Research Article

Corinne Emmenegger*, Meta-Lina Spohn, Timon Elmer, and Peter Bühlmann

Treatment effect estimation with observational network data using machine learning

https://doi.org/10.1515/jci-2023-0082
received December 17, 2023; accepted January 04, 2025

Abstract: Causal inference methods for treatment effect estimation usually assume independent units. However, this assumption is often questionable because units may interact, resulting in spillover effects between them. We develop augmented inverse probability weighting (AIPW) for estimation and inference of the expected average treatment effect (EATE) with observational data from a single (social) network with spillover effects. In contrast to overall effects such as the global average treatment effect, the EATE measures, in expectation and on average over all units, how the outcome of a unit is causally affected by its own treatment, marginalizing over the spillover effects from other units. We develop cross-fitting theory with plugin machine learning to obtain a semiparametric treatment effect estimator that converges at the parametric rate and asymptotically follows a Gaussian distribution. The asymptotics are developed using the dependency graph rather than the network graph, which makes explicit that we allow for spillover effects beyond immediate neighbors in the network. We apply our AIPW method to the Swiss StudentLife Study data to investigate the effect of hours spent studying on exam performance accounting for the students' social network.

Keywords: dependent data, interference, observed confounding, semiparametric inference, spillover effects

MSC 2020: 62D20, 62G20

1 Introduction

Classical causal inference approaches for treatment effect estimation with observational data usually assume independent units. This assumption is part of the common stable unit treatment value assumption (SUTVA) [1]. However, independence is often violated in practice due to interactions among units that lead to so-called spillover effects. For example, the vaccination against an infectious disease (treatment) of a person (unit) may not only influence this person's health status (outcome) but may also protect the health status of other people the person is interacting with [2,3]. In the presence of spillover effects, standard algorithms fail to separate correlation from causation, and spurious associations due to network dependence contribute to the replication crisis [4] and may yield biased causal effect estimators and invalid inference [2,4–8]. New approaches are required to guarantee valid causal inference from observational data with spillover effects.

We consider the following types of spillover effects: (i) causal effects of other units' treatments on a given unit's outcome, referred to as interference in the literature [5,9], and (ii) causal effects of other units' covariates

*** Corresponding author: Corinne Emmenegger,** Seminar for Statistics, ETH Zurich, Zurich, Switzerland,
e-mail: emmenegger@stat.math.ethz.ch
Meta-Lina Spohn, Peter Bühlmann: Seminar for Statistics, ETH Zurich, Zurich, Switzerland
Timon Elmer: Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland

on a given unit's treatment or outcome.[1] The spillover effects a unit receives are governed by proximity of this unit to other units in a known undirected network G. The edges of this network represent some kind of interaction or relationship of the respective units such as friendship, geographical closeness, or shared department in a company.

In this article, the causal effect of interest and target of inference is the expected average treatment effect (EATE) [3] in an observational setting. The EATE measures, in expectation and on average over all units, how the outcome of a unit is causally affected by its own treatment in the presence of spillover effects from other units. The EATE is the statistical parameter when the question is how, on average for all units, the outcome of a specific unit is influenced when only its own treatment is altered. In the infectious disease example, the EATE measures the average expected difference in health status of an individual assigned to the vaccination versus not, marginalizing over unit-specific covariates and spillover effects of other people. This corresponds to the medical effect of the vaccine in a person's body. This interpretation highlights that the EATE is not an estimand for policy evaluation, where, for example, one is interested in capturing the effect of jointly vaccinating a sample of the population.

We now formalize the EATE following the study by Sofrygin and van der Laan [10]. For each unit $i = 1, 2, \ldots, N$, let $W_i \in \{0, 1\}$ be the dichotomous treatment, Y_i be the response, and C_i be the covariates of unit i. The N units are connected in a fixed undirected network G in which they may exhibit spillover effects of the two aforementioned types (i) and (ii) from their immediate neighbors and/or units further away. Let P_L^N be the observational distribution of $O = (W_i, C_i, Y_i)_{i=1,\ldots,N}$, where L is the distribution of $W = (W_1, \ldots, W_N)$ given $C = (C_1, \ldots, C_N)$. Let $P_{\tilde{L}}^N$ be the distribution of $\widetilde{O} = (\widetilde{W}_i, C_i, \widetilde{Y}_i)_{i=1,\ldots,N}$, where the conditional distribution L of W given C has been replaced by the user-defined distribution $\tilde{L}$. This distribution $\tilde{L}$ describes the intervention on the treatment vector W that the researcher is interested in. We can then define the EATE as

$$\theta_N^0 = \theta_N^0(1) - \theta_N^0(0),$$

where

$$\theta_N^0(w) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{P^N_{\tilde{L}_i(w)}}[Y_i^{\mathrm{do}(W=\tilde{L}_i(w))}],$$

where we use the do-notation of Pearl [11] and

$$\tilde{L}_i(w) = (W_1, \ldots, W_{i-1}, w, W_{i+1}, \ldots, W_N), \quad \text{for } w \in \{0,1\},$$

represents the intervention on the unit-specific treatment W_i (setting it to constant w), but the distribution of treatments W_j for the other $N - 1$ units j in the network are left unchanged. In particular, the intervention $\tilde{L}_i(w)$ is independent of C. Thus, $\theta_N^0(1)$ evaluates a collection of unit-specific distributions, $(\tilde{L}_1(1), \ldots, \tilde{L}_N(1))$, which cannot be rewritten as a single intervention $\tilde{L}$ on the whole treatment vector W. By denoting the EATE by θ_N^0, it remains implicit that it is defined conditional on a specific network G, while it is explicit that it is a function of the given sample size N. Consequently, the EATE's true value can vary depending on the sample size and the network structure.

To simplify notation, we rewrite the EATE by

$$\theta_N^0 = \frac{1}{N}\sum_{i=1}^{N} \theta_i^0,$$

where

$$\theta_i^0 = \mathbb{E}_{W_{-i}, C_{-i}, C_i}[\mathbb{E}[Y_i^{do(W_i=1)} - Y_i^{do(W_i=0)} | W_{-i}, C_{-i}, C_i]],$$

and $W_{-i} = (W_1, \ldots, W_{i-1}, W_{i+1}, \ldots, W_N)$ and $C_{-i} = (C_1, \ldots, C_{i-1}, C_{i+1}, \ldots, C_N)$. Thus, the EATE equals the average of the unit-specific treatment effects θ_i^0, i.e., the expected difference in outcomes Y_i if the treatment was assigned to

1 Another notion of spillover effects is frequently used in the social sciences; please see Section B in the appendix for a discussion.

unit i versus if it was retained from unit i. The unit-specific treatment effects may not be the same for all units because the outcomes may have different distributions conditional on W_{-i} and C_{-i} across units due to the spillover effects. In the setting without spillover effects, the distribution of Y_i does not depend on W_{-i} and C_{-i}, for each $i = 1, \ldots, N$, and thus, the EATE coincides with the average treatment effect (ATE) if spillover effects are absent [12,13].

We impose the following key assumption (that is standard in this literature [6,10,14]): the spillover effects can be summarized by lower-dimensional features, i.e., we will use domain knowledge-informed features that are arbitrary functions of the network G and the treatment and covariate vectors of all units [15,16]. The features are assumed to capture all pathways through which spillover effects take place. For example, Cai et al. [17] and Leung [18] model the purchase of a weather insurance (outcome) of farmers in rural China as a function of attending a training session (treatment) and the proportion of friends who attend the session (feature on direct neighbors in the network).

In the following, we will assume a structural equation model (SEM) to impose our assumptions on the data-generating mechanism of the joint distribution of $(W_i, C_i, Y_i)_{i=1\ldots,N}$. The outcome and propensity score model of the SEM may be highly complex and nonsmooth and include interactions and high-dimensional variables. We then follow an augmented inverse probability weighting (AIPW) [19] approach to estimate the EATE θ_N^0 in the context of this model. We estimate the outcome and propensity score models with arbitrary machine learning algorithms and plug them into our AIPW estimand identifying θ_N^0. These machine learning estimators may be highly complex and suffer from regularization bias and slow converge rates. However, the use of sample splitting with cross-fitting [20] allows us to address these issues. Limiting the growth of dependencies between units, our estimator of the EATE is consistent, converges at the $\sqrt{N}$-rate, and asymptotically follows a Gaussian distribution. This allows us to construct confidence intervals and p-values.

1.1 Our contribution and comparison to literature

Our work is most related to the literature on semiparametric treatment effect estimation and inference with observational data from a single network. Tchetgen Tchetgen et al. [21] developed a network version of the g-formula [22] and performed outcome regression, assuming that the data can be represented as a chain graph, which is a graphical model that is generally incompatible with our SEM approach [23]. An SEM approach is also used by van der Laan [14], Sofrygin and van der Laan [10], and Ogburn et al. [6]. These works considered a similar model as we do and proposed semiparametric treatment effect estimation by targeted maximum likelihood (TMLE) [24–26]. van der Laan [14] and Ogburn et al. [6] primarily considered global effects that compare two hypothetical interventions on the whole treatment vector. An example of such an effect is the global average treatment effect (GATE), which contrasts the interventions of treating all units of the population versus treating no unit of the population. In contrast, we consider the EATE that is the average effect of assigning the treatment to one unit versus not and integrate out the treatment selections from the other units. Causal effects like the EATE summarizing the effect of N unit-specific interventions generally cannot be described using a single intervention on the whole treatment vector, as done for global effects. The behavior of estimators for the EATE under the wrong independent and identically distributed assumption is studied by Sävje et al. [3]. Sofrygin and van der Laan [10] mentioned a possible extension to estimate the EATE with TMLE, but all their results are for global effects such as the GATE. Their theory assumes some kind of a bounded entropy integral, which is difficult to verify for machine learning methods.

Our contribution includes the following. First, we present a semiparametric, machine learning-based approach to estimate the EATE with observational data from a single network. Our approach enables performing inference, including confidence intervals and p-values. Particularly, we do not require multiple disjoint networks. We develop a cross-fitting algorithm under interference and reason in terms of the dependency graph to explicitly allow for different interactions, also specifically ones that are beyond immediate neighbors in the network. Second, the limiting asymptotic Gaussian distribution and optimal $\sqrt{N}$-convergence

rate of the EATE estimator are achieved even if the number of ties of a unit may diverge asymptotically. To reach this optimal convergence rate to estimate global effects, Ogburn et al. [6] need to uniformly bound the neighborhood size of a unit. Third, our algorithm based on sample splitting is easy to understand and implement, and the user may choose any machine learning algorithm. Fourth, we analyze the Swiss StudentLife Study data [27,28] and estimate the effect of study time on the grade point average (GPA) of freshmen students after their first-year examinations at one of the world's leading universities.

Outline of this article. Section 2 presents the model assumptions, characterizes the treatment effect of interest, outlines the procedures for the point estimation of the EATE and estimation of its variance, and establishes asymptotic results. Section 3 demonstrates our methodological and theoretical developments in a simulation study and on empirical data of the StudentLife Study.

2 Framework and our network AIPW estimator

2.1 Model formulation

We consider $i = 1,\dots, N$ units interacting in a known undirected network G. For each unit i, we observe a binary treatment $W_i \in \{0,1\}$, a univariate outcome Y_i, and a possibly multivariate vector of observed covariates C_i that may causally affect W_i and Y_i. The outcome Y_i may be dichotomous or continuous, and the potentially multivariate covariates C_i may consist of discrete and continuous components. Irrespective of whether the outcomes are continuous or dichotomous, we can consider the following SEM with additive error terms for $i = 1,\dots, N$:

$$\begin{aligned} C_i &\leftarrow \varepsilon_{C_i}, \\ Z_i &\leftarrow (f_z^1(C_{-i}, G), \dots, f_z^t(C_{-i}, G)), \\ W_i &\leftarrow h^0(C_i, Z_i) + \varepsilon_{W_i}, \\ X_i &\leftarrow (f_x^1(W_{-i}, C_{-i}, G), \dots, f_x^r(W_{-i}, C_{-i}, G)), \\ Y_i &\leftarrow W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i}, \end{aligned} \tag{1}$$

where the errors ε_{W_i} and ε_{Y_i} are jointly independent, we have $\mathbb{E}[\varepsilon_{W_i}|C_i, Z_i] = 0$ and $\mathbb{E}[\varepsilon_{Y_i}|W_i, C_i, X_i] = 0$, the errors satisfy the assumptions in the study by Bühlmann [29] (required for the bootstrap variance results in Appendix F), and the ε_{W_i}'s are identically distributed and the ε_{Y_i}'s are identically distributed (required for the alternative variance results in Appendix G). We note that the identical distribution of the error terms is only required for our approach to estimate standard errors. The vector $C_{-i} = (C_1, C_2, \dots, C_{i-1}, C_{i+1}, \dots, C_N)$ denotes the vector of covariates of units $j \neq i$, and W_{-i} is similarly defined. The binary treatments W_i can be thought of as Bernoulli$(h^0(C_i, Z_i))$ realizations. A constant h^0 corresponds to a Bernoulli experiment. This SEM encodes the assumption that the covariates C_i and features X_i suffice to control for confounding of the effect of the treatment on the outcome. The propensity score function $h^0(\cdot,\cdot)$ and the outcome model consisting of $g_1^0(\cdot,\cdot)$ and $g_0^0(\cdot,\cdot)$ are fixed but unknown functions that all units share. Nevertheless, the distribution of the responses may differ across units due to the X-spillover that captures effects from, e.g., a unit's neighbors' covariates and treatment assignments as described next. Because every unit may have a different number of neighbors, the X_i's may follow a different distribution across different units, resulting in non-fixed distributions of the responses across units. Furthermore, the individual equations in (1) have to be understood in a distributional sense in that, if, e.g., $g_1^0 \equiv 0 \equiv g_0^0$, we have $Y_i = \varepsilon_{Y_i}$ in distribution only.

The functions f_z^l, $l \in [t]$ and f_x^l, $l \in [r]$, which are shared by all units and used to build the Z- and X-features, are assumed to be known, and their concatenations are assumed to be of fixed dimensions t and r, respectively. This is analogous to the in-practice considerations in Ogburn et al. [6]. We also allow for features of further degree neighbors: for example, f_x^1 might capture the fraction of treated units that are a distance of 2 from a given unit in the network G. Making use of an implied dependency graph gives a more

transparent formulation (Section 2.3). Since the network G is undirected, our spillover effects are assumed to be reciprocal, i.e., if unit i receives spillover effects from unit j through W_j and/or C_j, then unit j also receives spillover effects from unit i through W_i and/or C_i. Example 1 illustrates the construction of 2-dimensional X-features. Importantly, the X- and Z-features render the unit-level data dependent. In addition, the distributions of propensity scores and outcomes are not generally identical across units due to distributional differences of these features.

Example 1. Consider the network in Figure 1, where gray nodes take the treatment and white ones do not. We choose $r = 2$ many X-features and discard any influence of C_j in X_i, i.e., $f_x^l(\{(W_j, C_j)\}_{j\in[N]\setminus\{i\}}, G) = f_x^l(\{(W_j)\}_{j\in[N]\setminus\{i\}}, G)$ for $l = 1, 2$. Given a unit i, we choose the first feature in X_i as the fraction of treated neighbors of unit i and the second feature as the fraction of treated neighbors of neighbors of i. Let us consider unit $i = 6$ in Figure 1. Its neighbors are the units 2, 5, and 7, and its neighbors of neighbors are the units 1 and 3 (neighbors of unit 2) and unit 8 (neighbor of unit 7), where we exclude $i = 6$ from its second degree neighborhood by definition. Therefore, we have $X_6 = (1/3, 2/3)$ because one out of three neighbors is treated and two out of three neighbors of neighbors are treated. The whole 9×2 dimensional X-feature matrix is obtained by applying the same computations to all other units i.

2.2 Treatment effect and identification

Plugging in the outcome equation of the SEM (1), we can rewrite the treatment effect of interest, the EATE, as

$$\begin{aligned}\theta_N^0 &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{W_{-i},C_{-i},C_i}[\mathbb{E}[Y_i^{do(W_i=1)} - Y_i^{do(W_i=0)}|W_{-i}, C_{-i}, C_i]] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{C_i,X_i}[\mathbb{E}_{\varepsilon_{Y_i}}[Y_i|\,\mathrm{do}(W_i = 1), C_i, X_i] - \mathbb{E}_{\varepsilon_{Y_i}}[Y_i|\,\mathrm{do}(W_i = 0), C_i, X_i]] \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{C_i,X_i}[g_1^0(C_i, X_i) - g_0^0(C_i, X_i)],\end{aligned} \tag{2}$$

where we obtain that the unit-specific treatment effect of unit i is $\theta_i^0 = \mathbb{E}_{C_i,X_i}[g_1^0(C_i, X_i) - g_0^0(C_i, X_i)]$. Particularly, we assume that given the observable confounders C_i and features X_i, we can replace the do-operator by respective conditioning. The expectation $\mathbb{E}_{C_i,X_i}$ over C_i and X_i is with respect to the observational distributions of C_i and X_i, as defined by the SEM (1). This notation makes explicit that the EATE is conditional on N, whereas it remains implicit that it is also conditional on the network G. We refer to the study by Ogburn et al. [6] for a discussion of the interpretation of such conditional effects.

Estimating g_1^0 and g_0^0 by regression machine learning algorithms and plugging them into (2) would not result in a parametric convergence rate and an asymptotic Gaussian distribution of the so-obtained estimator. To obtain asymptotic normality with convergence at the $\sqrt{N}$-rate, a centered correction term involving the propensity score h^0 is added to $g_1^0(C_i, X_i) - g_0^0(C_i, X_i)$, and we can identify the EATE as follows.

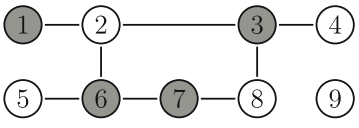

Figure 1: Network on nine units where the node label represents the number of a unit. Gray nodes receive the treatment, corresponding to $W_i = 1$, and white ones do not, corresponding to $W_i = 0$.

Lemma 2.1. *Let $i \in [N]$. Let*

$$S_i = (C_i, Z_i, W_i, X_i, Y_i) \tag{3}$$

be the concatenation of the observed variables for unit i. For concatenations $\eta = (g_1, g_0, h)$ of general nuisance functions g_1, g_0, and h , consider the score

$$\varphi(S_i, \eta) = g_1(C_i, X_i) - g_0(C_i, X_i) + \frac{W_i}{h(C_i, Z_i)}(Y_i - g_1(C_i, X_i)) - \frac{1 - W_i}{1 - h(C_i, Z_i)}(Y_i - g_0(C_i, X_i)), \tag{4}$$

including the aforementioned correction term. For the true nuisance functions $\eta^0 = (g_1^0, g_0^0, h^0)$, we have $\mathbb{E}[\varphi(S_i, \eta^0)] = \theta_i^0$ and can consequently identify the EATE (2) by

$$\theta_N^0 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\varphi(S_i, \eta^0)]. \tag{5}$$

The aforementioned expectation is with respect to the law of S_i , but we omit it for notational simplicity.

