Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **00**, 1–23
https://doi.org/10.1093/jrsssa/qnae014

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Original Article

Identification of causal diffusion effects using placebo outcomes under structural stationarity

Naoki Egami

Department of Political Science, Columbia University, New York, USA

Address for correspondence: Naoki Egami, Department of Political Science, Columbia University, 420 West 118th Street, New York, NY 10027, USA. Email: naoki.egami@columbia.edu

Abstract

Social and biomedical scientists have long been interested in the process through which ideas and behaviours 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 2 types mentioned above. We then propose a difference-in-differences style estimator that can directly correct biases under an additional causal assumption. Analysing 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

1 Introduction

Scientists have long been interested in how ideas and behaviours diffuse across space, networks, and time. For example, social scientists have studied the diffusion of policies and voting behaviours in political science (Graham et al., 2013; Jones et al., 2017; Sinclair, 2012), educational outcomes and crimes in economics (Duflo et al., 2011; Glaeser et al., 1996; Sacerdote, 2001), and innovations and job attainment in sociology (Granovetter, 1973; Rogers, 1962). Epidemiologists and researchers in public health have focused on the spread of infectious disease (Cai et al., 2019; Halloran & Struchiner, 1995; Morozova et al., 2018) and health behaviour (Christakis & 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

Received: January 22, 2022. **Revised:** August 20, 2023. **Accepted:** December 30, 2023

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 (Egami & Tchetgen Tchetgen, 2024; Manski, 1993; VanderWeele & 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. Angrist, 2014; Cohen-Cole & Fletcher, 2008; Lyons, 2011). 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 & 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; Ogburn & VanderWeele, 2014; Pearl, 2000). In particular, by extending a class of dynamic causal DAGs (Dean & Kanazawa, 1989; Pearl & 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 & 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 & 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, also known as causal peer effects (Goldsmith-Pinkham & Imbens, 2013; Ogburn, 2018; Shalizi & Thomas, 2011).[1] In particular, several papers develop methods specifically for network data. Some studies (e.g. An, 2015; Bramoullé et al., 2009; O'Malley et al., 2014) 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 analysing 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 (Miao & Tchetgen Tchetgen, 2017; Sofer et al., 2016) to identification of causal diffusion effects. Our work is different from and complementary to two recent papers utilizing negative controls. Egami and Tchetgen Tchetgen (2024) 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 in the dyadic data setting. First, unlike these two papers, our paper relies on a placebo outcome (a.k.a., negative control outcome), and thus, the placebo test, the bias-corrected estimator, and corresponding assumptions are different. Second, while both papers focus on the two-period network data, our method can handle panel data with both network and spatial settings and we analyse 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. Because the other two papers focus on the network data, they cannot be directly applied to the spatial data setting. We also offer comparisons of the underlying assumptions in Section 4.3.3 after we introduce our proposed bias-corrected estimator. Overall, these methods are complementary to each other and applicable to different application settings.

Finally, our approach based on causal DAGs and corresponding NPSEM is different and complementary to an alternative approach based on chain graphs. Recent papers (Ogburn et al., 2020; Tchetgen Tchetgen, Fulcher, et al., 2021) 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 (Myers, 2000; Wilson & Kelling, 1982). 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 behaviours 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

[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 & VanderWeele, 2014). See Halloran and Hudgens (2016) for a review of the interference literature.

weekly magazine *Stern*, which has been documenting anti-refugee violence in Germany since the beginning of 2014. The dataset we analyse in this paper is compiled by Dancygier et al. (2022), who extended these 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 1a 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 1b 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 (Braun, 2011; Myers, 2000). Indeed, Jäckle and König (2016) found that the incidence of hate crimes in one county predicts that of hate crimes 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 & Thomas, 2011; VanderWeele et al., 2012). This section introduces a setup for analysing such causal diffusion. We define the average causal diffusion effect (ACDE) 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.[2] Define *neighbours* $\mathcal{N}_i$ 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. Gleditsch and Ward (2006) use 500 km and Miguel and Kremer (2004) use 3 km and 6 km as the thresholds. We denote the outcome variables at time t of unit i's neighbours 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 neighbours over time, that is, the causal effect of neighbours' 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-

[2] Following the standard practice, this paper assumes that the distance matrix $\mathbf{W}$ is correctly specified. However, in practice, $\mathbf{W}$ might be misspecified. For example, the real underlying spatial matrix might be based on the travel distance even though researchers use the straight distance to define $\mathbf{W}$. Only recently this misspecification problem has received some methodological attention in the causal inference literature. Even in randomized experiments settings where there is no unobserved confounding, only few papers consider the misspecification (e.g. Aronow & Samii, 2017; Egami, 2021; Sävje, 2021), and there is no general approach to handle arbitrary misspecification. As far as we know, there is no prior work that handles misspecification in the observational spatial data settings even in the absence of unmeasured confounding. Developing methods to handle misspecification of $\mathbf{W}$ in this setting is an interesting topic for future work.

