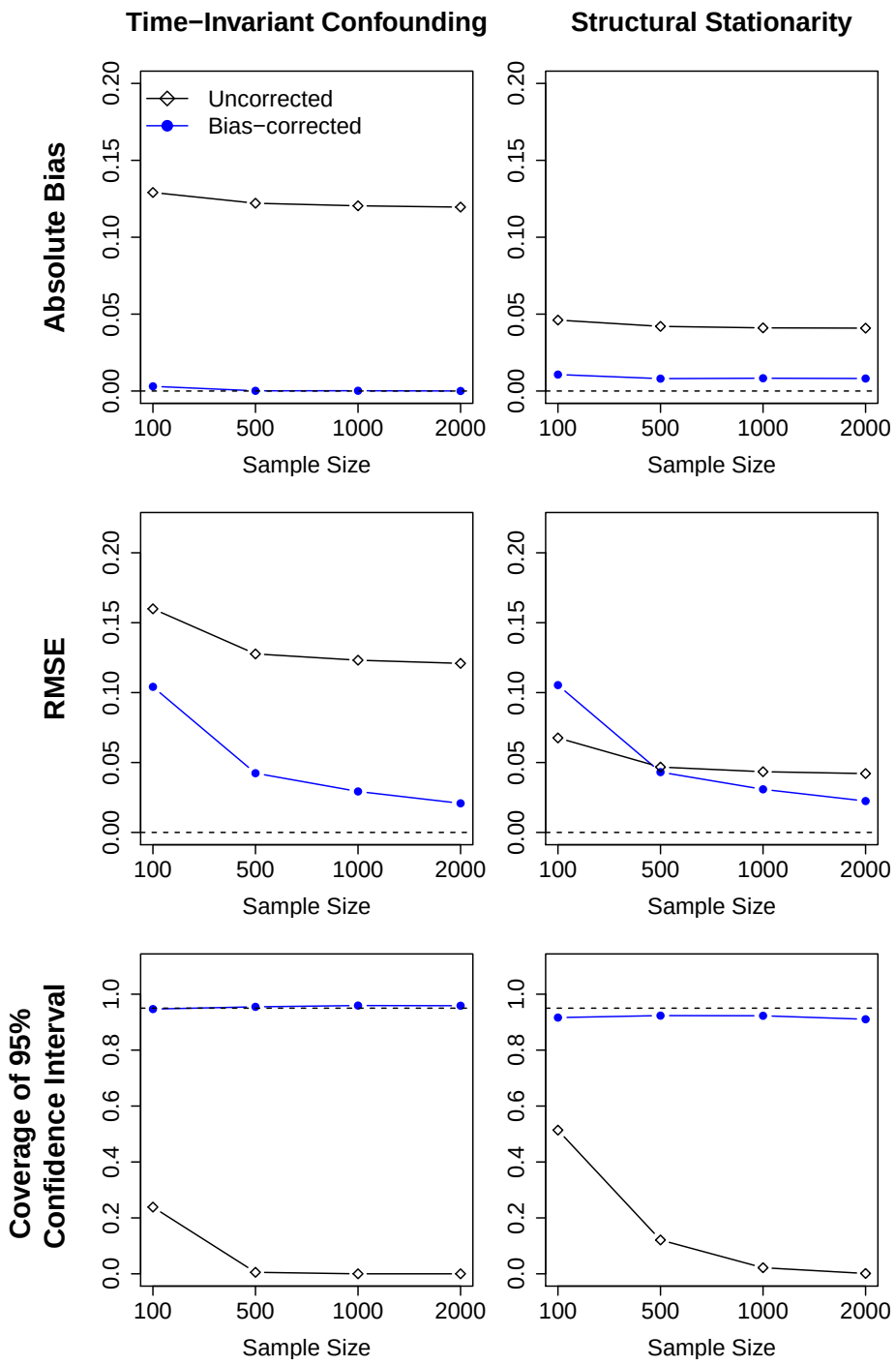


Figure A3: Simulation Results on Bias-Corrected Estimator. *Note:* The first row compares the absolute bias of the uncorrected estimator (empty black square) and the bias-corrected estimator (solid blue circle). The second row examines the root mean squared error (RMSE) and the third row shows the coverage of the 95% confidence interval. The first and second columns correspond to the time-invariant confounding and structural stationarity, respectively. Results are based on 5000 Monte Carlo draws using four sample sizes.