The proof of Lemma 2.1 is provided in Appendix E. Based on this lemma, we will present our estimator of θ_N^0 in Section 2.4. The true nuisance functions $\eta^0 = (g_1^0, g_0^0, h^0)$ are not of statistical interest, but have to be estimated to build an estimator of θ_N^0, and we will estimate them using regression machine learning algorithms. Such machine learning estimators might suffer from regularization bias and converge slower than at the $\sqrt{N}$-rate. However, the two correction terms $W_i/h(C_i, Z_i)(Y_i - g_1(C_i, X_i))$ and $(1 - W_i)/(1 - h(C_i, Z_i))(Y_i - g_0(C_i, X_i))$ make the score φ Neyman orthogonal, which counteracts the effect of regularization bias. Moreover, the machine learning estimators are only required to converge at a moderate rate; please see Section 2.4 for further details.

Scharfstein et al. [30] and Bang and Robins [31] considered a similar score φ for causal effect estimation and inference under the SUTVA assumption, and their function is based on the influence function for the mean for missing data from Robins and Rotnitzky [32]. Moreover, it is also used to compute the AIPW estimator under SUTVA, and our score φ defined in (4) coincides with one of the AIPW approaches under SUTVA if we omit the X- and Z-spillover features. In this case, we can reformulate φ as

$$\varphi(S_i, \eta^0) = \frac{W_i Y_i}{e(C_i)} - \frac{(1 - W_i)Y_i}{(1 - e(C_i))} - \frac{W_i - e(C_i)}{e(C_i)(1 - e(C_i))}((1 - e(C_i))\mathbb{E}[Y_i|W_i = 1, C_i] + e(C_i)\mathbb{E}[Y_i|W_i = 0, C_i]),$$

where $e(C_i) = \mathbb{E}[W_i|C_i] = h^0(C_i)$ denotes the propensity score, $\mathbb{E}[Y_i|W_i = 1, C_i] = g_1^0(C_i, X_i)$, and $\mathbb{E}[Y_i|W_i = 0, C_i] = g_0^0(C_i, X_i)$. This equivalence remains true if the true nuisance functions are replaced by their estimators.

2.3 Dependency graph

Depending on the feature functions that are used, if an edge connects two units in the network G, the units may be dependent. However, the absence of an edge in G does not necessarily imply independence of the respective units. Subsequently, we present a second graph where the presence of an edge represents dependence and its absence independence of the variables of the two respective units. Our theoretical results will be established based on this so-called dependency graph [3]. Example 2 illustrates the concept.

Definition 1. Dependency graph on $S_i, i \in [N]$ [3]: the dependency graph $G_D = (V, E_D)$ on the unit-level data S_i, $i \in [N]$ defined in (3), is an undirected graph on the node set V of the network $G = (V, E)$ with potentially larger edge set E_D than E. An undirected edge $\{i, j\}$ between two nodes i and j from V belongs to E_D if at least one of the following two conditions holds: (1) there exists an $m \in [N]\backslash\{i, j\}$ such that W_m and/or C_m are present in both X_i and X_j or are present in both Z_i and Z_j and (2) W_i is present in X_j, or C_i is present in X_j or in Z_j, i.e., units i and j receive spillover effects from at least one common third unit, or they receive spillover effects from each other.

Example 2. Consider the chain-shaped network G in Figure 2 on the left. We consider as 1-dimensional X-spillover effect the fraction of treated direct neighbors in the network G and no Z-spillover. The resulting dependency graph G_D is displayed in the middle of Figure 2. In G_D, unit 2 shares an edge with units 1 and 3

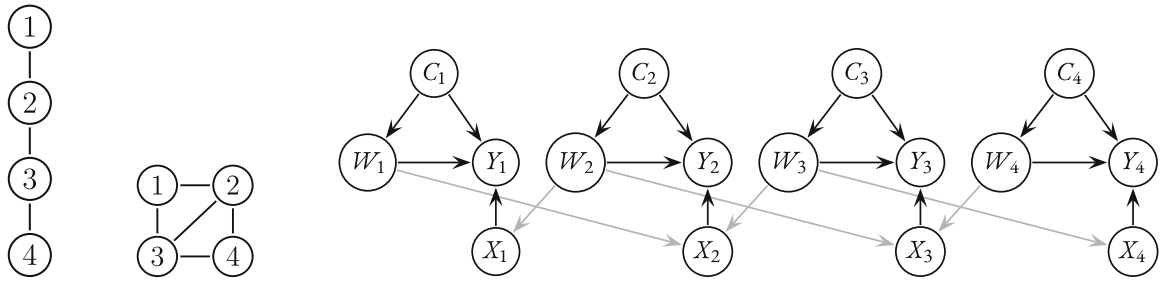

Figure 2: Network G on four units (left), where the spillover effects come from the treatments of the direct neighbors, which results in a distance-two dependence, which is displayed in the corresponding dependency graph G_D (middle). The underlying causal DAG is displayed on the right, where arrows due to X-spillover effects are gray.

because these units are neighbors of 2 in the network. Unit 2 also shares an edge with 4 in G_D because it shares its neighbor 3 with unit 4. The right-hand of Figure 2 displays the causal directed acyclic graph (DAG) on all units corresponding to this model, including confounders C. Due to the definition of the X-spillover effect, we have $X_1 = W_2$ and $X_4 = W_3$. Consequently, using graphical criteria [33–37], we infer that the unit-level data $S_1 = (C_1, W_1, X_1, Y_1)$ are independent of $S_4 = (C_4, W_4, X_4, Y_4)$.

The dependency graph is a function of the network G as well as the Z-and X-features. Constraining the growth of the maximal degree of this graph allows us to obtain a CLT result for our treatment effect estimator.

2.4 Estimation procedure and asymptotics

Subsequently, we describe our estimation procedure and its asymptotic properties. We use sample splitting and cross-fitting to estimate the EATE θ_N^0 identified by equation (5) as follows. We randomly partition $[N]$ into $K \geq 2$ sets of approximately equal size that we call $I_1, \ldots, I_K$. We split the unit-level data according to this partition into the sets $\mathcal{S}_{I_k} = \{S_i\}_{i \in I_k}$, $k \in [K]$. For each $k \in [K]$, we perform the following steps. First, we estimate the nuisance functions g_1^0, g_0^0, and h^0 on the complement set of $\mathcal{S}_{I_k}$, which we define as

$$\mathcal{S}_{I_k^c} = \{S_j\}_{j \in [N]} \backslash (\mathcal{S}_{I_k} \cup \{S_m | \exists i \in I_k : (i, m) \in E_D\}), \tag{6}$$

where E_D denotes the edge set of the dependency graph G_D. Particularly, $\mathcal{S}_{I_k^c}$ consists of unit-level data S_j from units j that do not share an edge with any unit $i \in I_k$ in the dependency graph. Consequently, the set $\mathcal{S}_{I_k^c}$ contains all S_j's that are independent of the data in $\mathcal{S}_{I_k}$. To estimate g_1^0, we select the S_i's from $\mathcal{S}_{I_k^c}$ whose W_i equals 1 and regress the corresponding outcomes Y_i on the confounders C_i and the features X_i, which yields the estimator $\hat{g}_1^{I_k^c}$. Similarly, to estimate g_0^0, we select the S_i's from $\mathcal{S}_{I_k^c}$ whose W_i equal 0 and perform an analogous regression, which yields the estimator $\hat{g}_0^{I_k^c}$. To estimate h^0, we use the whole set $\mathcal{S}_{I_k^c}$ and regress W_i on the confounders C_i and the features Z_i, which yields the estimator $\hat{h}^{I_k^c}$[2]. These regressions may be carried out with any machine learning algorithm. We concatenate these nuisance function estimators into the nuisance parameter estimator $\hat{\eta}^{I_k^c} = (\hat{g}_1^{I_k^c}, \hat{g}_0^{I_k^c}, \hat{h}^{I_k^c})$ and plug it into φ that is defined in (4). We then evaluate the so-obtained function $\varphi(\cdot, \hat{\eta}^{I_k^c})$ on the data $\mathcal{S}_{I_k}$, which yields the terms $\varphi(S_i, \hat{\eta}^{I_k^c})$ for $i \in I_k$, i.e., we evaluate $\varphi(\cdot, \hat{\eta}^{I_k^c})$ on unit-level data S_i that is independent of the data that were used to estimate the nuisance parameter $\hat{\eta}^{I_k^c}$. Finally, we estimate the EATE by the cross-fitting estimator

$$\hat{\theta} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{|I_k|} \sum_{i \in I_k} \varphi(S_i, \hat{\eta}^{I_k^c}) \right) \tag{7}$$

2 If the treatment is randomized with a known probability, we do not have to estimate the propensity function h^0 and set it to the randomization probability instead.

that averages over all K folds. The estimator $\hat{\theta}$ converges at the parametric rate, $N^{-1/2}$, and follows a Gaussian distribution asymptotically with limiting variance σ_∞^2 as stated in Theorem 2.2.

The partition $I_1,\ldots,I_K$ is random. To alleviate the effect of this randomness, the whole procedure is repeated a number of B times, and the median of the individual point estimators over the B repetitions is our final estimator of θ_N^0. The asymptotic results for this median estimator remain the same as for $\hat{\theta}$ (see Chernozhukov et al. [20]). For each repetition $b \in [B]$, we compute a point estimator $\hat{\theta}_b$, a variance estimator $\hat{\sigma}_{\infty,b}^2$ (for details, please see Section 2.5), and a p-value p_b for the two-sided test $H_0 : \theta_N^0 = 0$ versus $H_A : \theta_N^0 \neq 0$. The B many p-values $p_1,\ldots,p_B$ from the individual repetitions are aggregated according to

$$p_{\text{aggr}}^0 = 2\text{median}_{b\in[B]}(p_b).$$

This aggregation scheme yields a valid overall p-value for the same two-sided test [38]. The corresponding confidence interval is constructed as

$$\text{CI}(\hat{\theta}) = \{\theta \in \mathbb{R} | p_{\text{aggr}}^\theta \text{ of testing } H_0 : \theta_N^0 = \theta \text{ vs } H_A : \theta_N^0 \neq \theta \text{ satisfies } p_{\text{aggr}}^\theta > \alpha\}, \tag{8}$$

where typically $\alpha = 0.05$. This set contains all values θ for which the null hypothesis $H_0 : \theta_N^0 = \theta$ cannot be rejected at level α against the two-sided alternative $H_A : \theta_N^0 \neq \theta$.

Next, we describe how $\text{CI}(\hat{\theta})$ can easily be computed. Due to the asymptotic result of Theorem 2.2, the aggregated p-value p_{aggr}^θ for $\theta \in \mathbb{R}$ can be represented as

$$p_{\text{aggr}}^\theta = 4\text{median}_{b\in[B]}(1 - \Phi(\sqrt{N}\hat{\sigma}_{\infty,b}^{-1}|\hat{\theta}_b - \theta|)),$$

where Φ denotes the cumulative distribution function of a standard Gaussian random variable. Consequently, we have

$$p_{\text{aggr}}^\theta > \alpha \quad \Leftrightarrow \quad \Phi^{-1}(1 - \alpha/4) > \text{median}_{b\in[B]}(\sqrt{N}\hat{\sigma}_{\infty,b}^{-1}|\hat{\theta}_b - \theta|),$$

which can be solved for feasible values of θ using root search. A full description of our method is presented in Algorithm 1.

Algorithm 1: Estimating the EATE from observational data on networks with spillover effects using plugin machine learning

---

Input: N unit-level observations $S_i = (W_i, C_i, X_i, Z_i, Y_i)$ from the model (1), network G, feature functions f_x^l, $l \in [r]$ and $f_z^l, l \in [t]$, corresponding dependency graph G_D, natural number K, natural number B, significance level $\alpha \in [0,1]$, machine learning algorithms.

Output: Estimator of the EATE θ_N^0 and a valid p-value and confidence interval for the two-sided test $H_0 : \theta_N^0 = 0$ vs $H_A : \theta_N^0 \neq 0$.

for $b \in [B]$ **do**
  Randomly split the index set $[N]$ into K sets $I_1, \ldots, I_K$ of approximately equal size.
  for $k \in [K]$ **do**
    Compute nuisance function estimators $\hat{g}_1^{I_k^c}$, $\hat{g}_0^{I_k^c}$, and $\hat{h}^{I_k^c}$ with machine learning algorithm and data from $\mathcal{S}_{I_k^c}$.
  end
  Compute point estimator of θ_N^0 according to (7), and call it $\hat{\theta}_b$.
  Estimate asymptotic variance of $\hat{\theta}_b$ using the bootstrap procedure described in Section 2.5 (or according to Theorem G.1 in Appendix G), and call it $\hat{\sigma}_{\infty,b}^2$.
  Compute p − value p_b for the two − sided test $H_0 : \theta_N^0 = 0$ vs. $H_A : \theta_N^0 \neq 0$ using $\hat{\theta}_b$, $\hat{\sigma}_{\infty,b}^2$, and asymptotic Gaussian approximation.
end

10 Compute $\hat{\theta} = \text{median}_{s\in[B]}(\hat{\theta}_b)$.

11 Compute aggregated p-value $p^0_{\text{aggr}} = 2\text{median}_{b\in[B]}p_b$.

12 Compute confidence interval according to (8), call it $\text{CI}(\hat{\theta})$.

13 Return $\hat{\theta}$, p^0_{aggr}, $\text{CI}(\hat{\theta})$.

Before we present our main theorem we mentioned in the construction of confidence intervals earlier, we present and discuss key assumptions. First, we require that products of machine learning errors decay fast enough, namely,

$$||h^0(C_i, Z_i) - \hat{h}^{I_k^c}(C_i, Z_i)||_{P,2} \cdot (||g_1^0(C_i, X_i) - \hat{g}_1^{I_k^c}(C_i, X_i)||_{P,2} + ||g_0^0(C_i, X_i) - \hat{g}_0^{I_k^c}(C_i, X_i)||_{P,2} + ||h^0(C_i, Z_i) - \hat{h}^{I_k^c}(C_i, Z_i)||_{P,2}) \ll N^{-\frac{1}{2}}$$

(see Assumption A4 in the appendix for more details). In particular, the individual error terms may vanish at a rate smaller than $N^{-1/4}$. This is achieved by many machine learning methods under suitable assumptions (see, for instance, Chernozhukov et al. [20]): ℓ_1-penalized and related methods in a variety of sparse models [39–44], forward selection in sparse models [45], L_2-boosting in sparse linear models [46], a class of regression trees and random forests [47], and neural networks [48]. Second, to ensure enough sparsity in the dependency structure of the data, the maximal degree $d_{\max}$ in the dependency graph is assumed to grow at most at the rate $d_{\max} = o(N^{1/4})$, which implies that the dependencies are not too far reaching. This assumption allows us to bound the Wasserstein distance of our (centered and scaled) treatment effect estimator to a standard Gaussian random variable using Stein's method [49].

Assumption 1. The maximal degree $d_{\max}$ of a node in the dependency graph satisfies $d_{\max} = o(N^{1/4})$.

Ogburn et al. [6] only required $d_{\max} = o(N^{1/2})$, but achieved a slower convergence rate of their treatment effect estimator. To recover the $\sqrt{N}$-rate, they require that $d_{\max}$ is bounded by a constant, meaning $d_{\max} = O(1)$.

Furthermore, we require that this dependency structure is not too strong moment-wise in the sense that the variance term given in the following assumption converges.

Assumption 2. Let $\{\mathcal{P}_N\}_{N\geq 1}$ be a sequence of sets of probability distributions P of the N units. There exists σ_∞^2, possibly depending on $P \in \mathcal{P}_N$, satisfying $0 < L \leq \sigma_\infty^2 \leq U < \infty$ with fixed constants L and U, such that for all $P \in \mathcal{P}_N$, we have

$$\lim_{N\to\infty}\left(\text{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \theta_i^0, \eta^0)\right) - \sigma_\infty^2\right) = 0, \tag{9}$$

where $\psi(S_i, \theta_i^0, \eta^0) = \varphi(S_i, \eta^0) - \theta_i^0$ is a centered version of φ.

Assuming bounded second moments, $\sum_{j=1}^{N}|\text{Cov}(\psi(S_i, \theta_i^0, \eta^0), \psi(S_j, \theta_i^0, \eta^0))|$ can be bounded, up to constants, by $d(i)$, where $d(i)$ denotes the degree of node i in the dependency graph. Consequently, we have

$$\text{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \theta_i^0, \eta^0)\right) \leq \gamma \cdot \frac{1}{N}\sum_{i=1}^{N}d(i), \tag{10}$$

where γ denotes some universal constant. Subsequently, we consider two special cases. First, if the maximal degree of the dependency graph is uniformly bounded by some constant D, we can bound (10) by the constant γD. Second, assume that the dependency graph has some nodes with finite degree: $d(i) \leq D$ for i in some set $S_{\max}^c$; the other nodes' degree $d(i)$ for $i \in S_{\max}$ is bounded by $d_{\max} = o(N^{1/4})$ with $|S_{\max}| \geq O(N/d_{\max}) = O(N^{3/4})$. Then, (10) is also of bounded order $O(1)$.

Theorem 2.2. (Asymptotic distribution of $\hat{\theta}$) *Assume* Assumptions 1 *and* 2 *as well as* A3 *and* A4 *stated in the appendix in* Section A. *Then, the estimator* $\hat{\theta}$ *of the EATE* θ_N^0 *given in* (7) *converges at the parametric rate,* $N^{-1/2}$*, and asymptotically follows a Gaussian distribution, namely,*

$$\sqrt{N}\sigma_\infty^{-1}(\hat{\theta} - \theta_N^0) \xrightarrow{d} \mathcal{N}(0,1) \quad (N \to \infty), \tag{11}$$

where σ_∞ *is characterized in* Assumption 2. *The convergence in* (11) *is in fact uniformly over the law* $P \in \mathcal{P}_N$ $(N \to \infty)$.