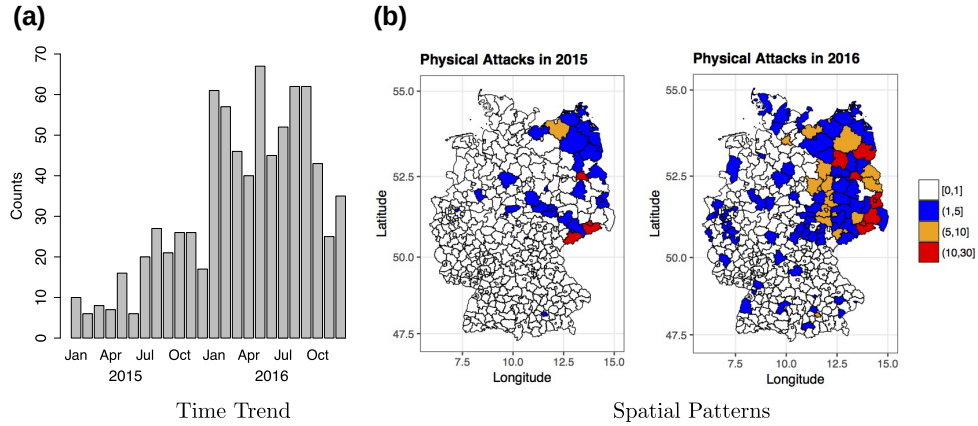


Figure 1. Temporal and spatial patterns of hate crimes in Germany. Note: The left figure shows the number of physical attacks each month. In the middle and right figures, we show the number of physical attacks in each county in 2015 and 2016, respectively. Each of 402 counties is coloured in white, blue, 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. (a) Time trend and (b) spatial patterns.

reducing assumption, known as the exposure mapping (Aronow & Samii, 2017), to define the treatment variable. In particular, we define the treatment variable D_{it} at time t as a function of relevant neighbours' 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 neighbours' 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; Robins, 1986; Rubin, 1974). $Y_{it}(d)$ 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 *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, 2018; Ogburn & VanderWeele, 2014). 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 (ACDE). The 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 this month would have changed if we had seen more hate crimes in neighbouring counties in the previous 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 neighbours' 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:

$$\tau_t(d^H, d^L) = \int_{\mathcal{C}} \left\{ \mathbb{E}[Y_{it} \mid D_{it} = d^H, \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}] - \mathbb{E}[Y_{it} \mid D_{it} = d^L, \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}] \right\} \mathrm{d}F_{\overline{\mathbf{C}}_{it}}(\overline{\mathbf{c}}), \tag{4}$$

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 $\Pr(D_{it} = d^H \mid \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}) > 0$ and $\Pr(D_{it} = d^L \mid \overline{\mathbf{C}}_{it} = \overline{\mathbf{c}}) > 0$ 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 $\mathbb{E}[Y_{it} \mid D_{it}, \overline{\mathbf{C}}_{it}]$ 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 & Thomas, 2011; VanderWeele & 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 DAG and its corresponding NPSEM (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 nonparametric 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 ϵ_k are mutually independent. In contrast to a linear structural equation model, nonparametric structural equations are entirely general—V_k may depend on any function of its parents and ϵ_k. The nonparametric structural equations encode counterfactual relationships between the variables that are represented on the graph. We review basic terminologies for NPSEM and DAGs in online supplementary material, 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 to be stable over time, while we allow for the change in the effect and sign of such relationships. For example, when the unemployment rate has none-zero causal effect on the incidence of hate crimes, this relationship satisfies structural stationarity, as long as this causal effect is nonzero over time, even if this causal effect changes from positive to negative over time. This can be seen as an extension of dynamic causal DAGs (Dean & Kanazawa, 1989; Pearl & 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:

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

where $(\epsilon^Y_{it}, \epsilon^L_{it}, \widetilde{\epsilon}^L_i, \epsilon^U_{gt})$ 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 noncausal associations 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