E Empirical Analysis in Section 5

E.1 Control Sets and Placebo Sets

We investigate five different control sets to illustrate how to use the proposed placebo test and bias-corrected estimator. Table A1 describes types of variables we use for those five control sets and their corresponding placebo sets. The column of "Main model" indicates variables used for control sets and the column of "Placebo model" indicates corresponding variables in placebo sets.

The first control set (C1) includes variables from "Basic Variables." The second control set (C2) adds variables from "Two-month Lags" to the first control set. The third control set adds state fixed effects to the second control set. The fourth control set adds all the variables from "Contextual Variables," which include variables on refugees, demographics, general crimes, economic indicators, education, and politics. Note that these contextual variables are measured only annually. The final fifth set adds the time trend variable as third-order polynomials to the fourth set.

Type	**Main Model**	**Placebo Model**
Outcome	Physical Attack$_{t+1}$	Physical Attack$_t$
Treatment	Physical Attack$_t$ in Neighbors	Physical Attack$_t$ in Neighbors
A Control Set/A Placebo Set		
Basic Variables	Physical Attack$_t$	Physical Attack$_{t-1}$
	Physical Attack$_{t-1}$ in Neighbors	Physical Attack$_{t-1,t-2}$ in Neighbors
	the number of neighbors	the number of neighbors
	variance of $\mathbf{W}_i$	variance of $\mathbf{W}_i$
Two-month Lags	Physical Attack$_{t-1}$	Physical Attack$_{t-2}$
Contextual Variables (annual)		
Refugee variables	Total number of refugees	Total number of refugees
	Total number of foreign born	Total number of foreign born
Population variables	Population size	Population size
	Share of male inhabitants	Share of male inhabitants
Crime variables	Number of general crimes per 100,000 inhabitants	Number of general crimes per 100,000 inhabitants
	Percent of general crimes solved	Percept of general crimes solved
Economic variables	Number of newly registered business	Number of newly registered business
	Number of newly deregistered business	Number of newly deregistered business
	Number of insolvency	Number of insolvency
	per capita income	per capita income
	Number of employees with social security	Number of employees with social security
	Unemployment rate	Unemployment rate
Education variables	Share of school leavers without lower secondary education graduation	Share of school leavers without lower secondary education graduation
Political variables	Turnout rate in 2013	Turnout rate in 2013
	Vote share of extreme right and populist right-wing parties in 2013	Vote share of extreme right and populist right-wing parties in 2013

Table A1: Five Control Sets and Placebo Sets: Spatial Diffusion of Hate Crimes.

E.2 Conditional ACDEs by Education

We present the distribution of proportions of school dropouts without a secondary school diploma, separately for East Germany and West Germany. Because these distributions are substantially different between them (Figure A4), we estimate the conditional ACDE by proportions of school dropouts, separately for the East and the West.