Please see Section E in the appendix for a proof of Theorem 2.2. The asymptotic variance σ_∞^2 in Theorem 2.2 can be consistently estimated using a bootstrap approach (see Section 2.5). Alternatively, it is possible to consistently estimate it using a plugin approach (see Theorem G.1 in the next Section G). However, empirical simulations have revealed that the bootstrap procedure described in the next section performs better.

Our estimator $\hat{\theta}$ is robust in two senses. First, it is $\sqrt{N}$-consistent and asymptotically normal if only the product property (9) of the machine learning estimators holds. Second, it can be shown that it remains consistent if either the propensity model or the outcome model is correctly specified. These properties are also called rate double robustness and model double robustness, respectively [50].

2.5 Bootstrap variance estimator

We use the residual bootstrap as follows to estimate the asymptotic variance. First, we use the estimated nuisance functions to compute the outcome regression residuals. More precisely, for $i \in [N]$, denote by $k(i)$ the index in $[K]$ specifying the partition unit i belongs to, namely, $i \in I_{k(i)}$. Then, we estimate the ε_Y's by $\hat{\varepsilon}_{Y_i} = \hat{\varepsilon}'_{Y_i} - \frac{1}{N}\sum_{j=1}^N \hat{\varepsilon}'_{Y_j}$, where $\hat{\varepsilon}'_{Y_i} = Y_i - W_i\hat{g}_1^{I_{k(i)}^c}(C_i, X_i) - (1 - W_i)\hat{g}_0^{I_{k(i)}^c}(C_i, X_i)$. Next, we sample confounders $\{C_i^*\}_{i\in[N]}$ with replacement from $\{C_i\}_{i\in[N]}$, and we sample $\hat{\varepsilon}_{Y_i}^*$ with replacement from $\{\hat{\varepsilon}_{Y_i}\}_{i\in[N]}$. These sampled covariates and error terms are now propagated through the SEM (1), i.e., we compute $Z_i^* = (f_z^1(C_{-i}^*, G), \ldots, f_z^t(C_{-i}^*, G))$, sample $W_i^* = \text{Bernoulli}(\hat{h}^{I_{k(i)}^c}(C_i^*, Z_i^*))$, compute $X_i^* = (f_x^1(W_{-i}^*, C_{-i}^*, G), \ldots, f_x^r(W_{-i}^*, C_{-i}^*, G))$, and build $Y_i^* = W_i^*\hat{g}_1^{I_{k(i)}^c}(C_i^*, X_i^*) - (1 - W_i^*)\hat{g}_0^{I_{k(i)}^c}(C_i^*, X_i^*) + \hat{\varepsilon}_{Y_i}^*$. Subsequently, we concatenate these values to obtain the bootstrap datapoints $S_i^* = (C_i^*, Z_i^*, W_i^*, X_i^*, Y_i^*), i \in [N]$. Then, we apply our treatment effect estimation procedure to the S_i^*'s to obtain a bootstrap estimator $\hat{\theta}^*$. This procedure is repeated R many times, and the bootstrap variance estimator is given by the empirical variance of the $\hat{\theta}_r^*$ over $r \in [R]$.

Theorem 2.3. *The bootstrap scheme described in* Section 2.5 *consistently estimates the asymptotic variance* (9) *under* Assumption A7 *stated in the Appendix*.

The proof of Theorem 2.3 can be found in Appendix F.

3 Empirical validation

We demonstrate our method in a simulation study and on a real-world dataset. In the simulation study, we validate the performance of our method on different network structures and compare it to two popular treatment effect estimators. Then, we investigate the effect of study time on exam performance in the Swiss StudentLife Study [27,28] taking into account the effect of social ties.

3.1 Simulation study

We investigate a fairly simple data-generating mechanism with 1-dimensional X-features and no Z-features. The X-interference effects a unit receives come from an interaction between treatments and control of its immediate neighbors in the network (we consider Erdős–Rényi and Watts–Strogatz). We compare the performance of our method to two popular off-the-shelf alternative schemes with respect to bias of the point estimator and coverage and length of respective two-sided confidence intervals: the Hájek estimator and an inverse propensity weighting estimator. Our aim is to see that these standard estimators may suffer in the presence of interference and to demonstrate that our easy-to-implement estimator overcomes their shortcomings.

We first describe the two competitors and then detail the simulation setting and present the results. Our code is available on GitHub (https://github.com/corinne-rahel/networkAIPW).

The **Hájek estimator** (denoted by "Hajek" in Figure 4) without incorporation of confounders [51] equals

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{\frac{1}{N}\sum_{j=1}^{N}W_i}+\frac{(1-W_i)Y_i}{\frac{1}{N}\sum_{j=1}^{N}(1-W_i)}\right).$$

The parametric convergence rate and asymptotic Gaussian distribution are preserved under X-spillover effects that equal the fraction of treated neighbors in a randomized experiment [52]. The **IPW estimator** [53] has been developed under SUTVA and uses observed confounding by creating a "pseudo-population" in which the treatment is independent of the confounders [54]. We compute it using sample splitting and cross-fitting according to

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\left(\frac{W_iY_i}{\hat{e}^{I_k^c}(C_i)}-\frac{(1-W_i)Y_i}{1-\hat{e}^{I_k^c}(C_i)}\right),$$

where $\hat{e}^{I_k^c}$ is the fitted propensity score obtained by regressing W_i on C_i on the data in $i \in \mathcal{S}_{I_k^c}$. In our simulation, $\hat{e}^{I_k^c}$ coincides with $\hat{h}^{I_k^c}$ because we consider no Z-features. We denote this estimator by "IPW" in Figure 4. These estimators are not designed for the interference structures we consider, but we would like to investigate the performance of these off-the-shelf and easy-to-implement estimators, also in comparison with our proposed method.

We investigate two network structures that govern our interference effects: Erdős–Rényi networks [56] and Watts–Strogatz networks [57]. Erdős–Rényi networks randomly form edges between units with a fixed probability and are a simple example of a random mathematical network model. These networks play an important role as a standard against which to compare more complicated models. Watts–Strogatz networks,

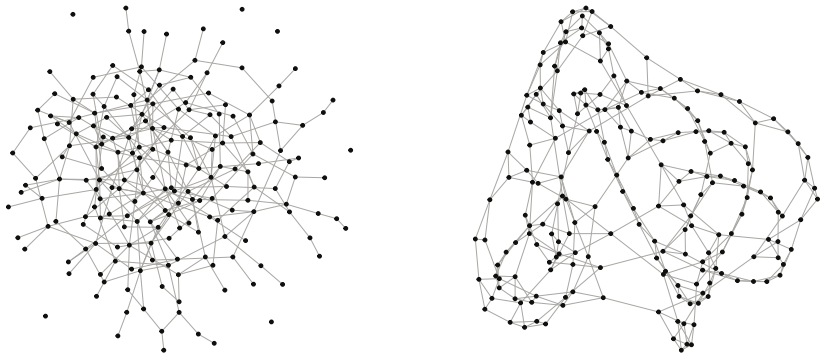

Figure 3: Different network structures on $N = 200$ units: Erdős–Rényi network (left) where two nodes are connected with probability $3/N$ (every node is connected to three other nodes in expectation); Watts–Strogatz network (right) with a rewiring probability of 0.05, a 1-dimensional ring-shaped starting lattice where each node is connected to two neighbors on both sides (i.e., every node is connected to four other nodes), no loops, and no multiple edges. The graphs are generated using the R-package `igraph` [55].

also called small-world networks, share two properties with many networks in the real world: a small average shortest path length and a large clustering coefficient. To construct such a network, the vertices are first arranged in a regular fashion and linked to a fixed number of their neighbors. Then, some randomly chosen edges are rewired with a constant rewiring probability. A representative of each network type is provided in Figure 3. For each of these two network types, we consider one case where the dependency in the network does not increase with N (denoted by "const" in Figure 4) and one where it increases with N (denoted by N^(1/15) in Figure 4).

The specific unit-level structural equations (1) we consider are as follows. For each unit $i \in [N]$, we sample independent and identically distributed confounders $C_i \sim \text{Unif}(0,1)$ from the uniform distribution. The treatment selections W_i are drawn from a Bernoulli distribution with arbitrarily chosen success probability $p_i = p_i(C_i) = 0.15\mathbb{1}_{C_i<0.33} + 0.5\mathbb{1}_{0.33\le C_i<0.66} + 0.85\mathbb{1}_{0.66\le C_i}$. Let $\alpha(i)$ denote the neighbors of unit i in the network (without i itself). Then, we let the 1-dimensional X-features X_i denote the shifted average number of neighbors assigned to treatment weighted by their confounder, namely,

$$X_i = \frac{1}{|\alpha(i)|} \sum_{j \in \alpha(i)} (\mathbb{1}_{W_j=1} - \mathbb{1}_{W_j=0}) C_j,$$

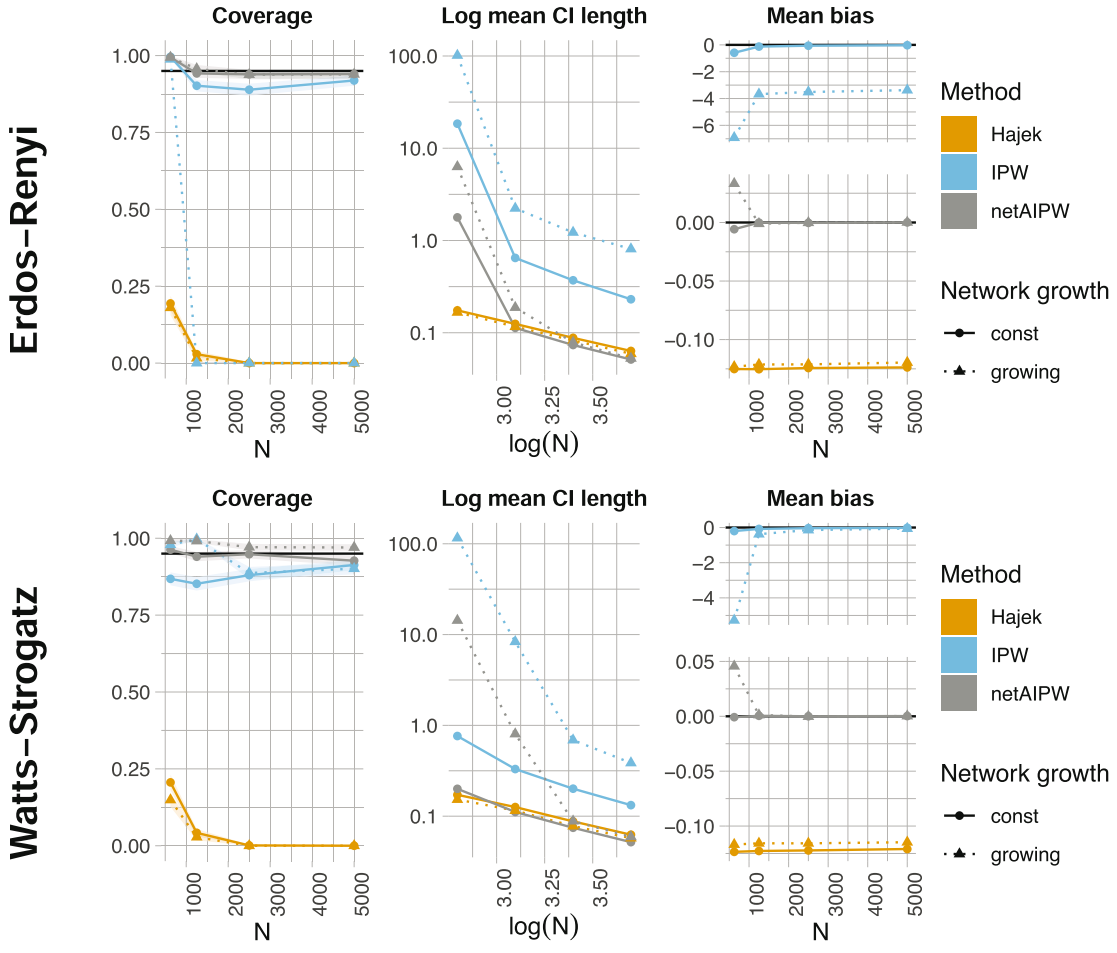

Figure 4: Coverage (fraction of times the true, and in general unknown, θ_N^0 was inside the confidence interval), log mean length of two-sided 95% confidence intervals for θ_N^0, and mean bias over 1,000 simulation runs for Erdős–Rényi and Watts–Strogatz networks of different complexities (Erdős–Rényi: expected degree 3 and $3N^{1/15}$ for "const" and "N ^ (1/15)," respectively; Watts–Strogatz: before rewiring, nodes have degree 4 and $4N^{1/15}$ for "const" and "N ^ (1/15)," respectively, and the rewiring probability is 0.05). We compare the performance of our method, netAIPW, with the Hájek and an IPW estimator, indicated by color. The variances of the competitors are empirical variances over the 1,000 repetitions, whereas we computed confidence intervals for netAIPW according to (8) with $B = 1$ and 300 bootstrap samples. The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1,000 simulation runs.

if $a(i)$ is non-empty, and 0 else. We do not consider Z-features. For real numbers x and c, we consider the arbitrary functions

$$g_1^0(x, c) = 1.5\mathbb{1}_{x\geq 0.5, c\geq -0.2, x<0.7} + 4\mathbb{1}_{c\geq -0.2, x\geq 0.7} + 0.5\mathbb{1}_{x\geq 0.5, c<-0.2} + 3.5\mathbb{1}_{x<0.5, c\geq -0.2} + 2.5\mathbb{1}_{x<0.5, c<-0.2}$$

and

$$g_0^0(x, c) = 0.5\mathbb{1}_{x\geq 0.4, c\geq 0.2} - 0.75\mathbb{1}_{x\geq 0.4, c<0.2} + 0.25\mathbb{1}_{x<0.4, c\geq 0.2} - 0.5\mathbb{1}_{x<0.4, c<0.2},$$

i.e., the functions g_1^0, g_0^0, and h^0 are step functions. For independent and identically distributed error terms $\varepsilon_{Y_i} \sim \text{Unif}(-\sqrt{0.12}/2, \sqrt{0.12}/2)$, we consider the outcomes $Y_i = W_i g_1^0(C_i, X_i) + (1 - W_i) g_0^0(C_i, X_i) + \varepsilon_{Y_i}$.

For the sample sizes $N = 625, 1{,}250, 2{,}500$, and $5{,}000$, we perform 1,000 simulation runs redrawing the data according to the SEM, and consider $B = 1$, $K = 5$, and $R = 300$ bootstrap samples to estimate the variance in Algorithm 1, i.e., we consider one split per generated dataset and consequently do not aggregate p-values in these simulations. However, the empirical analysis in Section 3.2 aggregates p-values over 100 datasplits. We estimate the nuisance functions by random forests consisting of 500 trees with a minimal node size of 5 and other default parameters using the R-package `ranger` [58]. To estimate the propensity score, we limit the depth of the trees to 2. Our results for the Erdős–Rényi and Watts–Strogatz networks are displayed in Figure 4. Two different panels are used to display the results for different ranges of the bias of the methods. For all network types and complexities, we observe the following. The IPW estimator incurs some bias as can be expected because it does not account for network spillover, and even under SUTVA, it is not Neyman orthogonal, which means we are not allowed to plug in machine learning estimators of nuisance functions. Furthermore, it is known to have a poor finite-sample performance due to estimated propensity scores $\hat{e}^{I_k^c}$ that may be close to 0 or 1. The Hájek estimator incurs some bias because it does not adjust for observed confounding and assumes a randomized treatment instead. The bias of our method (denoted by "netAIPW" in Figure 4) decreases as the sample size increases. As the dependency graph becomes more complex, our method requires more observations to achieve a small bias because the datasets $\mathcal{S}_{I_k^c}$ in (6), which are used to estimate the nuisance functions, are smaller in denser networks. In terms of coverage, the two competitors perform poorly, whereas our method guarantees coverage.

Simulation results involving spillover effects from second-degree neighbors and misspecified spillover effects are presented in Appendix C. Furthermore, for a Bernoulli(1/2) treatment assignment and with the "const" Watts-Strogatz setting presented in the main article, we found that the AIPW approach leads to variances that are of about a factor of 23 smaller than the ones obtained with IPW. This suggests that AIPW is helpful in reducing the variance of IPW even in the randomized case.

3.2 Empirical analysis: Swiss StudentLife study data

Subsequently, we estimate the causal effect of study time on academic success of university students with our newly developed estimator. We quantify this causal effect by the EATE that is the average of the difference in expected GPA of the final exam had a student studied much versus little, allowing for potential spillover effects from the student's friends on the student's study time. Among the factors that determine academic success are person-specific traits, such as intelligence [59], willingness to work hard [60], and socioeconomic background [61]. The Swiss StudentLife Study data [27,28] were collected to investigate the impact of various factors on academic achievement. It consists of observations from freshmen undergraduate students pursuing a degree in the natural sciences at a Swiss university. Instead of a university entrance test, these students had to pass a demanding examination after 1 year of studying. At several time points throughout this year, the students were asked to fill out questionnaires about their student life, social network, and well-being. The data consist of three cohorts of students. Cohort 1 was observed in 2016 and cohorts 2 and 3 in 2017. Importantly, for all three cohorts, the data contain friendship information among the students. We build the corresponding undirected network by drawing an edge between two students if at least one of them mentioned the other one as being a friend. We believe that spillover effects arise due to students interacting in this network, and

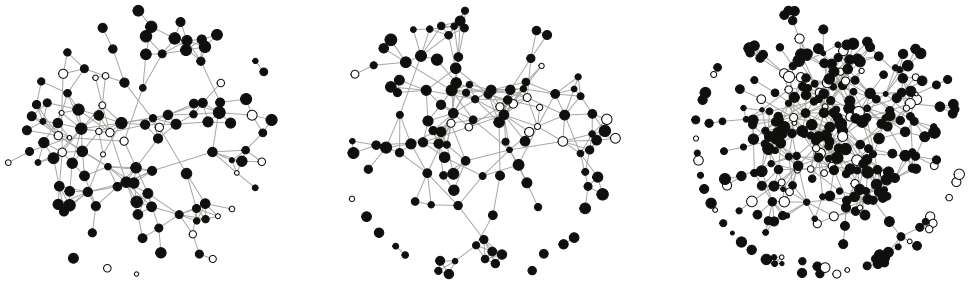

Figure 5: Friendship networks per cohort with black dots representing $W_i = 1$ and a weekly study time of at least 8 h, white for $W_i = 0$, and a weekly study time of less than 8 hours, and a bigger node size represents a higher GPA.

thus, we have to control for them when estimating the EATE described earlier. Figure 5 displays the resulting network consisting of the three cohorts.