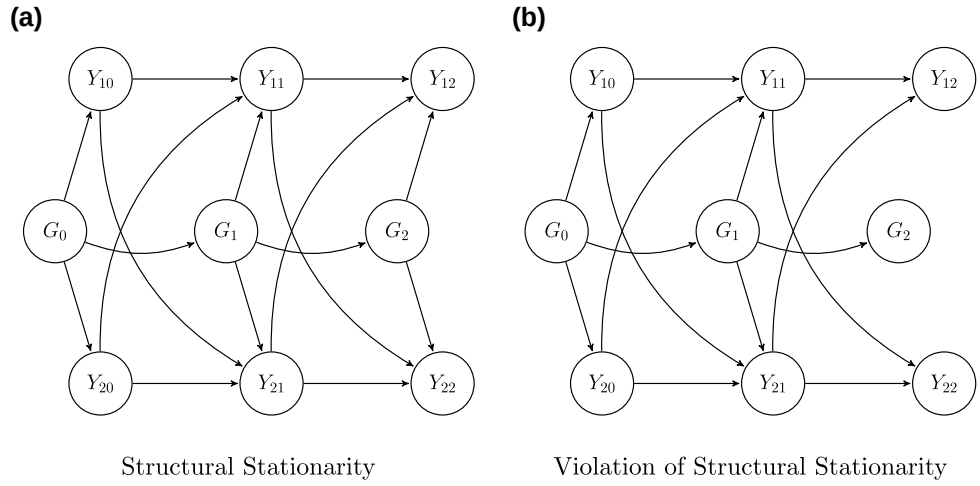


Figure 2. Illustration of structural stationarity. Note: Six nodes Y_{it} represent outcome variables for two individuals $i \in \{1, 2\}$ over three time periods $t \in \{0, 1, 2\}$. Three nodes G_t are contextual variables for $t \in \{0, 1, 2\}$. In the first panel, the causal structure between variables 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. (a) Structural stationarity and (b) violation of structural stationarity.

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 neighbours' outcomes at the last period $\mathbf{Y}_{\mathcal{N}_i,t-1}$ but also by the neighbours' 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 neighbours. 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 online supplementary material, Appendix C. □

Structural stationarity can accommodate many applied diffusion questions. Indeed, structural stationarity is often an implicit assumption researchers make in applied contexts. When analysing 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 (Ogburn & VanderWeele, 2014; O'Malley et al., 2014; Shalizi & Thomas, 2011) satisfy structural stationarity. Causal DAGs in the causal discovery literature often impose a similar but stronger condition (Danks & 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. If a new time-varying confounder arises in the middle of the time periods we analyse, 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. For example, suppose a new radical right party emerged in the middle of the time periods we analyse, and the party started a nation-wide political campaigns against refugees. Then, unless researchers can explicitly analyse data before and after this party emergence, structural stationarity is less plausible in this application.

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[3] to this underlying NPSEM.

Two points are worth noting about Assumption 2. First, it requires that the underlying data are 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 back-door 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.

[3] 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 DAG (Spirtes et al., 2000).

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 $\mathbf{Y}_{\mathcal{N}_i,t-1}$,

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

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 $\overline{\mathbf{C}}_{it}$,[4] $\mathbf{Y}_{\mathcal{N}_i,t-2}$ is a lag of the treatment variable, $\setminus$ is the set difference,[5] and $\mathrm{Des}(Y_{i,t-1})$ is a descendant of $Y_{i,t-1}$, i.e. variables affected by $Y_{i,t-1}$.

The proof of Theorem 1 is in online supplementary material, 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 unobservable. 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 neighbours' 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 $Y_{i,t-1} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{N}_i,t-1} \mid \overline{\mathbf{C}}^P_{it}$.

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

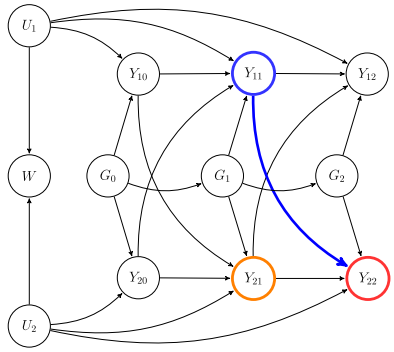
Although the proposed placebo test is applicable to any NPSEMs and corresponding causal DAGs that satisfy structural stationarity, we consider a causal DAG in Figure 3a as one concrete example. Suppose we are interested in the ACDE of Y_{11} on Y_{22} where Y_{11} is the treatment variable (blue), Y_{22}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

Based on this causal DAG in Figure 3a, table in Figure 3b 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, $Y_{11} \perp\!\!\!\perp Y_{21} \mid \overline{\mathbf{C}}^P$ 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.