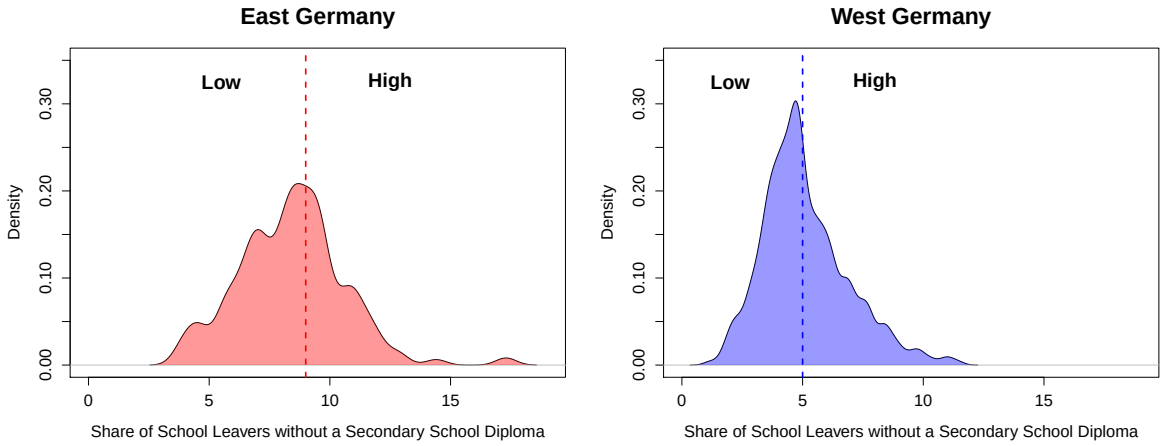
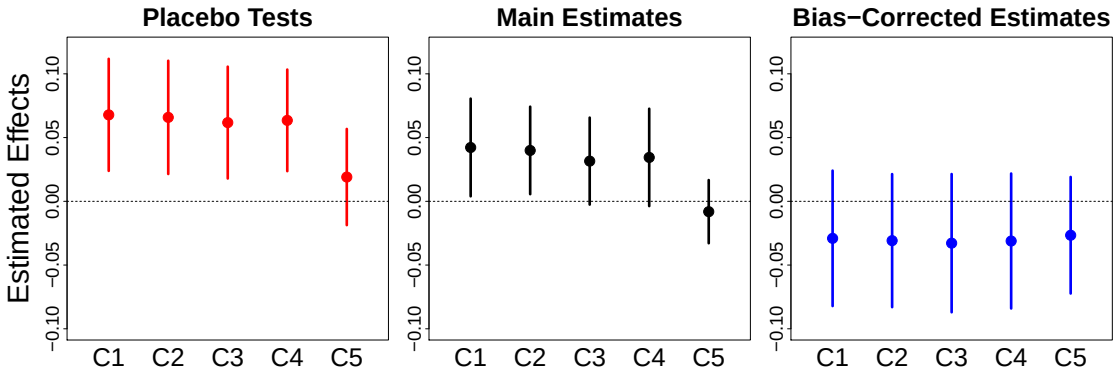


Figure A4: Distribution of Proportions of School Dropouts. Note: For East Germany, we use 9% as a cutoff for high and low proportions of school dropouts, which is approximately the median value in East Germany. For West Germany, we use 5% as a cutoff for high and low proportions of school dropouts, which is approximately the median value in West Germany.

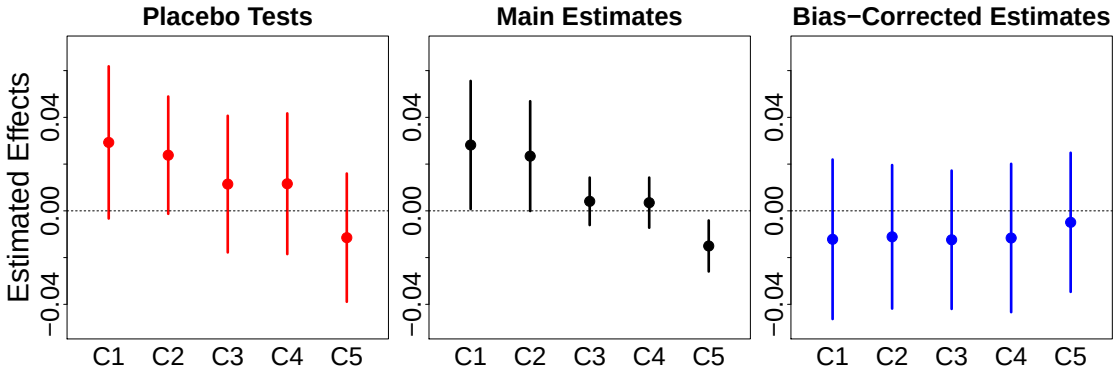
Next, we present the conditional ACDE for counties in East Germany with low proportions of school dropouts. In contrast to Figure 5, estimates are small.