GPA (Y_i) constitutes our outcome variable and represents the average grade of seven to nine exams, depending on study programs. It ranges from 1 to 6, with passing grades of 4 or higher. The average GPA in the data we used was 4.266 with a standard deviation of 0.872. The remaining variables were measured 5 to 6 months before the exam period and correspond to wave four of the Swiss StudentLife Study data. The self-reported number of hours spent studying per week during the semester (W_i) constitutes the treatment variable. It was dichotomized into studying many ($W_i = 1$) and few ($W_i = 0$) hours. We considered a setting where $W_i = 1$ corresponds to studying at least 8 hours per week, which is the 20% quantile, and one where $W_i = 1$ corresponds to studying at least 20 hours per week, which is the 80% quantile. We consider spillover effects from the friends of a student, which are a student's direct neighbors in the friendship network. We consider Z-spillover effects that account for the effect of befriended students' study motivation and stress variables on a student's treatment. We do not consider spillover effects on the outcome GPA (no X-features). The Z_i-spillover variable of a student i is a vector of length 6, where each entry corresponds to the average of the following six variables across the friends of the student: (a) study motivation, measured with the learning objectives subscale of the SELLMO-ST[3] [62]; (b) work avoidance, measured with the work avoidance subscale of the students version of the SELLMO-ST[3]; (c) the average of ten perceived stress items [63]; (d, e) two items specifically on exam related stress; and (f) whether one was perceived as clever by at least one other student. In addition to these network effects, we control on the unit level (C_i) for the just mentioned variables observed on an individual unit as well as the cohort number, gender, having Swiss nationality, speaking German, and the financial situation. From all the data of the three cohorts combined, we only considered individuals for whom all the mentioned variables, i.e., treatment, outcome, covariates, and Z-spillover variables, are observed. The final sample consisted of $N = 526$ individuals: 113 from cohort 1, 119 from cohort 2, and 294 from cohort 3. In our algorithm, we used $S = 1{,}000$ sample splits (from which we aggregate p-values as in (2.4)) with $K = 10$ groups each and random forests consisting of 5,000 trees to learn g_0^0, g_1^0, and h^0 whose leaf size was initially determined by fivefold cross-validation. Also, we used the variance estimator as in Appendix G that relies on fewer assumptions.

We estimated the EATE with two different definitions for $W_i = 1$, defined by a study time of either at least 8 or 20 h per week, corresponding to the 20 and 80% quantiles, respectively, and Table 1 displays the results. Table 1(a) displays our estimated EATE with $W_i = 1$ representing a weekly study time of at least 8 h. Our EATE estimator is positive and significant. On average, students received a 0.362 points higher GPA had they studied at least 8 h per week compared to studying less. Consequently, a significantly higher GPA can be achieved by studying more. If we apply the same procedure but exclude the Z-spillover covariates (no spillover), the EATE estimator is higher and also significant. Table 1(b) displays our results with $W_i = 1$ representing a weekly study

3 This is a scale to assess learning and achievement motivation, and the subscale consists of eight items measured on a 5-point Likert-scale from 1 ("completely disagree") to 5 ("completely agree").

Table 1: EATE and 95% confidence intervals for θ_N^0 for different settings with different control groups, namely, studying less than 8 (a) or less than 20 (b) hours per week

Spillover	EATE	95% CI for θ_N^0
(a) W_i = 1 if studied at least 8 h per week (20% quantile)		
Yes	0.362	[0.283, 0.442]
No	0.451	[0.364, 0.528]
(b) W_i = 1 if studied at least 20 h per week (80% quantile)		
Yes	0.078	[−0.096, 0.252]
No	0.163	[0.011, 0.311]

time of at least 20 hours. Our EATE estimator is positive but not significant anymore. Hence, our results suggest that GPA is not significantly higher had a student studied at least 20 h per week compared to studying less. Without spillover, the treatment effect is significant. In both cases in Table 1, the estimate of the EATE is higher under the assumption of no spillover effects, compared to the estimator that allows for possible Z-spillover effects. This potentially relevant difference highlights the importance of not *a priori* ruling out spillover effects. Overall, the model including spillover effects seems more realistic than the one excluding them. Finally, when interpreting the results, it is important to recall that study time captures the learning time during the semester. There is an additional 8-week lecture-free preparation period, and our study time does not reflect this preparation time. Consequently, our results only describe the EATE of study time during the semester on GPA.

4 Conclusion

Causal inference with observational data usually assumes independent units. However, having independent observations is often questionable, and so-called spillover effects among units are common in practice. Our aim was to develop point estimation and asymptotic inference for the expected average treatment effect (EATE) with observational data from a single (social) network. We would like to point out the hardness of this problem: we consider treatment effect estimation on data with increasing dependence among units, where the data-generating mechanism can be highly nonlinear and include confounders. We use an augmented inverse probability weighting (AIPW) principle and account for spillover effects that we capture by features, which are functions of the known network and the treatment and covariate vectors. There may be several features, and one feature may capture spillover effects from different units than another feature; these units might be direct neighbors to compute one feature and neighbors of neighbors to compute another feature. We consider the dependency graph to pose assumptions on these features in our asymptotic theory. Units may interact beyond their direct neighborhoods, interactions may become increasingly complex as the sample size increases, and we consider arbitrary networks. Using ideas of double machine learning [20], we develop a cross-fitting algorithm under interference that allows us to estimate the nuisance components of our model by arbitrary machine learning algorithms. Although we employ machine learning algorithms, our EATE estimator converges at the $\sqrt{N}$-rate and asymptotically follows a Gaussian distribution, which allows us to perform inference.

In a simulation study, we demonstrated that commonly employed methods for treatment effect estimation suffer from the presence of spillover effects, whereas our method could account for the complex dependence structures in the data so that the bias vanished with increasing sample size and coverage was guaranteed. In the Swiss StudentLife Study, we investigated the EATE of study time on the GPA of university examinations, accounting for spillover effects due to friendship relations. Omitting this spillover may lead to biased results due to spurious association.

In this work, we focused on estimating the EATE. Other effects may be estimated in a similar manner, for instance, the global average treatment effect (GATE) where all units are jointly intervened on. We develop an estimator of the GATE in Appendix H.

Acknowledgements: We thank the associate editor and reviewers for detailed and constructive comments. We also thank Leonard Henckel and Dominik Rothenhäusler for useful comments.

Funding information: CE and PB received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 786461), and M-LS received funding from the Swiss National Science Foundation (SNF) (Project No. 200021_172485). The Swiss StudentLife data collection was supported by Swiss National Science Foundation Grant 10001A 169965 and the rectorate of ETH Zurich.

Author contributions: All authors have accepted responsibility for the en4re content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The data and code used in the simula4on study are available on GitHub (https://github.com/corinne-rahel/networkAIPW). The Swiss StudentLife data analyzed in the empirical analysis is cited in the main text of our paper.

References

[1] Rubin D. Comment on: Randomization analysis of experimental data in the fisher randomization test by D. Basu. J Amer Stat Assoc. 1980;75:591–3.
[2] Perez-Heydrich C, Hudgens MG, Halloran E, Clemens JD, Ali M, Emch ME. Assessing effects of cholera vaccination in the presence of interference. Biometrics. 2014;70(3):731–41.
[3] Sävje F, Aronow PM, Hudgens MG. Average treatment effects in the presence of unknown interference. Ann Stat. 2021;49(2):673–701.
[4] Lee Y, Ogburn EL. Network dependence can lead to spurious associations and invalid inference. J Amer Stat Assoc. 2021;116(535):1060–74.
[5] Sobel ME. What do randomized studies of housing mobility demonstrate? J Amer Stat Assoc. 2006;101(476):1398–407.
[6] Ogburn EL, Sofrygin O, Diiiaz I, van der Laan MJ. Causal inference for social network data. J Amer Stat Assoc. 2022;0(0):1–15.
[7] Eckles D, Bakshy E. Bias and high-dimensional adjustment in observational studies of peer effects. J Amer Stat Assoc. 2021;116(534):507–17.
[8] Ogburn EL, VanderWeele TJ. Vaccines, contagion, and social networks. Ann Appl Stat. 2017;11(2):919–48.
[9] Hudgens MG, Halloran E. Toward causal inference with interference. J Amer Stat Assoc. 2008;103(482):832–42.
[10] Sofrygin O, van der Laan MJ. Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. J Causal Inference. 2017;5(1):1–35.
[11] Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669–88.
[12] Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. Stat Sci. 1990;5(4):465–72.
[13] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol. 1974;66(5):688–701.
[14] van der Laan M. Causal inference for a population of causally connected units. J Causal Inference. 2014;2(1):13–74.
[15] Manski CF. Identification of endogenous social effects: the reflection problem. Rev Econ Stud. 1993;60(3):531–42.
[16] Chin A. Regression adjustments for estimating the global treatment effect in experiments with interference. J Causal Inference. 2019;7(2):20180026.
[17] Cai J, De Janvry A, Sadoulet E. Social networks and the decision to insure. Amer Econ J Appl Econ. 2015;7(2):81–108.
[18] Leung M. Treatment and spillover effects under network interference. Rev Econ Stat. 2020;102(2):368–80.
[19] Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Amer Stat Assoc. 1995;90(429):106–21.

[20] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21(1):C1–68.
[21] Tchetgen Tchetgen EJ, Fulcher IR, Shpitser I. Auto-G-computation of causal effects on a network. J Amer Stat Assoc. 2021;116(534):833–44.
[22] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. Math Model. 1986;7(9):1393–512.
[23] Lauritzen SL, Richardson TS. Chain graph models and their causal interpretations. J R Stat Soc Ser B (Stat Methodol). 2002;64(3):321–48.
[24] van der Laan MJ, Rubin D. Targeted maximum likelihood learning. Int J Biostat. 2006;2(1).
[25] van der Laan MJ, Rose S. Targeted learning. Springer Series in Statistics. New York: Springer; 2011.
[26] van der Laan MJ, Rose S. Targeted learning in data science. Springer Series in Statistics. New York: Springer; 2018.
[27] Stadtfeld C, Vörös A, Elmer T, Boda Z, Raabe IJ. Integration in emerging social networks explains academic failure and success. Proc Nat Acad Sci. 2019;116(3):792–7.
[28] Vörös A, Boda Z, Elmer T, Hoffman M, Mepham K, Raabe IJ, et al. The Swiss StudentLife Study: Investigating the emergence of an undergraduate community through dynamic, multidimensional social network data. Soc Netw. 2021;65:71–84.
[29] Bühlmann P. Sieve bootstrap for time series. Bernoulli. 1997;3(2):123–48.
[30] Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Amer Stat Assoc. 1999;94(448):1096–120.
[31] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–72.
[32] Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. J Amer Stat Assoc. 1995;90(429):122–9.
[33] Lauritzen SL. Graphical models. Oxford statistical science series. Oxford: Clarendon Press; 1996.
[34] Pearl J. Graphs, causality, and structural equation models. Sociol Meth Res. 1998;27(2):226–84.
[35] Pearl J. Causality: Models, reasoning, and inference. 2nd ed. Cambridge: Cambridge University Press; 2009.
[36] Pearl J. An introduction to causal inference. Int J Biostat. 2010;6(2):7.
[37] Perković E, Textor J, Kalisch M, Maathuis MH. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. J Machine Learn Res. 2018;18(220):1–62.
[38] Meinshausen N, Meier L, Bühlmann P. *p*-Values for high-dimensional regression. J Amer Stat Assoc. 2009;104(488):1671–81.
[39] Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of lasso and dantzig selector. Ann Stat. 2009;37(4):1705–32.
[40] Bühlmann P, van de Geer S. Statistics for high-dimensional data: methods, theory and applications. Springer Series in Statistics. Heidelberg: Springer; 2011.
[41] Belloni A, Chernozhukov V, Wang L. Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika. 2011;98(4):791–806.
[42] Belloni A, Chernozhukov V. Ell-penalized quantile regression in high-dimensional sparse models. Ann Stat. 2011;39(1):82–130.
[43] Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica. 2012;80(6):2369–429.
[44] Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models. Bernoulli. 2013;19(2):521–47.
[45] Kozbur D. Analysis of testing-based forward model selection. Econometrica. 2020;88(5):2147–73.
[46] Luo Y, Spindler M. High-Dimensional L2 Boosting: Rate of Convergence; 2016. Preprint arXiv:1602.08927.
[47] Wager S, Walther G. Adaptive Concentration of Regression Trees, with Application to Random Forests; 2016. Preprint arXiv:1503.06388.
[48] Chen X, White H. Improved rates and asymptotic normality for nonparametric neural network estimators. IEEE Trans Inform Theory. 1999;45:682–91.
[49] Stein C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory. vol. 6. University of California Press; 1972. p. 583–603.
[50] Smucler E, Rotnitzky A, Robins JM. A unifying approach for doubly-robust ell1 regularized estimation of causal contrasts; 2019. Preprint arXiv:1904.03737.
[51] Hájek J. Comment on “An essay on the logical foundations of survey sampling, part one” by Basu. In: Godambe VP, Sprott DA, editors. Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston; 1971. p. 236.
[52] Li S, Wager S. Random graph asymptotics for treatment effect estimation under network interference. Ann Stat. 2022;50(4):2334–58.
[53] Rosenbaum PR. Model-based direct adjustment. J Amer Stat Assoc. 1987;82(398):387–94.
[54] Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003;71(4):1161–89.
[55] Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695. https://igraph.org.
[56] Erdős P, Rényi A. On random graphs I. Publicationes Mathematicae. 1959;6:290–7.
[57] Watts DJ, Strogatz S. Collective dynamics of ‘small-world’ networks. Nature. 1998;393:440–2.
[58] Wright MN, Ziegler A.A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77(1):1–17.

[59] Chamorro-Premuzic T, Furnham A. Personality, intelligence and approaches to learning as predictors of academic performance. Personality Individual Differ. 2008;44(7):1596–603.
[60] Los R, Schweinle A. The interaction between student motivation and the instructional environment on academic outcome: a hierarchical linear model. Soc Psychol Educat. 2019;22(2):471–500.
[61] Heckman JJ. Skill formation and the economics of investing in disadvantaged children. Science. 2006;312(5782):1900–2.
[62] Spinath B, Stiensmeier-Pelster J, Schoene C, Dickhäuser O. Skalen zur Erfassung der Lern- und Leistungsmotivation: SELLMO. Bern: Hogrefe Verlag; 2002.
[63] Cohen S, Williamson G. Perceived stress in a probability sample of the United States. In: Spacapam S, Oskamp S, editors. The Social Psychology of Health: Claremont Symposium on Applied Social Psychology. Newbury Park, CA: Sage; 1988.
[64] Lattimore T, Szepesvári C. Bandit algorithms. Cambridge: Cambridge University Press; 2020.
[65] Ugander J, Karrer B, Backstrom L, Kleinberg J. Graph cluster randomization: network exposure to multiple universes; 2013. Preprint arXiv:1305. 6979.
[66] Eckles D, Karrer B, Ugander J. Design and analysis of experiments in networks: reducing bias from interference. J Causal Infer. 2017;5(1):20150021.
[67] Robins G, Pattison P, Elliott P. Network models for social influence processes. Psychometrika. 2001;66(2):161–89.
[68] Daraganova G, Robins G. Autologistic actor attribute models. In: Lusher D, Koskinen J, Robins G, editors. Structural analysis in the social sciences. Cambridge: Cambridge University Press; 2012. p. 102–14.
[69] Snijders TAB. Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S, editors. Structural Analysis in the Social Sciences. Cambridge: Cambridge University Press; 2005. p. 215–47.
[70] Snijders TAB, van de Bunt GG, Steglich CEG. Introduction to stochastic actor-based models for network dynamics. Soc Netw. 2010;32(1):44–60.
[71] Steglich C, Snijders TAB, Pearson M. Dynamic networks and behavior: separating selection from influence. Sociol Methodol. 2010;40(1):329–93.
[72] Peters J, Janzing D, Schölkopf B. Elements of causal inference: Foundations and learning algorithms. Adaptive Comput Machine Learn. Cambridge, MA: The MIT Press; 2017.
[73] Maathuis M, Drton M, Lauritzen S, Wainwright M, editors. Handbook of graphical models. Handbooks of Modern Statistical Methods. Boca Raton, FL: Chapman & Hall/CRC; 2019.
[74] Chin A. Central limit theorems via Stein's method for randomized experiments under interference; 2018. Preprint arXiv:1804.03105.
[75] Ross N. Fundamentals of Stein's method. Probability surveys. 2011;8(none):210–93.
[76] Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981;9(6):1196–217.

Appendix A

A1 Assumptions and additional definitions

We consider the following notation. We denote by $[N]$ the set $\{1, 2, \ldots, N\}$. We add the probability law as a subscript to the probability operator $\mathbb{P}$ and the expectation operator $\mathbb{E}$ whenever we want to emphasize the corresponding dependence. We denote the $L^p(P)$-norm by $\|\cdot\|_{P,p}$ and the Euclidean or operator norm by $|\cdot|$, depending on the context. We implicitly assume that given expectations and conditional expectations exist. We denote by $\xrightarrow{d}$ convergence in distribution. The symbol $\perp\!\!\!\perp$ denotes independence of random variables.

We observe N units according to the structural equations (1) that are connected by an underlying network. For each unit $i \in [N]$, we concatenate $S_i = (W_i, C_i, X_i, Z_i, Y_i)$ that are relevant for unit i. For notational simplicity, we abbreviate $D_i = (C_i, X_i)$ and $U_i = (C_i, Z_i)$ for $i \in [N]$.

Let the number of sample splits $K \geq 2$ be a fixed integer independent of N. We assume that $N \geq K$ holds. Consider a partition $I_1, \ldots, I_K$ of $[N]$. We assume that all sets $I_1, \ldots, I_K$ are of equal cardinality n. We make this assumption for the sake of notational simplicity, but our results hold without it.

Let $\{\delta_N\}_{N \geq K}$ and $\{\Delta_N\}_{N \geq K}$ be two sequences of non-negative numbers that converge to 0 as $N \to \infty$. Let $\{\mathcal{P}_N\}_{N \geq 1}$ be a sequence of sets of probability distributions P of the N units.

For completeness, we recall the following two assumptions from the main text. Assumption A1 limits the growth rate of the maximal degree of a node in the dependency graph. Assumption A2 characterizes the asymptotic variance in Theorem G.1 as the limit of the population variance on the N units.

Assumption A1. The maximal degree $d_{\max}$ of a node in the dependency graph satisfies $d_{\max} = o(N^{1/4})$.