[4] For example, suppose $\overline{\mathbf{C}}_{it} = (\mathbf{X}_{it}, \mathbf{X}_{i,t-1}, \widetilde{\mathbf{X}}_i)$ where $\mathbf{X}_{it}$ and $\mathbf{X}_{i,t-1}$ are time-varying covariates measured for unit i at time t and $t-1$, respectively, and $\widetilde{\mathbf{X}}_i$ are time-invariant covariates. In this case, $\overline{\mathbf{C}}^{(-1)}_{it} = (\mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2})$ because $\overline{\mathbf{C}}^{(-1)}_{it}$ only contains a lag of the time-varying variables in $\overline{\mathbf{C}}_{it}$. Importantly, recall that the overline indicates a history of variables, and thus $\overline{\mathbf{C}}_{it}$ include variables not only measured at time t but also those measured before time t.

[5] The set difference of B and A, denoted by $B \setminus A$, is the set of elements in B that are not in A.

(a)



Example of Placebo Test

(b)

	$\overline{C}$	$\overline{C}^P$	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

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 coloured blue. The placebo outcome Y_{21} is coloured orange. (a) Example of placebo test and (b) adjustment and placebo sets. DAG = directed acyclic graph; ACDE = average causal diffusion effect.

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 $Y_{11} \perp\!\!\!\perp Y_{21} \mid \overline{\mathbf{C}}^P$. In Figure 3a, 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; $Y_{11} \not\perp\!\!\!\perp Y_{21} \mid \overline{\mathbf{C}}^P$ due to a back-door path $(Y_{11} \leftarrow G_1 \to Y_{21})$. In online supplementary material, Appendix, we also provide an illustration with homophily bias.

4.2.3 Parametric placebo test via the spatial autoregressive model

As Theorem 1 is nonparametric, researchers can employ a variety of non-, semiparametric, 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 SAR model.

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

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 SAR 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 SAR coefficient δ serves as a test statistic of the placebo test. By testing whether this SAR coefficient is zero, researchers can

assess the no omitted confounders assumption and thus detect biases, including contextual confounding and homophily bias.

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 SAR coefficient as the strength of the spatial correlation, the proposed placebo test uses the SAR 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 Statistical power of the proposed placebo test is important because some falsification tests can have low power and cannot detect the violation of the key assumption in a finite sample. To investigate statistical power of the proposed placebo test, we provide simulation studies in online supplementary material, Appendix D.1, where we compare properties of the proposed placebo test and an 'oracle' test (a test that is possible only in simulations and this test checks whether an estimated causal effect is statistically distinguishable from the true causal effect). This 'oracle' test provides one type of theoretical upper bound and serves as a benchmark; if we cannot distinguish an estimated causal effect from the true causal effect even when we know the true effect in simulations, it is unlikely that bias can be detected by any feasible falsification test that does not assume knowledge of the true causal effect. We find that statistical power of the proposed approach is comparable to this theoretical upper bound and achieves about 80% of statistical power 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 practical tool to detect biases in applied settings. Examining how to further improve statistical power of the proposed test is an interesting topic for future studies. □

Remark It is important to note that if the parametric assumptions of the model are violated, the SAR 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 online supplementary material, 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.

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

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 SAR model as in equation (8).

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

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 SAR 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 ACDE 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 neighbourhood, i.e. d^H% of neighbouring 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 online supplementary material, 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 $\overline{\mathbf{c}}$,

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

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

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

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 neighbourhoods (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 & 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 & 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.

Remark It is important to emphasize that the time-invariance assumption can be violated in practice. This assumption is violated when the effects and imbalance of unobserved confounders change over time. For example, suppose again that the unemployment rate is the unobserved variable. Then, the time-invariance assumption is less plausible when the effect of the unemployment rates on the risk of hate crimes change over time and the strength of association between the treatment variable and the unemployment rates also change over time. To understand the potential influence of the violation of the time-invariance assumption, we provide a simulation study calibrated to the hate crime data. We show that the proposed bias-corrected estimator is not unbiased when the time-invariance assumption is violated, but it can reduce the bias and root mean squared error in a variety of settings where violation of the required time-invariance assumption is minor (online supplementary material, Appendix D.2). We also introduce a sensitivity analysis method in online supplementary material, Appendix A.4, such that researchers can quantitatively investigate the robustness of the bias-corrected estimates to the potential violation of the time-invariance assumption (Assumption 3).

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

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

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 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.

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

The proof is in online supplementary material, 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 (online supplementary material, Appendix D.2). We also introduce a sensitivity analysis method in online supplementary material, Appendix A.4, to investigate the robustness of the bias-corrected estimates to the potential violation of the time-invariance assumption (Assumption 3).