(a) (b) (c)

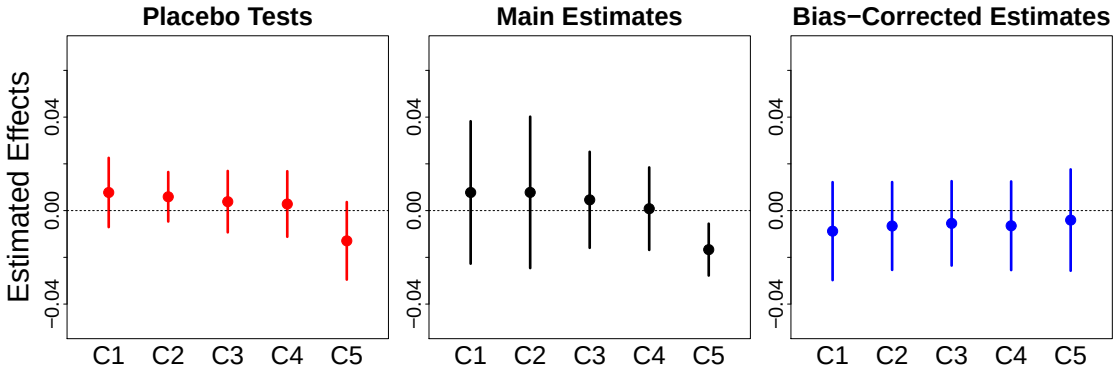
Figure A5: Results of the conditional ACDE (Low Proportion of School Dropouts, East). Note: Figure (a) shows that the last fifth set produces the smallest placebo estimate. Focusing on this fifth control set, a point estimate of the ACDE in Figure (b) is close to zero and its 95% confidence interval covers zero. Figure (c) shows that bias-corrected estimates are similar regardless of the selection of control variables and all of their 95% confidence intervals cover zero.

Now, we present the conditional ACDEs for counties in West Germany with high and low proportions of school dropouts. Given that proportions of school dropouts are lower in West Germany, estimates of the conditional ACDEs are small, in contrast to Figure 5.



(a) (b) (c)

Figure A6: Results of the conditional ACDE (High Proportion of School Dropouts, West). Note: Figure (a) shows that the third, fourth and fifth sets produce small placebo estimates. Focusing on these sets, point estimates of the ACDE in Figure (b) are close to zero and sometimes negative. Figure (c) shows that bias-corrected estimates are similar regardless of the selection of control variables and all of their 95% confidence intervals cover zero.



(a) (b) (c)

Figure A7: Results of the conditional ACDE (Low Proportion of School Dropouts, West). Note: Figure (a) shows that all the sets produce small placebo estimates. This is partly because there are few hate crimes in this area and hence, there is no variation in outcomes and treatments. In addition, point estimates of the ACDE in Figure (b) are close to zero and sometimes negative. Figure (c) shows that bias-corrected estimates are similar regardless of the selection of control variables and all of their 95% confidence intervals cover zero.

References

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic Blockmodels: First Steps. *Social Networks*, **5**(2), 109–137.

Ogburn, E. L., Sofrygin, O., Diaz, I., and van der Laan, M. J. (2017). Causal Inference for Social Network Data. *arXiv preprint arXiv:1705.08527*.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, **82**(4), 669–688.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge.

Sävje, F., Aronow, P. M., and Hudgens, M. G. (2017). Average Treatment Effects in the Presence of Unknown Interference. *arXiv preprint arXiv:1711.06399*.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2012). On the Validity of Covariate Adjustment for Estimating Causal Effects. In *Proceedings of the 26th Conference on Uncertainty and Artificial Intelligence*, pages 527–536, Corvallis, WA. AUAI Press.