Assumption A2. Let $\{\mathcal{P}_N\}_{N \geq 1}$ be a sequence of sets of probability distributions P of the N units. There exists σ_∞^2, possibly depending on $P \in \mathcal{P}_N$, satisfying $0 < L \leq \sigma_\infty^2 \leq U < \infty$ with fixed constants L, U, such that for all $P \in \mathcal{P}_N$, we have

$$\lim_{N \to \infty} \left(\mathrm{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0) \right) - \sigma_\infty^2 \right) = 0, \tag{A1}$$

where $\psi(S_i, \theta_i^0, \eta^0) = \varphi(S_i, \eta^0) - \theta_i^0$ is a centered version of φ.

We make the following additional sets of assumptions. Assumption A3 recalls that we use the model (1) and specifies regularity assumptions on the involved random variables. Assumptions 3.2 and 3.3 ensure that the random variables are integrable enough. Assumption 3.4 ensures that the true underlying function h^0 of the treatment selection model is bounded away from 0 and 1, which allows us to divide by h^0 and $1 - h^0$.

Assumption A3. Let $p \geq 4$. For all N, all $i \in [N]$, all $P \in \mathcal{P}_N$, and all $k \in [K]$, we have the following:

3.1 The structural equations (1) hold, where the treatment $W_i \in \{0, 1\}$ is binary.

3.2 There is a finite real constant C_1 independent of P satisfying $\|W_i\|_{P,p} + \|C_i\|_{P,p} + \|X_i\|_{P,p} + \|Z_i\|_{P,p} + \|Y_i\|_{P,p} \leq C_1$.

3.3 There is a finite real constant C_2 independent of P such that we have $\|Y_i\|_{P,\infty} + \|g_1^0(D_i)\|_{P,\infty} + \|g_0^0(D_i)\|_{P,\infty} + \|h^0(U_i)\|_{P,\infty} \leq C_2$.

3.4 There is a finite real constant C_3 independent of P such that $P(C_3 \leq h^0(U_i) \leq 1 - C_3) = 1$ holds.

3.5 There is a finite real constant C_4 such that we have $|\theta_i^0| \leq C_4$.

Assumption A4 characterizes the realization set of the nuisance functions and the $N^{-1/2}$ convergence rate of products of the machine learning errors from estimating the nuisance functions g_1^0, g_0^0, and h^0.

Assumption A4. Consider the $p \geq 4$ from Assumption A3. For all $N \geq K$ and all $P \in \mathcal{P}_N$, consider a nuisance function realization set $\mathcal{T}$ such that the following conditions hold:

4.1 The set $\mathcal{T}$ consists of P-integrable functions $\eta = (g_1, g_0, h)$ whose pth moment exists and whose $\|\cdot\|_{P,\infty}$-norm is in fact uniformly bounded, and $\mathcal{T}$ contains $\eta^0 = (g_1^0, g_0^0, h^0)$. Furthermore, there is a finite real constant C_5 such that for all $i \in [N]$ and all elements $\eta = (g_0, g_1, h) \in \mathcal{T}$, we have

$$\|h^0(W_i) - h(W_i)\|_{P,2} \cdot (\|g_1^0(D_i) - g_1(D_i)\|_{P,2} + \|g_0^0(D_i) - g_0(D_i)\|_{P,2} + \|h^0(W_i) - h(W_i)\|_{P,2}) \leq \delta_N N^{-\frac{1}{2}}.$$

4.2 Assumption 3.4 also holds with h^0 replaced by h.

4.3 Let κ be the largest real number such that for all $i \in [N]$ and all $\eta \in \mathcal{T}$, we have

$$\|h^0(W_i) - h(W_i)\|_{P,2} + \|g_1^0(D_i) - g_1(D_i)\|_{P,2} + \|g_0^0(D_i) - g_0(D_i)\|_{P,2} \lesssim \sqrt{\delta_N} N^{-\kappa},$$

i.e., κ represents the slowest convergence rate of our machine learners. Then, there is a finite real constant C_6 such that $d_{\max} N^{-2\kappa} \leq C_6$ holds, where $d_{\max}$ denotes the maximal degree of the dependency graph.

4.4 For all $k \in [K]$, the nuisance parameter estimate $\hat{\eta}^{I_k^c} = \hat{\eta}^{I_k^c}(\mathcal{S}_{I_k^c})$ belongs to the nuisance function realization set $\mathcal{T}$ with P-probability no less than $1 - \Delta_N$.

Assumption A5 and A6 are only required to establish that our plugin estimator of the asymptotic variance is consistent in Appendix G. (However, please recall that we recommend using the bootstrap procedure presented in Section 2.5 unless the sample size is large). They are not required to establish the asymptotic Gaussian distribution of our plugin machine learning estimator. Assumption A5 characterizes the order of the minimal size of the sets $\mathcal{A}_d$ for $d \geq 0$. These sets are required to contain a sufficient number of units such that the degree-specific treatment effects θ_d^0 for $d \geq 0$ can be estimated at a fast enough rate. These estimators are required to give a consistent estimator of the asymptotic variance σ_∞^2.

Assumption A5. For $d \geq 0$, the order of $|\mathcal{A}_d|$ is at least $N^{3/4}$, denoted by $\Omega(N^{3/4})$ according to the Bachmann–Landau notation [64].

Assumption A6 specifies that all individual machine learning estimators of the nuisance functions converge at a rate faster than $N^{-1/4}$.

Assumption A6. The slowest convergence rate κ in Assumption 4.3 satisfies $\kappa \geq 1/4$.

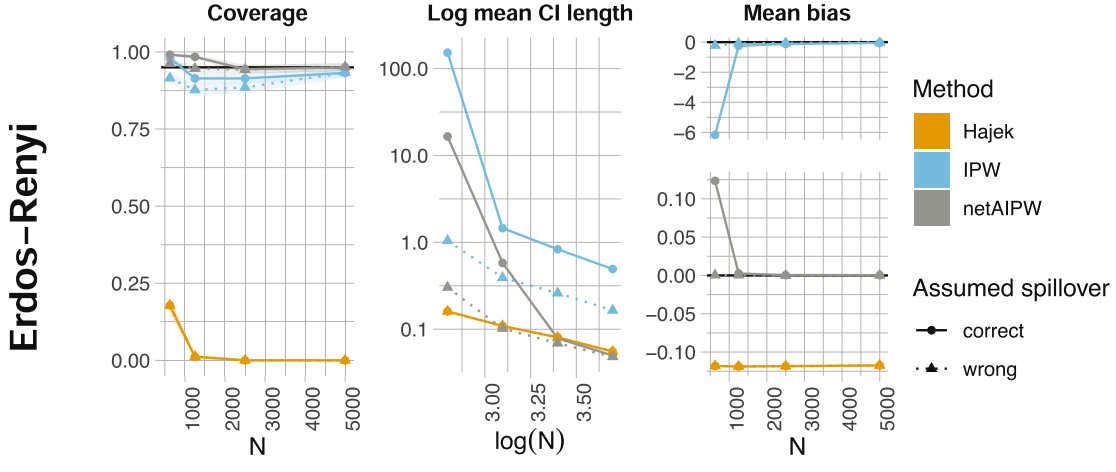

Figure A1: Coverage (fraction of times the true, and in general unknown, θ_N^0 was inside the confidence interval), log mean length of two-sided 95% confidence intervals for θ_N^0, and mean bias over 1,000 simulation runs for the "const" Erdős–Rényi network as in Section 3.1, except for the average degree of 2.5. We compare the performance of our method, netAIPW, with the Hájek and an IPW estimator, indicated by color, for correctly and incorrectly specifying the spillover effects from second-degree neighbors. The variance of the competitors are empirical variances over the 1,000 repetitions, whereas we computed confidence intervals for netAIPW according to (8) with $B = 1$ and 300 bootstrap samples. The shaded regions in the coverage plot represent 95% confidence bands with respect to the 1,000 simulation runs.

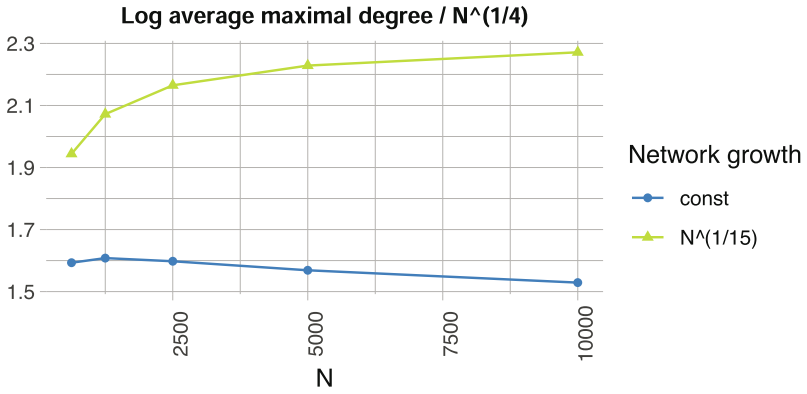

Figure A2: Simulated maximal degree of the dependency graph from second-degree spillover on Erdős–Rényi networks with an expected degree of either 2.5 ("const") or $2.5N^{1/15}$ ("N ^ (1/15)") divided by $N^{1/4}$, averaged over 1,000 simulation runs.

B Network effects in the social sciences

We consider models related to spillover effects. However, another notion of spillover effects has prevailed within the social science networks literature, namely, social influence effects. In this appendix, we describe social influence effects and how their modeling differs from our approach. Whereas spillover effects represent new covariates on the unit level that are built from variables of other units along network paths, social influence effects mostly concern effects that a specific variable A_j of neighboring units has on A_i of the ith unit. In the statistics literature, this is called contagion [65,66]. In the social sciences, there are two important models to investigate social influence/contagion processes: the autologistic actor attribute model (ALAAM; Robins et al. [67], Daraganova and Robins [68]) and the stochastic actor-oriented model (SAOM; Snijders [69], Snijders et al. [70], Steglich et al. [71]). Both models aim at estimating the degree to which a variable A_i of a focal individual is associated with the values of its neighbors' values of A. Whereas ALAAMs only considers cross-sectional data, SAOMs additionally allow estimating longitudinal social influence effects.

In contrast, the spillover features that we consider summarize variables from neighboring units. They represent a new variable that is used for the treatment or outcome regression models. For example, in our empirical analysis, we consider the spillover effect of study motivation of unit i's neighbors on the learning hours of unit i. We do not consider spillover from the learning hours of unit i's neighbors on unit i's own learning hours (i.e. social influence/contagion). Instead, we model such associations of the individual units' learning hours by constructing features from other variables and units that act as observed confounders. Moreover, we are not interested in estimating the effect as such of, say, other units' study motivation on the learning hours of unit i. However, this is possible with ALAAMs and SAOMs. We are not interested in estimating spillover as such, but we consider spillover as a tool to control for spurious associations due to the network structure to estimate treatment effects.

C Additional simulation results

First, we present simulation results involving spillover effects from second-degree neighbors and misspecified spillover effects. We consider the same data-generating mechanism and estimation framework as in Section 3.1 apart from the following change: the "neighborhood" $\alpha(i)$ defining X_i contains all second-degree neighbors of unit i, i.e., all units that are a distance 2 away from unit i in the network (neighbors of neighbors). For incorrectly specified spillover effects, we assumed that $\alpha(i)$ contains the direct neighbors of unit i instead. We consider an Erdős–Rényi network as in Section 3.1 except that the average degree of a unit is now 2.5. The results are displayed in Figure A1. Our method, netAIPW, does not seem to suffer much from the misspecified

spillover effects in terms of coverage, whereas the other methods do. In general, we observed that it is advantageous to include spillover effects even if they are not entirely correctly specified.

Next, we present networks and different kinds of spillover effects to show when Assumption A2 holds and when it fails to hold. We consider the same second-degree spillover effects and Erdős–Rényi network as mentioned earlier in this section. The "const" network has an expected degree of 2.5, and the "N^(1/15)" one has an expected degree of $2.5N^{1/15}$ in Figure A2. The maximal degree, divided by $N^{1/4}$ of the dependency graph of the "const" network, decreases with N, whereas the respective quantity increases with N for the non-constant-degree network, i.e., only the constant-degree network satisfies Assumption A2. The non-constant-degree network implies a dependency graph that is "too dense" to satisfy this assumption. We would like to remark that satisfying Assumption A2 is an interplay of the underlying network and the chosen spillover effects because they determine the dependency graph, and hence its maximal degree, together. A given network might lead to a dependency graph satisfying Assumption A2 with one kind of spillover effects (e.g., only from neighbors), whereas the same network might lead to a dependency graph violating this assumption with another kind of spillover effects (e.g., also including second-degree neighbors, i.e., neighbors of neighbors).

D Supplementary lemmata

In this section, we prove two results on conditional independence relationships of the variables from our model. We argue for the DAG of our model (1) and use graphical criteria [33–37,72,73]. We denote the direct causes of W_i by $\mathrm{pa}(W_i)$, the parents of W_i. Analogously, we denote the parents of Y_i by $\mathrm{pa}(Y_i)$; please see for instance Lauritzen [33]. We assume that $\mathrm{pa}(W_i)$ consists of C_i and the variables used to compute the spillover feature Z_i and that $\mathrm{pa}(Y_i)$ consists of W_i, C_i, and the variables used to compute the spillover feature X_i.

Lemma D.1. *Let $i \in [N]$, and let $C_j \notin \mathrm{pa}(Y_i)$. Then, we have $Y_i \perp\!\!\!\perp C_j | \mathrm{pa}(Y_i)$.*

Proof of Lemma D.1. The parents of Y_i are a valid adjustment set [35]. Because Y_i has no descendants, the claim follows. □

Lemma D.2. *Let $i \in [N]$, and let $C_j \notin \mathrm{pa}(W_i)$. Then, we have $W_i \perp\!\!\!\perp C_j | \mathrm{pa}(W_i)$. Furthermore, for $j \neq i$, we have $W_i \perp\!\!\!\perp W_j | \mathrm{pa}(W_i)$.*

Proof of Lemma D.2. The parents of W_i are a valid adjustment set [35]. The treatment variable W_i has no descendants apart from outcomes Y, which are colliders on any path from W_i to C_j or W_j, and thus, the empty set blocks these paths. Consequently, the two claims follow. □

E Proof of Theorem 2.2

Proof of Lemma 2.1. Let $i \in [N]$. We have

$$\mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)] = \mathbb{E}\left[\frac{W_i}{h^0(U_i)}(Y_i - g_1^0(D_i))\right] - \mathbb{E}\left[\frac{1 - W_i}{1 - h^0(U_i)}(Y_i - g_0^0(D_i))\right].$$

We have

$$\begin{aligned}
\mathbb{E}\left[\frac{W_i}{h^0(U_i)}(Y_i - g_1^0(D_i))\right] &= \mathbb{E}\left[\frac{W_i}{h^0(U_i)}(\mathbb{E}[Y_i | \mathrm{pa}(Y_i) \cup \mathrm{pa}(W_i)] - g_1^0(D_i))\right] \\
&= \mathbb{E}\left[\frac{1}{h^0(U_i)}\mathbb{E}[W_i Y_i - W_i g_1^0(D_i) | \mathrm{pa}(Y_i)]\right] \\
&= \mathbb{E}\left[\frac{W_i}{h^0(U_i)}\mathbb{E}[\varepsilon_{Y_i} | \mathrm{pa}(Y_i)]\right] \\
&= 0
\end{aligned} \tag{A2}$$

due to Lemma D.1 and because $\mathbb{E}[\varepsilon_{Y_i}|\mathrm{pa}(Y_i)] = 0$ holds by assumption. Analogous computations for $\mathbb{E}[(1 - W_i)/(1 - h^0(U_i))(Y_i - g_0^0(D_i))]$ conclude the proof. □

The following lemma shows that the score function φ is Neyman orthogonal in the sense that its Gateaux derivative vanishes [20].

Lemma E.1. (Neyman orthogonality) *Assume that the assumptions of* Theorem 2.2 *hold. Let* $\eta \in \mathcal{T}$, *and let* $i \in [N]$. *Then, we have*

$$\frac{\partial}{\partial r}\Big|_{r=0} \mathbb{E}[\varphi(S_i, \eta^0 + r(\eta - \eta^0))] = 0.$$

Proof of Lemma E.1. Let $r \in (0,1)$, let $i \in [N]$, and let $\eta \in \mathcal{T}$. Then, we have

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}[\varphi(S_i, \eta^0 + r(\eta - \eta^0))] \\
&= \frac{\partial}{\partial r}\mathbb{E}\Bigg[g_1^0(D_i) - g_0^0(D_i) + r(g_1(D_i) - g_0(D_i) - g_1^0(D_i) + g_0^0(D_i)) \\
&\quad + \frac{W_i}{h^0(U_i) + r(h(U_i) - h^0(U_i))}(Y_i - g_1^0(D_i) - r(g_1(D_i) - g_1^0(D_i))) \\
&\quad - \frac{1 - W_i}{1 - h^0(U_i) - r(h(U_i) - h^0(U_i))}(Y_i - g_0^0(D_i) - r(g_0(D_i) - g_0^0(D_i)))\Bigg] \\
&= \mathbb{E}\Bigg[(g_1(D_i) - g_0(D_i)) - (g_1^0(D_i) - g_0^0(D_i)) \\
&\quad + \frac{W_i}{(h^0(U_i) + r(h(U_i) - h^0(U_i)))^2}(-(g_1(D_i) - g_1^0(D_i))(h^0(U_i) + r(h(U_i) - h^0(U_i))) \\
&\quad - (Y_i - g_1^0(D_i) - r(g_1(D_i) - g_1^0(D_i)))(h(U_i) - h^0(U_i))) \\
&\quad - \frac{1 - W_i}{(1 - h^0(U_i) - r(h(U_i) - h^0(U_i)))^2}(-(g_0(D_i) - g_0^0(D_i))(1 - h^0(U_i) - r(h(U_i) - h^0(U_i))) \\
&\quad + (Y_i - g_0^0(D_i) - r(g_0(D_i) - g_0^0(D_i)))(h(U_i) - h^0(U_i)))\Bigg].
\end{aligned} \tag{A3}$$

We evaluate this expression at $r = 0$ and obtain

$$\begin{aligned}
&\frac{\partial}{\partial r}\Big|_{r=0}\mathbb{E}[\varphi(S_i, \eta^0 + r(\eta - \eta^0))] \\
&= \mathbb{E}\Bigg[(g_1(D_i) - g_0(D_i)) - (g_1^0(D_i) - g_0^0(D_i)) \\
&\quad - \left(1 + \frac{\varepsilon_{W_i}}{h^0(U_i)}\right)(g_1(D_i) - g_1^0(D_i)) - \frac{W_i}{(h^0(U_i))^2}(Y_i - g_1^0(D_i))(h(U_i) - h^0(U_i)) \\
&\quad + \left(1 - \frac{\varepsilon_{W_i}}{1 - h^0(U_i)}\right)(g_0(D_i) - g_0^0(D_i)) - \frac{1 - W_i}{(1 - h^0(U_i))^2}(Y_i - g_0^0(D_i))(h(U_i) - h^0(U_i))\Bigg] = 0
\end{aligned}$$

due to (A2) and because

$$\begin{aligned}\mathbb{E}\left[\frac{\varepsilon_{W_i}}{h^0(U_i)}(g_1(D_i) - g_1^0(D_i))\right] &= \mathbb{E}\left[(\mathbb{E}[W_i|\mathrm{pa}(W_i) \cup \mathrm{pa}(Y_i)] - h^0(U_i))\frac{1}{h^0(U_i)}(g_1(D_i) - g_1^0(D_i))\right] \\ &= \mathbb{E}\left[\mathbb{E}[W_i - h^0(U_i)|\mathrm{pa}(W_i)]\frac{1}{h^0(U_i)}(g_1(D_i) - g_1^0(D_i))\right] \\ &= \mathbb{E}\left[\mathbb{E}[\varepsilon_{W_i}|\mathrm{pa}(W_i)]\frac{1}{h^0(U_i)}(g_1(D_i) - g_1^0(D_i))\right] \\ &= 0\end{aligned}$$

holds due to Lemma D.2 and because we assumed $\mathbb{E}[\varepsilon_{W_i}|\mathrm{pa}(W_i)] = 0$, and similarly, for $\mathbb{E}[\varepsilon_{W_i}/(1 - h^0(U_i))(g_0(D_i) - g_0^0(D_i))]$. □

The following lemma bounds the second directional derivative of the score function. Its proof uses that products of the errors of the machine learners are of a smaller order than $N^{-1/2}$.