Remark Our work is closely connected to Liu and Tchetgen Tchetgen (2020) and Egami and Tchetgen Tchetgen (2024) in that our and these papers use placebo variables (also known as negative controls) in the presence of interdependence and interference between units. However, there are several key differences that make our approach and existing alternatives complementary to each other. First, both Liu and Tchetgen Tchetgen (2020) and Egami and Tchetgen Tchetgen (2024) focus on the network data rather than the spatial data setting. Thus, these existing methods are not directly applicable to our application of hate crime spatial diffusion. Second, to identify causal effects, each method makes different assumptions that are tailored to each application setting. While our paper requires the time-invariance assumption (Assumption 3), Liu and Tchetgen Tchetgen (2020) assume the correct specification of network formation process, and Egami and Tchetgen Tchetgen (2024) assume that researchers can find both a placebo outcome and a placebo treatment. Finding two placebo variables is relatively easier in the network data setting using a general strategy proposed in Egami and Tchetgen Tchetgen (2024), but the strategy does not apply to the spatial data setting and it could be much more difficult to find two credible placebo variables in the spatial applications. Thus, overall, these methods are complementary to each other and applicable to different application settings. □

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 the centre of counties i and j as $\widetilde{W}_{ij}$. We then row-standardize the initial matrix $\widetilde{\mathbf{W}}$ and obtain a final distance matrix $\mathbf{W}$. In this application, $\mathbf{W}$ is time-invariant and $W_{ii} = 0$. 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 neighbouring counties that experience the incidence of physical attacks in month $t-1$. The first causal quantity

Table 1. Five different adjustment sets

C1	$Y_{i,t-1}$, $D_{i,t-1}$, summary statistics of $\mathbf{W}_i(\lvert\mathcal{N}_i\rvert, \mathrm{Var}(\mathbf{W}_i))$
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)

of interest is the ACDE, which quantifies how much the probability of having hate crimes changes due to the increase in the proportion of neighbouring 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 neighbours 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 neighbouring 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 & Fowler, 2013; Eckles & Bakshy, 2017). The third adjustment set adds state fixed effects. Although the state fixed effects are often excluded from existing studies (e.g. Jäckle & König, 2016), we show how much these fixed effects help remove biases. Then, the fourth set adds a list of contextual 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 online supplementary material, 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 media, which we do not measure, are confounders, structural stationarity requires that such discussions in media, not only in Germany but also across the world, 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 (Angrist & Pischke, 2008; Athey & Imbens, 2006; Sofer et al., 2016), that the effect of debates about the refugee crisis in media is stable over time and the imbalance of unobserved media coverage is stable over time after adjusting for observed time-varying confounders. It is important to note that this assumption can be violated if the influence of media coverage in Germany and across the world becomes larger over time.

5.2 Estimation of ACDE

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.

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

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 neighbours experiencing hate crimes (common for safe areas in West Germany). Formally, $\hat{\tau} \equiv \int \{\widehat{\Pr}(Y_{it} = 1 \mid D_{it} = 0.27, \overline{\mathbf{C}}_{it}) - \widehat{\Pr}(Y_{it} = 1 \mid D_{it} = 0, \overline{\mathbf{C}}_{it})\}\, \mathrm{d}F_{\overline{\mathbf{C}}_{it}}(\overline{\mathbf{c}})$.

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.

Figure 4a and b presents 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 4a 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. While the placebo test cannot confirm the absence of unmeasured confounding (as no statistical test can confirm it), this also suggests that other unobserved confounders, such as debates about the refugee crisis in media, do not have strong confounding effects, after adjusting 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 4b. 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 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 4c 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 & 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

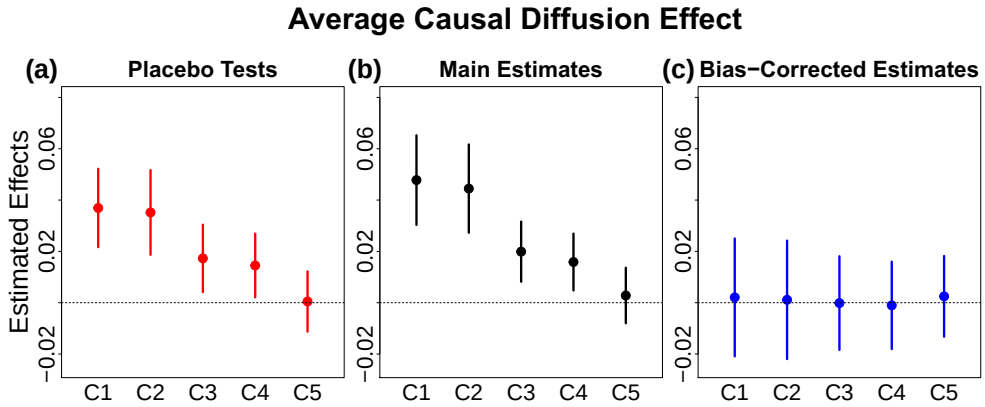


Figure 4. Placebo tests, main estimates, and bias-corrected estimates of the ACDE. Note: Figures (a), (b), and (c) present results from the placebo tests, estimates of the ACDE under the no omitted confounders assumption, and estimates from bias-corrected estimators with 95% confidence intervals, respectively. ACDE = average causal diffusion effect.

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 & 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 & Hiscox, 2007). Finally, schooling has incapacitation effects—keeping adolescents busy and off the street, thereby directly reducing the chances of committing crimes (Jacob & 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 analyse 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 online supplementary material, Appendix E. In particular, we estimate the 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 5a 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 5b. 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 5c 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.

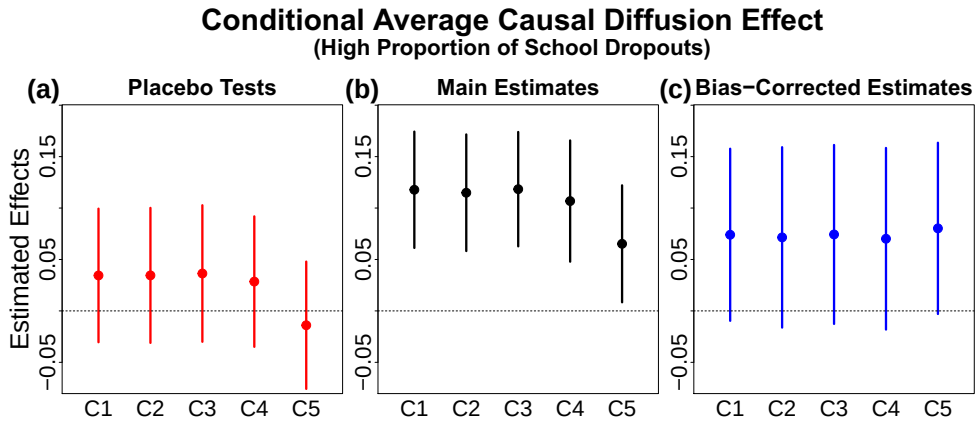


Figure 5. Placebo tests, main estimates, and bias-corrected estimates of the conditional ACDE for counties with a high proportion of school dropouts. Note: Figures (a), (b), and (c) present results from the placebo tests, estimates of the conditional ACDE under the no omitted confounders assumption,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

When we estimate the conditional ACDE for counties that have a lower proportion of school dropouts, effects are close to zero and their 95% confidence intervals cover zero, as the education hypothesis expects (see online supplementary material, 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 (Ogburn, 2018; Shalizi & Thomas, 2011; VanderWeele & An, 2013), 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 & König, 2016). This empirical analysis demonstrates the large differences in substantive conclusions that can result from contextual confounding.

For any method, it is important to understand its potential limitations. First, it is critical to assess the underlying assumptions (the structural stationarity assumption for the placebo test and the time-invariance assumption for the bias-corrected estimator) as they can be violated in practice. In particular, the time-invariance assumption is more plausible when the interval between each measured time period is shorter and researchers can adjust for a wide range of observed confounders.

On the other hand, it is less plausible when the effects and imbalance of unobserved confounders change over time quickly. To relax the time-invariance assumption, future studies can explore how to extend the double negative control framework (Egami & Tchetgen Tchetgen, 2024; Tchetgen Tchetgen, Ying, et al., 2020) to the spatial data setting. Second, it is important to emphasize that passing the placebo test does not guarantee the validity of the no omitted confounders assumption, as no statistical test can confirm the absence of unmeasured confounding. Specifically, passing the placebo test can be due to low power of the placebo test. Therefore, in practice, it is important to report point estimates and confidence intervals for the placebo test (not only the resulting p-value) and assess whether confidence intervals are large compared to confidence intervals for the ACDE main estimates. By doing so, we can check whether large p-values for the placebo test are due to large standard errors or small point estimates. We also show in online supplementary material, Appendix D.1, that the proposed placebo test has power comparable to a type of theoretical upper bound. When it is used with care, the placebo test can serve as a powerful practical tool to detect a wide range of biases in the causal diffusion settings.

Acknowledgments

I thank P.M. 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.

Conflict of interests: None declared.

Funding

None declared.