Lemma E.2. (Product property) *Assume the assumptions of* Theorem 2.2 *hold. Let* $r \in (0,1)$, *let* $\eta \in \mathcal{T}$, *and let* $i \in [N]$. *Then, we have*

$$\left|\frac{\partial^2}{\partial r^2}\mathbb{E}[\varphi(S_i, \eta^0 + r(\eta - \eta^0))]\right| \lesssim \delta_N N^{-\frac{1}{2}}.$$

Proof of Lemma E.2. We use the first directional derivative we derived in (A3) to compute the second directional derivative

$$\begin{aligned}&\frac{\partial^2}{\partial r^2}\mathbb{E}[\varphi(S_i, \eta^0 + r(\eta - \eta^0))] \\ &= 2\mathbb{E}\Bigg[\frac{W_i}{(h^0(U_i) + r(h(U_i) - h^0(U_i)))^4}((g_1(D_i) - g_1^0(D_i))(h^0(U_i) + r(h(U_i) - h^0(U_i))) \\ &\quad + (Y_i - g_1^0(D_i) - r(g_1(D_i) - g_1^0(D_i)))(h(U_i) - h^0(U_i))) \\ &\quad \cdot (h^0(U_i) + r(h(U_i) - h^0(U_i)))(h(U_i) - h^0(U_i))\Bigg] \\ &\quad + 2\mathbb{E}\Bigg[\frac{1 - W_i}{(1 - h^0(U_i) - r(h(U_i) - h^0(U_i)))^4}((g_0(D_i) - g_0^0(D_i))(1 - h^0(U_i) - r(h(U_i) - h^0(U_i))) \\ &\quad - (Y_i - g_0^0(D_i) - r(g_0(D_i) - g_0^0(D_i)))(h(U_i) - h^0(U_i))) \\ &\quad \cdot (1 - h^0(U_i) - r(h(U_i) - h^0(U_i)))(h(U_i) - h^0(U_i))\Bigg].\end{aligned}$$

Due to Hölder's inequality and Assumptions 3.1, 3.3, 3.4, and 4.1, we have

$$\begin{aligned}&\left|\frac{\partial^2}{\partial r^2}\mathbb{E}[\varphi(S_i, \eta^0 + r(\eta - \eta^0))]\right| \\ &\quad \lesssim (\|g_1(D_i) - g_1^0(D_i)\|_{P,2} + \|h(U_i) - h^0(U_i)\|_{P,2})\|h(U_i) - h^0(U_i)\|_{P,2} \\ &\qquad + (\|g_0(D_i) - g_0^0(D_i)\|_{P,2} + \|h(U_i) - h^0(U_i)\|_{P,2})\|h(U_i) - h^0(U_i)\|_{P,2}.\end{aligned}$$

Due to Assumption 4.1, both summands mentioned earlier are bounded by $\delta_N N^{-1/2}$, and hence, we conclude the proof. □

The following lemma describes how we apply Stein's method [74] to obtain the asymptotic Gaussian distribution of our estimator although the data are highly dependent.

Lemma E.3. (Asymptotic distribution with Stein's method) *Assume the assumptions of* Theorem 2.2 *hold. Denote by*

$$\sigma_N^2 = \mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \theta_i^0, \eta^0)\right).$$

Observe that by Assumption A2, *we have* $\lim_{N\to\infty}(\sigma_N^2 - \sigma_\infty^2) = 0$. *Then, we have*

$$\sigma_N^{-1} \cdot \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \theta_i^0, \eta^0) \xrightarrow{d} \mathcal{N}(0,1).$$

Proof of Lemma E.3. According to Lemma 2.1, we have $\mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)] = 0$. According to Assumption A3, the fourth moment of $\psi(S_i, \theta_i^0, \eta^0)$ exists for all $i \in [N]$ and is uniformly bounded over $i \in [N]$. Recall that we denote by $d_{\max}$ the maximal degree in the dependency graph on $S_i, i \in [N]$. Based on Ross [75, Theorem 3.6], we can thus bound the Wasserstein distance of $\sigma_N^{-1} \cdot \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \theta_i^0, \eta^0)$ to $\mathcal{N}(0,1)$ as follows: there exist finite real constants c_1 and c_2 such that we have

$$\begin{aligned} & d_{\mathcal{W}}\left(\sigma_N^{-1} \cdot \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \theta_i^0, \eta^0)\right) \\ & \leq c_1 \cdot \frac{d_{\max}^{3/2}}{\sigma_N^2} \cdot \sqrt{\sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{1}{\sqrt{N}}\psi(S_i, \theta_i^0, \eta^0)\right)^4\right]} + c_2 \cdot \frac{d_{\max}^2}{\sigma_N^3} \cdot \sum_{i=1}^{N}\mathbb{E}\left[\left|\frac{1}{\sqrt{N}}\psi(S_i, \theta_i^0, \eta^0)\right|^3\right] \\ & = c_1 \cdot \frac{d_{\max}^{3/2} \cdot \frac{1}{\sqrt{N}}}{\sigma_N^2} \cdot \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\psi^4(S_i, \theta_i^0, \eta^0)]} + c_2 \cdot \frac{d_{\max}^2 \cdot \frac{1}{\sqrt{N}}}{\sigma_N^3} \cdot \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[|\psi(S_i, \theta_i^0, \eta^0)|^3]. \end{aligned} \tag{A4}$$

By assumption, we have $d_{\max} = o(N^{1/4})$. Thus, we have $d_{\max}^{3/2} \cdot \frac{1}{\sqrt{N}} = o(N^{-1/8})$ and $d_{\max}^2 \cdot \frac{1}{\sqrt{N}} = o(1)$. Because the terms $\mathbb{E}[\psi^4(S_i, \theta_i^0, \eta^0)]$ and $\mathbb{E}[|\psi(S_i, \theta_i^0, \eta^0)|^3]$ are uniformly bounded over all $i \in [N]$ and because $\sigma_N \to \sigma_\infty$ as $N \to \infty$ according to Assumption A2, the Wasserstein distance in (A4) is of order $o(1)$. Consequently, we infer the statement of the lemma. □

Lemma E.4. (Vanishing covariance due to sparse dependency graph) *Assume the assumptions of* Theorem 2.2 *hold. Let* $k \in [K]$, *and recall that* $n = |I_k|$ *holds. Then, we have*

$$\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\varphi(S_i, \hat{\eta}^{I_k^c}) - \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]) - \frac{1}{\sqrt{n}}\sum_{i\in I_k}(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)])\right| = o_P(1).$$

Proof of Lemma E.4. Let $k \in [K]$. We have

$$\begin{aligned} & \mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\varphi(S_i, \hat{\eta}^{I_k^c}) - \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]) - \frac{1}{\sqrt{n}}\sum_{i\in I_k}(\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)])\right|^2 \middle| \mathcal{S}_{I_k^c}\right] \\ & = \frac{1}{n}\sum_{i\in I_k}\mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))^2|\mathcal{S}_{I_k^c}] - \frac{1}{n}\sum_{i\in I_k}\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)|\mathcal{S}_{I_k^c}]^2 \\ & \quad + \frac{1}{n}\sum_{i,j\in I_k, i\neq j}\mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))(\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0))|\mathcal{S}_{I_k^c}] \\ & \quad - \frac{1}{n}\sum_{i,j\in I_k, i\neq j}\mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)|\mathcal{S}_{I_k^c}]\mathbb{E}[\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)|\mathcal{S}_{I_k^c}]. \end{aligned} \tag{A5}$$

Let $i \in [N]$. The nuisance parameter estimator $\hat{\eta}^{I_k^c}$ belongs to $\mathcal{T}$ with P-probability at least $1 - \Delta_N$ by Assumption 4.4. Therefore, with P-probability at least $1 - \Delta_N = 1 - o(1)$, we have

$$\begin{aligned}
&\sqrt{\mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))^2 | \mathcal{S}_{I_k^c}]} \\
&\lesssim \sup_{\eta \in \mathcal{T}} \Bigg\| -g_1^0(D_i) + g_1(D_i) + g_0^0(D_i) - g_0(D_i) + \frac{W_i}{h^0(U_i)} \varepsilon_{Y_i} \\
&\quad - \frac{W_i}{h(U_i)} (g_1^0(D_i) - g_1(D_i) + \varepsilon_{Y_i}) - \frac{1 - W_i}{1 - h^0(U_i)} \varepsilon_{Y_i} + \frac{1 - W_i}{1 - h(U_i)} (g_0^0(D_i) - g_0(D_i) + \varepsilon_{Y_i}) \Bigg\|_{P,2} \\
&\lesssim \sup_{\eta \in \mathcal{T}} \|g_1^0(D_i) - g_1(D_i)\|_{P,2} + \sup_{\eta \in \mathcal{T}} \|g_0^0(D_i) - g_0(D_i)\|_{P,2} \\
&\quad + \sup_{\eta \in \mathcal{T}} \left\| \frac{h(U_i) - h^0(U_i)}{h^0(U_i) h(U_i)} W_i \varepsilon_{Y_i} \right\|_{P,2} + \sup_{\eta \in \mathcal{T}} \left\| \frac{W_i}{h(U_i)} (g_1^0(D_i) - g_1(D_i)) \right\|_{P,2} \\
&\quad + \sup_{\eta \in \mathcal{T}} \left\| \frac{h^0(U_i) - h(U_i)}{(1 - h^0(U_i))(1 - h(U_i))} (1 - W_i) \varepsilon_{Y_i} \right\|_{P,2} + \sup_{\eta \in \mathcal{T}} \left\| \frac{1 - W_i}{1 - h(U_i)} (g_0^0(D_i) - g_0(D_i)) \right\|_{P,2}.
\end{aligned}$$

Assumptions 3.1, 3.3, 3.4, and 4.2 bound the terms $\|W_i \varepsilon_{Y_i} / (h^0(U_i) h(U_i))\|_{P,\infty}$, $\|W_i / h(U_i)\|_{P,\infty}$, $\|(1 - W_i) \varepsilon_{Y_i} / ((1 - h^0(U_i))(1 - h(U_i)))\|_{P,\infty}$, and $\|(1 - W_i)/(1 - h(U_i))\|_{P,\infty}$. Assumption 4.3 specifies that the error terms $\|h^0(W_i) - h(W_i)\|_{P,2}$, $\|g_1^0(D_i) - g_1(D_i)\|_{P,2}$, and $\|g_0^0(D_i) - g_0(D_i)\|_{P,2}$ are upper bounded by $\sqrt{\delta_N} N^{-\kappa}$. Due to Hölder's inequality, we infer

$$\sqrt{\mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))^2 | \mathcal{S}_{I_k^c}]} \lesssim \sqrt{\delta_N} N^{-\kappa}, \tag{A6}$$

with P-probability at least $1 - \Delta_N$.

Subsequently, we bound the summands in (A5). Due to (A6), we have

$$\frac{1}{n} \sum_{i \in I_k} \mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))^2 | \mathcal{S}_{I_k^c}] - \frac{1}{n} \sum_{i \in I_k} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}]^2 \lesssim \delta_N N^{-2\kappa}$$

with P-probability at least $1 - \Delta_N$. Observe that we have

$$\begin{aligned}
&\frac{1}{n} \sum_{i,j \in I_k, i \neq j} \mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))(\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)) | \mathcal{S}_{I_k^c}] \\
&\quad - \frac{1}{n} \sum_{i,j \in I_k, i \neq j} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}] \mathbb{E}[\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}] \\
&= \frac{1}{n} \sum_{i,j \in I_k, i \neq j} \mathrm{Cov}(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0), \varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}) \\
&= \frac{1}{n} \sum_{i,j \in I_k, \{i,j\} \in E_D} \mathrm{Cov}(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0), \varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}),
\end{aligned}$$

where E_D denotes the edge set of the dependency graph, because the S_i with $i \in I_k$ are independent of data in $\mathcal{S}_{I_k^c}$ and because, given $\mathcal{S}_{I_k^c}$, $\varphi(S_i \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0)$ and $\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)$ are uncorrelated if there is no edge between i and j in the dependency graph. In the dependency graph, each node has a maximal degree of $d_{\max}$. Thus, there are at most $1/2 \cdot N \cdot d_{\max}$ many edges in E_D. With P-probability at least $1 - \Delta_N$, the term

$$\mathrm{Cov}(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0), \varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c})$$

can be bounded by $\delta_N N^{-2\kappa}$ up to constants for all i and j due to (A6). Therefore, with P-probability at least $1 - \Delta_N$, we have

$$\begin{aligned}
&\frac{1}{n} \sum_{i,j \in I_k, i \neq j} \mathbb{E}[(\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0))(\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0)) | \mathcal{S}_{I_k^c}] \\
&\quad - \frac{1}{n} \sum_{i,j \in I_k, i \neq j} \mathbb{E}[\varphi(S_i, \hat{\eta}^{I_k^c}) - \varphi(S_i, \eta^0) | \mathcal{S}_{I_k^c}] \mathbb{E}[\varphi(S_j, \hat{\eta}^{I_k^c}) - \varphi(S_j, \eta^0) | \mathcal{S}_{I_k^c}] \\
&\lesssim \delta_N d_{\max} N^{-2\kappa} \\
&\lesssim \delta_N,
\end{aligned}$$

where the last bound holds due to Assumption 4.3. Consequently, we have

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\varphi(S_i,\hat{\eta}^{I_k^c})-\mathbb{E}[\varphi(S_i,\hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}])-\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\varphi(S_i,\eta^0)-\mathbb{E}[\varphi(S_i,\eta^0)])\right|^2\middle|\mathcal{S}_{I_k^c}\right]\lesssim\delta_N,$$

with P-probability at least $1-\Delta_N$, and we infer the statement of the lemma based on Chernozhukov et al. [20, Lemma 6.1]. □

Lemma E.5. (Taylor expansion) *Assume the assumptions of* Theorem 2.2 *hold. Let* $k\in[K]$. *We have*

$$\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\mathbb{E}[\varphi(S_i,\hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]-\mathbb{E}[\varphi(S_i,\eta^0)])\right|=o_P(1).$$

Proof of Lemma E.5. Let $k\in[K]$. For $r\in[0,1]$, let us define the function

$$f_k(r)=\frac{1}{n}\sum_{i\in I_k}(\mathbb{E}[\varphi(S_i,\eta^0+r(\hat{\eta}^{I_k^c}-\eta^0))|\mathcal{S}_{I_k^c}]-\mathbb{E}[\varphi(S_i,\eta^0)].$$

We have

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\mathbb{E}[\varphi(S_i,\hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]-\mathbb{E}[\varphi(S_i,\eta^0)])\right|\ \mathcal{S}_{I_k^c}\right]\\&=\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\mathbb{E}[\varphi(S_i,\hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]-\mathbb{E}[\varphi(S_i,\eta^0)])\right|\\&=\sqrt{n}|f_k(1)|.\end{aligned}$$

We apply a Taylor expansion to $f_k(1)$ at 0 and obtain

$$f_k(1)=f_k(0)+f_k'(0)+\frac{1}{2}f_k''(\tilde{r}),$$

for some $\tilde{r}\in(0,1)$. Thus, we have

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\mathbb{E}[\varphi(S_i,\hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]-\mathbb{E}[\varphi(S_i,\eta^0)])\right|\middle|\mathcal{S}_{I_k^c}\right]\lesssim\sqrt{n}\left(|f_k(0)|+|f_k'(0)|+\sup_{r\in(0,1)}\frac{1}{2}|f_k''(r)|\right).$$

Due to the definition of f_k, we have $f_k(0)=0$. Due to Neyman orthogonality that we established in Lemma E.1, we have $f_k'(0)=0$. Due to the product property that we established in Lemma E.2, we have $\sup_{r\in(0,1)}\frac{1}{2}|f_k''(r)|\lesssim\delta_N N^{-1/2}$ with P-probability at least $1-\Delta_N$ because $\hat{\eta}^{I_k^c}$ belongs to $\mathcal{T}$ with P-probability at least $1-\Delta_N$. Consequently, we have

$$\mathbb{E}\left[\left|\frac{1}{\sqrt{n}}\sum_{i\in I_k}(\mathbb{E}[\varphi(S_i,\hat{\eta}^{I_k^c})|\mathcal{S}_{I_k^c}]-\mathbb{E}[\varphi(S_i,\eta^0)])\right|\middle|\mathcal{S}_{I_k^c}\right]\lesssim\delta_N,$$

with P-probability at least $1-\Delta_N$. We infer the statement of the lemma based on Chernozhukov et al. [20, Lemma 6.1]. □

Proof of Theorem 2.2. We have

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta_N^0) &= \sqrt{N} \cdot \frac{1}{nK} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) \\ &= \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n}} \sum_{i \in I_k} (\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \psi(S_i, \theta_i^0, \eta^0)) + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0)\end{aligned}$$

because the disjoint sets I_k are of equal size n, so that we have $N = nK$. Let $k \in [K]$. We have

$$\begin{aligned}&\left| \frac{1}{\sqrt{n}} \sum_{i \in I_k} (\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \psi(S_i, \theta_i^0, \eta^0)) \right| \\ &\leq \left| \frac{1}{\sqrt{n}} \sum_{i \in I_k} (\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \mathbb{E}[\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}]) - \frac{1}{\sqrt{n}} \sum_{i \in I_k} (\psi(S_i, \theta_i^0, \eta^0) - \mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)]) \right| \\ &\quad + \left| \frac{1}{\sqrt{n}} \sum_{i \in I_k} (\mathbb{E}[\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) | \mathcal{S}_{I_k^c}] - \mathbb{E}[\psi(S_i, \theta_i^0, \eta^0)]) \right| \\ &= o_P(1),\end{aligned}$$

due to Hölder's inequality and Lemmas E.4 and E.5. Because K is a constant independent of N, we have

$$\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{1}{\sqrt{n}} \sum_{i \in I_k} (\psi(S_i, \theta_i^0, \hat{\eta}^{I_k^c}) - \psi(S_i, \theta_i^0, \eta^0)) = o_P(1).$$

Due to Lemma E.3, we have $\frac{1}{\sqrt{N} \cdot \sigma_N} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0) \xrightarrow{d} \mathcal{N}(0,1)$ as $N \to \infty$. Due to Assumption A2, we therefore have

$$\frac{1}{\sqrt{N}\sigma_\infty^{-1}} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0) = \frac{1}{\sqrt{N} \cdot \sigma_N} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0) \cdot \sigma_N \sigma_\infty^{-1} \xrightarrow{d} \mathcal{N}(0,1),$$

as $N \to \infty$. Consequently, we have $\sqrt{N}\sigma_\infty^{-1}(\hat{\theta} - \theta_N^0) \xrightarrow{d} \mathcal{N}(0,1)$ as claimed. □

F Bootstrap variance estimator

We use the following assumption to establish the consistency of the bootstrap variance estimator. It is a high-level assumption, and we will not verify it in terms of the model (1); yet, assuming some form of continuity (as below) seems to be essentially necessary for the bootstrap to be consistent.