Data availability

The replication data and codes will be published in the Harvard Dataverse.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

An W. (2015). Instrumental variables estimates of peer effects in social networks. *Social Science Research*, *50*, 382–394. https://doi.org/10.1016/j.ssresearch.2014.08.011

Angrist J. D. (2014). The perils of peer effects. *Labour Economics*, *30*, 98–108. https://doi.org/10.1016/j.labeco.2014.05.008

Angrist J. D., & Pischke J. -S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Anselin L. (2013). *Spatial econometrics: Methods and models*. Springer.

Aral S., Muchnik L., & Sundararajan A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(51), 21544–21549. https://doi.org/10.1073/pnas.0908800106

Aronow P. M., & Samii C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics*, *11*(4), 1912–1947. https://doi.org/10.1214/16-AOAS1005

Athey S., & Imbens G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–497. https://doi.org/10.1111/ecta.2006.74.issue-2

Bramoullé Y., Djebbari H., & Fortin B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, *150*(1), 41–55. https://doi.org/10.1016/j.jeconom.2008.12.021

Braun R. (2011). The diffusion of racist violence in the Netherlands: Discourse and distance. *Journal of Peace Research*, *48*(6), 753–766. https://doi.org/10.1177/0022343311419238

Cai X., Loh W. W., & Crawford F. W. (2019). 'Identification of causal intervention effects under contagion', arXiv, arXiv:1912.04151, preprint: not peer reviewed.

Christakis N. A., & Fowler J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*, *32*(4), 556–577. https://doi.org/10.1002/sim.v32.4
Cohen-Cole E., & Fletcher J. M. (2008). Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics*, *27*(5), 1382–1387. https://doi.org/10.1016/j.jhealeco.2008.04.005
Conley T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, *92*(1), 1–45. https://doi.org/10.1016/S0304-4076(98)00084-0
Dancygier R. M., Egami N., Jamal A. A., & Rischke R. (2022). Hate crimes and gender imbalances: Fears over mate competition and violence against refugees. *American Journal of Political Science*, *66*(2), 501–515. https://doi.org/10.1111/ajps.12595
Danks D., & Plis S. (2013). Learning causal structure from undersampled time series. In *NIPS 2013 workshop on causality*.
Dean T., & Kanazawa K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, *5*(2), 142–150. https://doi.org/10.1111/coin.1989.5.issue-2
Duflo E., Dupas P., & Kremer M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739
Eckles D., & Bakshy E. (2017). 'Bias and high-dimensional adjustment in observational studies of peer effects', arXiv, arXiv:1706.04692, preprint: not peer reviewed.
Egami N. (2021). Spillover effects in the presence of unobserved networks. *Political Analysis*, *29*(3), 287–316. https://doi.org/10.1017/pan.2020.28
Egami N., & Tchetgen Tchetgen E. J. (2024). Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. https://doi.org/10.1093/jrsssb/qkad132
Fowler J. H., Heaney M. T., Nickerson D. W., Padgett J. F., & Sinclair B. (2011). Causality in political networks. *American Politics Research*, *39*(2), 437–480. https://doi.org/10.1177/1532673X10396310
Glaeser E. L., Sacerdote B., & Scheinkman J. A. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, *111*(2), 507–548. https://doi.org/10.2307/2946686
Gleditsch K. S., & Ward M. D. (2006). Diffusion and the international context of democratization. *International Organization*, *60*(4), 911–933. https://doi.org/10.1017/S0020818306060309
Goldsmith-Pinkham P., & Imbens G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, *31*(3), 253–264. https://doi.org/10.1080/07350015.2013.801251
Graham E. R., Shipan C. R., & Volden C. (2013). The diffusion of policy diffusion research in political science. *British Journal of Political Science*, *43*(03), 673–701. https://doi.org/10.1017/S0007123412000415
Granovetter M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. https://doi.org/10.1086/225469
Hainmueller J., & Hiscox M. J. (2007). Educated preferences: Explaining attitudes toward immigration in Europe. *International Organization*, *61*(2), 399–442. https://doi.org/10.1017/S0020818307070142
Halloran M. E., & Hudgens M. G. (2016). Dependent happenings: A recent methodological review. *Current Epidemiology Reports*, *3*(4), 297–305. https://doi.org/10.1007/s40471-016-0086-4
Halloran M. E., & Struchiner C. J. (1995). Causal inference in infectious diseases. *Epidemiology*, *6*(2), 142–151. https://doi.org/10.1097/00001648-199503000-00010
Hyttinen A., Plis S., Järvisalo M., Eberhardt F., & 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)* (pp. 216–227).
Jäckle S., & 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. https://doi.org/10.1080/01402382.2016.1215614
Jacob B. A., & Lefgren L. (2003). Are idle hands the devil's workshop? incapacitation, concentration, and juvenile crime. *American Economic Review*, *93*(5), 1560–1577. https://doi.org/10.1257/000282803322655446
Jones J. J., Bond R. M., Bakshy E., Eckles D., & 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. https://doi.org/10.1371/journal.pone.0173851
Lipsitch M., Tchetgen Tchetgen E. J., & Cohen T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, *21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb
Liu L., & Tchetgen Tchetgen E. (2020). 'Regression-based negative control of homophily in dyadic peer effect analysis', arXiv, arXiv:2002.06521, preprint: not peer reviewed.
Lochner L., & Moretti E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, *94*(1), 155–189. https://doi.org/10.1257/000282804322970751
Lyons R. (2011). The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, *2*(1). https://doi.org/10.2202/2151-7509.1024