Assumption A7. To make the dependence of σ_∞^2 in (9) on the law of the response error terms ε_Y, the law of the covariates C, the nuisance functions η^0, and the network G, we introduce the functional

$$\sigma_\infty^2(P_\varepsilon, P_C, \eta^0;\, G) = \lim_{N \to \infty} \mathrm{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(S_i, \theta_i^0, \eta^0) \right),$$

which can be represented as

$$\sigma_\infty^2(P_\varepsilon, P_C, \eta^0;\, G) = \lim_{N \to \infty} \mathrm{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varphi(S_i, \eta^0) \right),$$

due to $\psi(S_i, \theta_i^0, \eta^0) = \varphi(S_i, \eta^0) - \theta_i^0$ and because the θ_i^0's are non-random.

We assume that $\sigma_\infty^2(P_\varepsilon, P_C, \eta^0;\, G)$ is continuous with respect to Mallows' distance $d_2(\cdot,\cdot)$ in the first and second argument and with respect to $\|\cdot\|_{P,2}$ in the third argument.

Proof of Theorem 2.3. The bootstrap variance relies on the same dependency structure induced by the network as σ_∞^2 and can be represented by

$$\sigma_\infty^2(\hat{P}_{\hat{\varepsilon}}, \hat{P}_C, \hat{\eta}; G) = \lim_{N\to\infty} \mathrm{Var}^*\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i^*, \hat{\theta}_i^0, \hat{\eta})\right),$$

where the construction of S_i^* is described in Section 2.5. Similarly, we can rewrite this bootstrap variance as

$$\sigma_\infty^2(\hat{P}_{\hat{\varepsilon}}, \hat{P}_C, \hat{\eta}; G) = \lim_{N\to\infty} \mathrm{Var}^*\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varphi(S_i^*, \hat{\eta})\right).$$

Due to Assumption 3.1 and [76], we have $d_2(\hat{P}_C, P_C) \xrightarrow{P} 0$, where $d_2(\cdot,\cdot)$ denotes Mallows' distance. Furthermore, due to $\|\hat{\varepsilon}_Y - \varepsilon_Y\|_{P,2} \xrightarrow{P} 0$, we also have $d_2(\hat{P}_{\hat{\varepsilon}}, P_{\hat{\varepsilon}}) \xrightarrow{P} 0$ (see [76] and [29, Lemma 5.4]). Due to $\|\hat{\eta}^{I_k^c} - \eta^0\|_{P,2} \xrightarrow{P} 0$ for $k \in [K]$, we obtain

$$\lim_{N\to\infty}\left|\mathrm{Var}^*\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varphi(S_i^*, \hat{\eta})\right) - \mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(S_i, \eta^0)\right)\right| \xrightarrow{P} 0,$$

which consequently establishes consistency of the bootstrap variance. □

G Consistent plugin variance estimator

An alternative to the bootstrap variance estimator can be constructed as described below. We do not recommend this estimator unless the sample size is large relative to the network connectivity, but its consistency can be derived under different and more explicit conditions than in (A7).

The challenge is that the unit-level effects θ_i^0 for $i \in [N]$ are not all equal. This is because the unit-level data points S_i are typically not identically distributed. The difference in distributions originates from the X- and Z-features that generally depend on a varying number of other units. If two unit-level data points S_i and S_j have the same distribution, then their unit-level treatment effects θ_i^0 and θ_j^0 coincide. If enough of these unit-level treatment effects coincide, we can use the corresponding unit-level data to estimate them. Subsequently, we describe this procedure.

We partition $[N]$ into sets $\mathcal{A}_d$ for $d \geq 0$ such that all unit-level data points S_i for $i \in \mathcal{A}_d$ have the same distribution. Provided that the sets $\mathcal{A}_d$ are large enough, we can consistently estimate the corresponding θ_d^0 for $d \geq 0$ by

$$\hat{\theta}_d = \frac{1}{|\mathcal{A}_d|}\sum_{i\in\mathcal{A}_d}\varphi(S_i, \hat{\eta}^{I_{k(i)}^c}), \tag{A7}$$

where $k(i)$ denotes the index in $[K]$ such that $i \in I_{k(i)}$. The convergence rate of these estimators is at least $N^{-1/4}$ (see Lemma G.3 in Section G.1 in the appendix). To achieve this rate, we require that the sets $\mathcal{A}_d$ contain at least of order $N^{3/4}$ many indices (see Assumption A5 in Section A in the appendix). The parametric convergence rate cannot be achieved in general because $\mathcal{A}_d$ is of smaller size than N, but the corresponding units may have the maximal $d_{\max}$ many ties in the network.

Subsequently, we characterize a situation in which the index d corresponds to the degree in the dependency graph G_D. This is the case if two unit-level data points S_i and S_j have the same distribution if and only if the units i and j have the same degree in G_D. We assume, given a unit i and some $m \in [N]\backslash\{i\}$, that (1) if C_m is part of Z_i, then C_m is also part of X_i, and vice versa; and (2) if W_m is part of X_i, then C_m is part of X_i and Z_i and vice versa. Consequently, if two units $i \neq j$ have the same degree in the dependency graph, then their X- and their Z-features are computed using the same number of random variables. Hence, X_i and X_j as well as Z_i and Z_j are identically distributed, and therefore, S_i and S_j have the same distribution. Thus, the sets $\mathcal{A}_d$ form partition of the units according to their degree in the dependency graph, i.e., $\mathcal{A}_d = \{i \in [N] : d(i) = d\}$ for

$d \geq 0$, where $d(i)$ denotes the degree of i in the dependency graph. There are $d_{\max} + 1 = o(N^{1/4})$ many such sets, and each of them is required to be of size at least of order $N^{3/4}$ in Lemma G.3. This is feasible because there are N units in total. Provided that the machine learning estimators of the nuisance functions converge at a rate faster than $N^{1/4}$ as specified by Assumption A6 in the appendix, we have the following consistent estimator of the asymptotic variance given in Theorem G.1. Algorithm 1 summarizes the whole procedure of point estimation and inference for the EATE where the variance is estimated as given in Theorem G.1. Nevertheless, this estimation scheme can be extended to general sets $\mathcal{A}_d$.

Theorem G.1. *Denote by $G_D = (V, E_D)$ the dependency graph on S_i, $i \in [N]$. For a unit $i \in [N]$, denote by $d(i)$ its degree in G_D and by $k(i)$ the number in $[K]$ such that $S_i \in I_{k(i)}$. In addition to the assumptions made in* Theorem 2.2*, also assume that* Assumptions A5 *and* A6 *stated in* Section A *in the appendix hold. Based on φ defined in* (4)*, we define the score function $\psi(S_i, \theta, \eta) = \varphi(S_i, \eta) - \theta$ for some general $\theta \in \mathbb{R}$ and the nuisance function triple $\eta = (g_1, g_0, h)$. Then,*

$$\frac{1}{N}\sum_{i=1}^{N}\psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) + \frac{2}{N}\sum_{\{i,j\}\in E_D}\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}})\psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}})$$

is a consistent estimator of the asymptotic variance σ^2_∞ in Theorem 2.2.

G.1 Proof of Theorem G.1

Lemma G.2. *Assume the assumptions of* Theorem G.1 *hold. Let $i \in [N]$. There exists a finite real constant C_7 independent of i such that $\|\psi(S_i, \theta^0_{d(i)}, \eta^0)\|_{P,4} \leq C_7$ holds. Consequently, for $i, j, m, r \in [N]$, we can also bound the following terms by finite uniform constants*:

- $\|\psi(S_i, \theta^0_{d(i)}, \eta^0)\|_{P,2}$,
- $\mathrm{Var}(\varphi(S_i, \eta^0))$,
- $\mathrm{Var}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0))$,
- $\mathrm{Cov}(\varphi(S_i, \eta^0), \varphi(S_j, \eta^0))$,
- $\mathrm{Var}(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0))$,
- $\mathrm{Cov}(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0), \psi(S_m, \theta^0_{d(m)}, \eta^0)\psi(S_r, \theta^0_{d(r)}, \eta^0))$.

Moreover, we have $\varphi^2(S_i, \eta^0) = O_P(1)$. Furthermore, we have $\psi^2(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) = O_P(1)$.

Proof of Lemma G.2. We have

$$\begin{aligned}\|\psi(S_i, \theta^0_{d(i)}, \eta^0)\|_{P,4} \leq \|g_1^0(D_i)\|_{P,4} + \|g_0^0(D_i)\|_{P,4} + \left\|\frac{W_i}{h^0(U_i)}\right\|_{P,4}\|Y_i - g_1^0(D_i)\|_{P,\infty} \\ + \left\|\frac{1 - W_i}{1 - h^0(U_i)}\right\|_{P,2}\|Y_i - g_0^0(D_i)\|_{P,\infty} + |\theta^0_{d(i)}|.\end{aligned} \tag{A8}$$

All individual summands in the aforementioned decomposition are bounded by a finite real constant independent of i due to Assumption A3. Therefore, there exists a finite real constant C_7 independent of i such that $\|\psi(S_i, \theta^0_i, \eta^0)\|_{P,4} \leq C_7$ holds.

The other terms in the statement of the present lemma are bounded as well by finite real constants independent of $i, j, m, r \in [N]$ due to Hölder's inequality.

Moreover, we have $\psi^2(S_i, \eta^0) = O_P(1)$ because $\|\psi^2(S_i, \eta^0)\|_{P,2}$ is bounded by a constant that is independent of i.

Furthermore, with P-probability at least $1 - \Delta_N$, we have

$$\mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}})|\mathcal{S}_{I^c_{k(i)}}] \leq \sup_{\eta\in\mathcal{T}} \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta)] = \sup_{\eta\in\mathcal{T}} ||\psi(S_i, \theta^0_{d(i)}, \eta)||^2_{P,2}.$$

The term $||\psi(S_i, \theta^0_{d(i)}, \eta)||^2_{P,2}$ is bounded by a real constant that is independent of i and η because the derivation in (A8) also holds with η^0 replaced by $\eta \in \mathcal{T}$ due to Assumption A4. □

Lemma G.3. (Convergence rate of unit-level effect estimators) *Assume the assumptions of* Theorem G.1 *hold. Let* $d \geq 0$, *and assume that all assumptions of* Section A *in the appendix hold. Then, we have* $\hat{\theta}_d - \theta^0_d = o_P(N^{-1/4})$, *where* $\hat{\theta}_d$ *is as in* (A7).

Proof of Lemma G.3. Let $d \geq 0$. Due to the definition of $\hat{\theta}_d$ given in (A7) and Lemma 2.1, we have

$$\begin{aligned} N^{\frac{1}{4}}(\hat{\theta}_d - \theta^0_d) &= \frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i\in\mathcal{A}_d} (\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \mathbb{E}[\varphi(S_i, \eta^0)]) \\ &= \frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i\in\mathcal{A}_d} (\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)) + \frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i\in\mathcal{A}_d} (\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)]). \end{aligned} \tag{A9}$$

Subsequently, we show that the two sets of summands in (A9) are of order $o_P(1)$. We start with the first set of summands. Let $i \in \mathcal{A}_d$. With P-probability at least $1 - \Delta_N$, we have

$$\sqrt{\mathbb{E}[(\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0))^2|\mathcal{S}_{I^c_k}]} \lesssim \sqrt{\delta_N} N^{-\kappa},$$

due to equation (A6). Hence, we have $|\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| = O_P(\sqrt{\delta_N} N^{-\kappa})$ based on Chernozhukov et al. [20, Lemma 6.1]. Consequently, we have

$$\frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i\in\mathcal{A}_d} |\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| = O_P(\sqrt{\delta_N} N^{\frac{1}{4}-\kappa}) = o_P(1),$$

because we have $\kappa \geq 1/4$ by Assumption A6. Next, we show that the second set of summands in (A9) is of order $o_P(1)$. Let $\varepsilon > 0$. We have

$$\begin{aligned} &P\left(\left|\frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i\in\mathcal{A}_d} (\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)])\right|^2 > \varepsilon^2\right) \\ &\leq \frac{N^{\frac{1}{2}}}{\varepsilon^2|\mathcal{A}_d|^2}\left(\sum_{i\in\mathcal{A}_d} \mathrm{Var}(\varphi(S_i, \eta^0)) + \sum_{i,j\in\mathcal{A}_d, i\neq j} \mathrm{Cov}(\varphi(S_i, \eta^0), \varphi(S_i, \eta^0))\right) \\ &= \frac{N^{\frac{1}{2}}}{\varepsilon^2|\mathcal{A}_d|^2}(|\mathcal{A}_d| + 2|E_D \cap \mathcal{A}^2_d|)O(1), \end{aligned}$$

because $\mathrm{Var}(\varphi(S_i, \eta^0))$ and $\mathrm{Cov}(\varphi(S_i, \eta^0), \varphi(S_i, \eta^0))$ are bounded by constants uniformly over i due to Lemma G.2, and because $\mathrm{Cov}(\varphi(S_i, \eta^0), \varphi(S_i, \eta^0))$ does not equal 0 only if $\{i, j\} \in E_D \cap \mathcal{A}^2_d$, where E_D denotes the edge set of the dependency graph. There are $|\mathcal{A}_d|$ many nodes in $\mathcal{A}_d$, and each node has a maximal degree of $d_{\max}$. Thus, we have $|E_D \cap \mathcal{A}^2_d| \leq 1/2|\mathcal{A}_d|d_{\max}$. Due to $d_{\max} = o(N^{1/4})$ and $|\mathcal{A}_d| = \Omega(N^{3/4})$, which hold according to Assumptions A1 and A5, we obtain

$$\frac{N^{\frac{1}{2}}}{\varepsilon^2|\mathcal{A}_d|^2}(|\mathcal{A}_d| + 2|E \cap \mathcal{A}^2_d|)O(1) = o(1).$$

Consequently, we also have

$$\left|\frac{N^{\frac{1}{4}}}{|\mathcal{A}_d|} \sum_{i\in\mathcal{A}_d} (\varphi(S_i, \eta^0) - \mathbb{E}[\varphi(S_i, \eta^0)])\right| = o_P(1).$$

□

Lemma G.4. (Consistent variance estimator part I) *Assume the assumptions of* Theorem G.1 *hold. We have*

$$\left| \frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) - \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta^0)]) \right| = o_P(1).$$

Proof of Lemma G.4. We have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) - \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta^0)]) \\
&\quad = \frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) - \psi^2(S_i, \hat{\theta}_{d(i)}, \eta^0)) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \hat{\theta}_{d(i)}, \eta^0) - \psi^2(S_i, \theta^0_{d(i)}, \eta^0)) \\
&\qquad + \frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0) - \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta^0)]).
\end{aligned} \tag{A10}$$

We bound the three sets of summands in (A10) individually. The first set of summands can be expressed as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) - \psi^2(S_i, \hat{\theta}_{d(i)}, \eta^0)) \\
&\quad = \frac{1}{N}\sum_{i=1}^{N}(\varphi^2(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i, \eta^0)) - \frac{2}{N}\sum_{i=1}^{N}\hat{\theta}_{d(i)}(\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)).
\end{aligned}$$

We have

$$\left| \frac{1}{N}\sum_{i=1}^{N}(\varphi^2(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i, \eta^0)) \right| = o_P(1) \tag{A11}$$

because the function $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is continuous and due to equation (A6). Indeed, let $\varepsilon > 0$. Because the function $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is continuous, there exists $\delta > 0$ such that if $|\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| < \delta$, then also $|\varphi^2(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i, \eta^0)| < \varepsilon$. Consequently, we have

$$\begin{aligned}
&P(|\varphi^2(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi^2(S_i, \eta^0)| > \varepsilon | \mathcal{S}_{I^c_{k(i)}}) \\
&\quad \leq P(|\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| > \delta | \mathcal{S}_{I^c_{k(i)}}) \\
&\quad \leq \frac{1}{\delta}\sup_{\eta \in \mathcal{T}} ||\varphi(S_i, \eta) - \varphi(S_i, \eta^0)||_{P,1},
\end{aligned}$$

with P-probability at least $1 - \Delta_N$, and we infer (A11) due to (A6). The estimator $\hat{\theta}_{d(i)}$ is a consistent estimator of $\theta^0_{d(i)}$ due to Lemma G.3, and $\theta^0_{d(i)}$ is bounded independent of i due to Assumption 3.5. Moreover, we have $|\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)| = o_P(1)$ due to (A6) and Chernozhukov et al. [20, Lemma 6.1]. Consequently, we have

$$\left| \frac{2}{N}\sum_{i=1}^{N}\hat{\theta}_{d(i)}(\varphi(S_i, \hat{\eta}^{I^c_{k(i)}}) - \varphi(S_i, \eta^0)) \right| = o_P(1),$$

due to Hölder's inequality. Hence, the first set of summands in (A10) is of order $o_P(1)$. The second set of summand in (A10) can be decomposed as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}(\psi^2(S_i, \hat{\theta}_{d(i)}, \eta^0) - \psi^2(S_i, \theta^0_{d(i)}, \eta^0)) \\
&\quad = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}^2_{d(i)} - (\theta^0_{d(i)})^2) - \frac{2}{N}\sum_{i=1}^{N}(\hat{\theta}_{d(i)} - \theta^0_{d(i)})\varphi(S_i, \eta^0).
\end{aligned}$$

We have $|\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}^2_{d(i)} - (\theta^0_{d(i)})^2)| = o_P(1)$ due to Lemma G.3. Lemma G.2 bounds $\varphi^2(S_i, \eta^0)$ in probability. Due to Hölder's inequality, we obtain

$$\left| \frac{2}{N} \sum_{i=1}^{N} (\hat{\theta}_{d(i)} - \theta^0_{d(i)}) \varphi(S_i, \eta^0) \right| = o_P(1).$$

Consequently, the second set of summands in (A10) is of order $o_P(1)$. Finally, we bound the third set of summands in (A10). Let $\varepsilon > 0$. We have

$$\begin{aligned}
&P\left(\left| \frac{1}{N} \sum_{i=1}^{N} (\psi^2(S_i, \theta^0_{d(i)}, \eta^0) - \mathbb{E}[\psi^2(S_i, \theta^0_{d(i)}, \eta^0)]) \right|^2 > \varepsilon^2 \right) \\
&\leq \frac{1}{\varepsilon^2 N^2} \left(\sum_{i=1}^{N} \mathrm{Var}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0)) + \sum_{i,j \in [N], \{i,j\} \in E_D} \mathrm{Cov}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0), \psi^2(S_j, \theta^0_{d(j)}, \eta^0)) \right) \\
&\leq \frac{1}{\varepsilon^2 N^2} (NO(1) + N d_{\max} O(1)) \\
&= o(1)
\end{aligned}$$

because $\mathrm{Var}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0))$ and $\mathrm{Cov}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0), \psi^2(S_j, \theta^0_{d(j)}, \eta^0))$ are bounded uniformly over i and j by Lemma G.2, because $\mathrm{Cov}(\psi^2(S_i, \theta^0_{d(i)}, \eta^0), \psi^2(S_j, \theta^0_{d(j)}, \eta^0))$ does not vanish only if $\{i, j\} \in E_D$, and because $d_{\max} = o(N^{1/4})$ by Assumption A1. Consequently, also the third set of summands in (A10) is of order $o_P(1)$, and we have established the statement of the present lemma. □

Lemma G.5. (Consistent variance estimator part II) *Assume the assumptions of* Theorem G.1 *hold. Denote by E_D the edge set of the dependency graph. We have*

$$\left| \frac{1}{N} \sum_{i,j \in [N], \{i,j\} \in E_D} (\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)]) \right| = o_P(1).$$

Proof of Lemma G.5. We have the decomposition

$$\begin{aligned}
&\frac{1}{N} \sum_{i,j \in [N], \{i,j\} \in E_D} (\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)]) \\
&= \frac{2}{N} \sum_{\{i,j\} \in E_D} (\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) - \psi(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}})) \\
&\quad + \frac{2}{N} \sum_{\{i,j\} \in E_D} (\psi(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) - \psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)) \\
&\quad + \frac{2}{N} \sum_{\{i,j\} \in E_D} (\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0) - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0) \psi(S_j, \theta^0_{d(j)}, \eta^0)]).
\end{aligned} \tag{A12}$$

Subsequently, we bound the three sets of summands in (A12) individually. We start by bounding the first set of summands. We have

$$\begin{aligned}
&\frac{1}{N} \sum_{\{i,j\} \in E_D} (\psi(S_i, \hat{\theta}_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \hat{\theta}_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) - \psi(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}})) \\
&= \frac{2}{N} \sum_{\{i,j\} \in E_D} (\theta^0_{d(i)} - \hat{\theta}_{d(i)}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) + \frac{1}{N} \sum_{\{i,j\} \in E_D} (\theta^0_{d(i)} - \hat{\theta}_{d(i)})(\theta^0_{d(j)} - \hat{\theta}_{d(j)}).
\end{aligned}$$

We have

$$\left| \frac{1}{N} \sum_{\{i,j\} \in E_D} (\theta^0_{d(i)} - \hat{\theta}_{d(i)}) \psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) \right|$$

$$
\begin{aligned}
&\leq \sqrt{\frac{1}{N}\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)} - \hat{\theta}_{d(i)})^2}\sqrt{\frac{1}{N}\sum_{\{i,j\}\in E_D}\psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}})} \\
&= \frac{1}{N}|E_D| o_P(N^{-1/4}) \\
&= d_{\max} o_P(N^{-1/4}) \\
&= o_P(1)
\end{aligned}
$$

due to Hölder's inequality, Lemma G.3, Lemma G.2, and Assumption A1. Moreover, we have

$$
\left|\frac{1}{N}\sum_{\{i,j\}\in E_D}(\theta^0_{d(i)} - \hat{\theta}_{d(i)})(\theta^0_{d(j)} - \hat{\theta}_{d(j)})\right| = \frac{1}{N}|E_D| o_P(N^{-1/2}) = o_P(1)
$$

due to Hölder's inequality, Lemma G.3, and Assumption A1. Consequently, the first set of summands in (A12) is of order $o_P(1)$. We proceed to bound the second set of summands in (A12). Let $\{i,j\} \in E_D$. Due to the construction of $\mathcal{S}_{I_{k(i)}}$ and $\mathcal{S}_{I^c_{k(i)}}$, we have $S_i = (W_i, C_i, X_i, Z_i, Y_i) \in \mathcal{S}_{I_{k(i)}}$, and none of W_i, C_i, Y_i, or the variables used to compute X_i belong to $\mathcal{S}_{I^c_{k(i)}}$. Moreover, the variables W_i, C_i, Y_i, and the variables used to compute X_i also cannot belong to $\mathcal{S}_{I^c_{k(j)}}$ as otherwise we would have $S_i \perp\!\!\!\perp S_j$, and consequently, $\{i,j\} \notin E_D$. Therefore, we have

$$
\begin{aligned}
&\mathbb{E}[|\psi(S_i, \theta^0_{d(i)}, \hat{\eta}^{I^c_{k(i)}})\psi(S_j, \theta^0_{d(j)}, \hat{\eta}^{I^c_{k(j)}}) - \psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0)||\mathcal{S}_{I^c_{k(i)}}, \mathcal{S}_{I^c_{k(j)}}] \\
&\quad\leq \sup_{\eta_1,\eta_2\in\mathcal{T}} \mathbb{E}[|\psi(S_i, \theta^0_{d(i)}, \eta_1)\psi(S_j, \theta^0_{d(j)}, \eta_2) - \psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0)|] \\
&\quad\leq \sup_{\eta_1\in\mathcal{T}} ||\varphi(S_i, \eta_1) - \varphi(S_i, \eta^0)||_{P,2}||\psi(S_j, \theta^0_{d(j)}, \eta^0)||2 \\
&\qquad + \sup_{\eta_2\in\mathcal{T}} ||\psi(S_i, \theta^0_{d(i)}, \eta^0)||_{P,2}||\varphi(S_j, \eta_2) - \varphi(S_j, \eta^0)||_{P,2} \\
&\qquad + \sup_{\eta_1,\eta_2\in\mathcal{T}} ||\varphi(S_i, \eta_1) - \varphi(S_i, \eta^0)||_{P,2}||\varphi(S_j, \eta_2) - \varphi(S_j, \eta^0)||_{P,2},
\end{aligned}
$$

with P-probability at least $1 - \Delta_N$ due to Hölder's inequality. Because all terms mentioned earlier are uniformly bounded due to Lemma G.2, we infer that the second set of summands in (A12) is of order $o_P(1)$ based on Chernozhukov et al. [20, Lemma 6.1]. Finally, we bound the third set of summands in (A12). Let $\varepsilon > 0$. We have

$$
\begin{aligned}
&P\left(\left|\frac{1}{N}\sum_{\{i,j\}\in E_D}(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0) - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0)])\right|^2 > \varepsilon^2\right) \\
&\leq \frac{1}{\varepsilon^2 N^2}\Bigg(\sum_{\{i,j\}\in E_D}\mathrm{Var}(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_i, \theta^0_{d(j)}, \eta^0)) \\
&\qquad + \sum_{\{i,j\},\{m,r\}\in E_D,\text{unequal}}\mathrm{Cov}(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0), \psi(S_m, \theta^0_{d(m)}, \eta^0)\psi(S_r, \theta^0_{d(r)}, \eta^0))\Bigg).
\end{aligned} \tag{A13}
$$

Due to Lemma G.2, the variance and covariance terms in (A13) are uniformly bounded by constants. Furthermore, the covariance terms do only not equal 0 if S_i depends on S_m or S_r, or if S_j depends on S_m or S_r. In order to better describe these dependency relationships, we build a graph on the edge set of the dependency graph. We consider the graph $G' = (V', E')$ with $V' = E_D$ and such that an edge $\{\{i,j\},\{m,r\}\} \in E'$ if and only if at least one of $\{i,m\},\{i,r\},\{j,m\},\{j,r\}$ belongs to E_D. Consequently, $\{\{i,j\},\{m,r\}\} \in E'$ if and only if $(S_i, S_j) \not\perp\!\!\!\perp (S_m, S_r)$, in which case the covariance term in (A13) corresponding to $\{i,j\}$ and $\{m,r\}$ does not vanish. Furthermore, we have $|E'| = 1/2|E_D|d'_{\max}$, where $d'_{\max}$ denotes the maximal degree of a node in G'. We have $d'_{\max} \leq 2d_{\max}$. Consequently, we have

$$
P\left(\left|\frac{1}{N}\sum_{\{i,j\}\in E_D}(\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0) - \mathbb{E}[\psi(S_i, \theta^0_{d(i)}, \eta^0)\psi(S_j, \theta^0_{d(j)}, \eta^0)])\right|^2 > \varepsilon^2\right)
$$

$$
\begin{aligned}
&\leq \frac{1}{\varepsilon^2 N^2}(|E_D| + |E'|)O(1) \\
&\leq \frac{1}{\varepsilon^2 N^2}(Nd_{\max} + Nd_{\max}^2)O(1) \\
&= \frac{1}{\varepsilon^2 N}(o(N^{1/4}) + o(N^{1/2}))O(1) \\
&= o(1)
\end{aligned}
$$

due to Assumption A1. Therefore, we have established the statement of the present lemma because we have verified that all three sets of summands in (A12) are of order $o_P(1)$. □

Proof of Theorem G.1. The proof follows from Lemmas G.4 and G.5. □

H Extension to estimate global effects

So far, we focused on the EATE. We intervened on each individual unit and left the treatment selections of the other units as they were.

Subsequently, we consider another type of treatment effect where we assess the effect of a single intervention that intervenes on all subjects simultaneously. Instead of the EATE in (2), we subsequently consider the GATE with respect to the binary vector $\pi \in \{0,1\}^N$ of treatment selections

$$
\xi_N^0(\pi) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[Y_i^{do(\boldsymbol{W}=\pi)} - Y_i^{do(\boldsymbol{W}=1-\pi)}], \tag{A14}
$$

where $\boldsymbol{W} = (W_1, \ldots, W_N)$ denotes the complete vector of treatment selections of all units. In practice, the most common choice is where all components of π equal 1, i.e., the treatment effect comes from comparing the situation where all units are assigned to the treatment versus where no-one gets the treatment.

We use the same definition for S_i, $i \in [N]$ as before and denote the dependency graph on S_i, $i \in [N]$ by $G_D = (V, E_D)$. Furthermore, we let $\alpha(i) = \{j \in [N] : \{i, j\} \in E_D\} \cup \{i\}$ for $i \in [N]$ denote the nodes that share an edge with i in the dependency graph together with i itself. For some real number $\xi \in \mathbb{R}$ and a nuisance function triple $\eta = (g_1, g_0, h)$, consider the score function

$$
\begin{aligned}
\psi(S_i, \theta, \xi) = {} & g_1(C_i, X_i) - g_0(C_i, X_i) + \left(\prod_{j\in\alpha(i)} \frac{W_j}{h(C_j, Z_j)}\right)(Y_i - g_1(C_i, X_i)) \\
& - \left(\prod_{j\in\alpha(i)} \frac{1 - W_j}{1 - h(C_j, Z_j)}\right)(Y_i - g_0(C_i, X_i)) - \xi.
\end{aligned} \tag{A15}
$$

In contrast to the score that we used for the EATE, this score includes additional factors $\frac{W_j}{h(C_j, Z_j)}$ and $\frac{1 - W_j}{1 - h(C_j, Z_j)}$ for units j that share an edge with i in the dependency graph. With the GATE, when we globally intervene on all treatment selections at the same time, this also influences the X_i that are present in g_1 and g_0. In the score (A15), the "correction terms" $(\prod_{j\in\alpha(i)} \frac{W_j}{h(C_j, Z_j)})(Y_i - g_1(C_i, X_i))$ and $(\prod_{j\in\alpha(i)} \frac{1 - W_j}{1 - h(C_j, Z_j)})(Y_i - g_0(C_i, X_i))$ are only active if i and the units from which it receives spillover effects have the same observed treatment selection.

Let us denote by

$$
\xi_i^0 = \mathbb{E}[Y_i^{do(\boldsymbol{W}=\pi)} - Y_i^{do(\boldsymbol{W}=1-\pi)}] = \mathbb{E}[g_1^0(C_i, X_i^\pi) - g_0^0(C_i, X_i^{1-\pi})]
$$

the ith contribution in (A14). Here,

$$
X_i^\pi = (f_x^1(\pi_{-i}, C_{-i}, A), \ldots, f_x^r(\pi_{-i}, C_{-i}, A))
$$

denotes the feature vector where W_j is replaced by π_j, and

$$X_i^{1-\pi} = (f_x^1(1-\pi_{-i}, C_{-i}, A), \ldots, f_x^r(1-\pi_{-i}, C_{-i}, A))$$

denotes the feature vector where W_j is replaced by $1-\pi_j$. The features Z_i^{π} and $Z_i^{1-\pi}$ are defined analogously. Similar to Lemma 2.1, it can be shown that $\mathbb{E}[\psi(S_i, \xi_i^0, \eta^0)] = 0$ holds, which lets us identify the global treatment effect ξ_N^0 by

$$\xi_N^0 = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\varphi(S_i, \eta^0)],$$

where

$$\varphi(S_i, \eta) = g_1(C_i, X_i) - g_0(C_i, X_i) + \left(\prod_{j\in\alpha(i)}\frac{W_j}{h(C_j, Z_j)}\right)(Y_i - g_1(C_i, X_i)) - \left(\prod_{j\in\alpha(i)}\frac{1-W_j}{1-h(C_j, Z_j)}\right)(Y_i - g_0(C_i, X_i)).$$

To estimate ξ_N^0, we apply the same procedure as for the ATE. The only difference is that when we evaluate the machine learning estimates, we do not use the observed treatment selections, but instead insert the respective components of π and $1-\pi$. However, we insert the actually observed treatment selections in the product terms $\prod_{j\in\alpha(i)}\frac{W_j}{h(C_j, Z_j)}$ and $\prod_{j\in\alpha(i)}\frac{1-W_j}{1-h(C_j, Z_j)}$. This gives the estimator $\hat{\xi}$. Analogously to Theorem 2.2 for the EATE, also the GATE with respect to π converges at the parametric rate and follows a Gaussian distribution asymptotically.

Theorem H.1. (Asymptotic distribution of $\hat{\xi}$) *Assume* Assumption A3 (*with θ replaced by ξ*), 1, *and* A4 *in the appendix in* Section A *hold. Furthermore, assume that there exists a finite real constant L such that $|\alpha(i)| \leq L$ holds for all $i \in [N]$.*

Then, the estimator $\hat{\xi}$ of the GATE with respect to $\pi \in \{0,1\}^N$, ξ_N^0, satisfies

$$\sqrt{N}(\hat{\xi} - \xi_N^0) \xrightarrow{d} \mathcal{N}(0, \sigma_\infty),$$

where σ_∞ is characterized in Assumption A2 *with the ψ in* (A15). *The convergence in* (H.1) *is in fact uniformly over the law P of the observations.*

This theorem requires that the number of spillover effects a unit receives is bounded. Theorem 2.2 that establishes the parametric convergence rate and asymptotic Gaussian distribution of the EATE estimator did not require such an assumption. The reason is that $h^0(C_i, Z_i)$ represents the conditional expectation of W_i given C_i and Z_i and consequently a probability taking values in the interval (0,1). If we allowed $|\alpha(i)|$ to grow with N, the products $\prod_{j\in\alpha(i)}\frac{W_j}{h(C_j, Z_j)}$ and $\prod_{j\in\alpha(i)}\frac{1-W_j}{1-h(C_j, Z_j)}$ would diverge.

To estimate σ_∞^2 in Theorem H.1, we can apply the procedure described in Section G, where we replace ψ, φ, and the point estimators by the respective new quantities. Also an analog of Theorem G.1 holds, but where we assume the setting of Theorem H.1 holds and that $|\mathcal{A}_d| \to \infty$ as $N \to \infty$ for all $d \geq 0$. In particular, we do not require Assumptions A5 and A6 formulated in the appendix in Section A. Furthermore, to prove consistency of the variance estimator, it is sufficient to establish that the degree-specific causal effect estimators $\hat{\xi}_d$, which are defined analogously to $\hat{\theta}_d$, are consistent. In particular, they are not required to converge at a particular rate.

Also van der Laan [14], Sofrygin and van der Laan [10], and Ogburn et al. [6] consider semiparametric estimation of the GATE using TMLE. They also require a uniform bound of the number of spillover effects a unit receives to achieve the parametric convergence rate of their estimator.