Manski C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, *60*(3), 531–542. https://doi.org/10.2307/2298123

McFowland III E., & Shalizi C. R. (2021). Estimating causal peer influence in homophilous social networks by inferring latent locations. *Journal of the American Statistical Association*, *118*(541), 707–718. https://doi.org/10.1080/01621459.2021.1953506

Miao W., & 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. https://doi.org/10.1093/aje/kwx012

Miguel E., & Kremer M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217. https://doi.org/10.1111/ecta.2004.72.issue-1

Morozova O., Cohen T., & Crawford F. W. (2018). Risk ratios for contagious outcomes. *Journal of The Royal Society Interface*, *15*(138), 20170696. https://doi.org/10.1098/rsif.2017.0696

Myers D. J. (2000). The diffusion of collective violence: Infectiousness, susceptibility, and mass media networks. *American Journal of Sociology*, *106*(1), 173–208. https://doi.org/10.1086/303110

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. http://www.jstor.org/stable/2245382

Ogburn E. L. (2018). Challenges to estimating contagion effects from observational data. In S. Lehmann, & Y-Y. Ahn (Eds.), *Complex spreading phenomena in social systems* (pp. 47–64). Springer.

Ogburn E. L., Shpitser I., & Lee Y. (2020). Causal inference, social networks, and chain graphs. *The Journal of the Royal Statistical Society, Series A*, *183*(4), 1659–1676. https://doi.org/10.1111/rssa.12594

Ogburn E. L., & VanderWeele T. J. (2014). Causal diagrams for interference. *Statistical Science*, *29*(4), 559–578. https://doi.org/10.1214/14-STS501

O'Malley A. J., Elwert F., Rosenquist J. N., Zaslavsky A. M., & Christakis N. A. (2014). Estimating peer effects in longitudinal dyadic data using instrumental variables. *Biometrics*, *70*(3), 506–515. https://doi.org/10.1111/biom.v70.3

Pearl J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl J., & Russell S. (2001). Bayesian networks. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 157–160). 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. https://doi.org/10.1016/0270-0255(86)90088-6

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), 688–701. https://doi.org/10.1037/h0037350

Sacerdote B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly Journal of Economics*, *116*(2), 681–704. https://doi.org/10.1162/00335530151144131

Sävje F. (2021). 'Causal inference with misspecified exposure mappings', arXiv, arXiv:2103.06471, preprint: not peer reviewed.

Shalizi C. R., & Thomas A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239. https://doi.org/10.1177/0049124111404820

Sinclair B. (2012). *The social citizen: Peer networks and political behavior*. University of Chicago Press.

Sofer T., Richardson D. B., Colicino E., Schwartz J., & Tchetgen Tchetgen E. J. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical Science*, *31*(3), 348. https://doi.org/10.1214/16-STS558

Spirtes P., Glymour C. N., & Scheines R. (2000). *Causation, prediction, and search*. MIT Press.

Tchetgen Tchetgen E. J., Fulcher I., & Shpitser I. (2021). Auto-g-computation of causal effects on a network. *Journal of American Statistical Association*, *116*(534), 833–844. https://doi.org/10.1080/01621459.2020.1811098

Tchetgen Tchetgen E. J., Ying A., Cui Y., Shi X., & Miao W. (2020). 'An introduction to proximal causal learning', arXiv, arXiv:2009.10982, preprint: not peer reviewed.

United Nations High Commissioner for Refugees (2017). *Global trends: Forced displacement in 2017*.

VanderWeele T. J., & An W. (2013). Social networks and causal inference. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 353–374). Springer.

VanderWeele T. J., Ogburn E. L., & 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). https://doi.org/10.1515/2151-7509.1050

Wilson J. Q., & Kelling G. L. (1982). Broken windows. *Atlantic Monthly*, *249*(3), 29–38. https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/