Learning and Testing Exposure Mappings of Interference using Graph Convolutional Autoencoders

Martin Huber MARTIN.HUBER@UNIFR.CH
University of Fribourg, Department of Economics

Jannis Kueck KUECK@DICE.HHU.DE
Heinrich Heine University Düsseldorf, Düsseldorf Institute for Competition Economics

Mara Mattes MATTES@DICE.HHU.DE
Heinrich Heine University Düsseldorf, Düsseldorf Institute for Competition Economics

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Interference or spillover effects arise when an individual's outcome (e.g., health) is influenced not only by their own treatment (e.g., vaccination) but also by the treatment of others, creating challenges for evaluating treatment effects. Exposure mappings provide a framework to study such interference by explicitly modeling how the treatment statuses of contacts within an individual's network affect their outcome. Most existing research relies on a priori exposure mappings of limited complexity, which may fail to capture the full range of interference effects. In contrast, this study applies a graph convolutional autoencoder to learn exposure mappings in a data-driven way, which exploit dependencies and relations within a network to more accurately capture interference effects. As our main contribution, we introduce a machine learning-based test for the validity of exposure mappings and thus test the identification of the direct effect. In this testing approach, the learned exposure mapping is used as an instrument to test the validity of a simple, user-defined exposure mapping. The test leverages the fact that, if the user-defined exposure mapping is valid (so that all interference operates through it), then the learned exposure mapping is statistically independent of any individual's outcome, conditional on the user-defined exposure mapping. Conversely, a violation of this conditional independence indicates that the researcher-defined mapping fails to capture some relevant interference. We propose a conditional mean (rather than full) independence test sufficient for identifying average direct effects and study its finite-sample performance via simulation.

Keywords: causal inference, network interference, graph convolutional autoencoder, conditional independence, instrumental variables

1. Introduction

Interference or spillover effects, where an individual's outcome (e.g., health) is influenced not only by their own treatment (e.g., vaccination) but also by the treatment of others, pose substantial challenges for causal inference, particularly when arbitrary forms of interference are allowed; see Manski (2013). Exposure mappings, as discussed in Aronow and Samii (2017) and Eckles et al. (2017), provide a structured framework to study such interference by explicitly modeling the mechanisms through which the treatment statuses of social contacts within an individual's network influence their outcome. For example, an exposure mapping might specify that an interference effect occurs if at least one contact is treated, but does not depend on the number of treated contacts. By defining such mappings, researchers can separate interference effects from the direct effect of a treatment,

provided that they appropriately account for differences in the probabilities of specific exposure mappings across individuals - for instance, due to variation in network structure. However, most existing research relies on exposure mappings of limited complexity that are a priori defined by the researcher in an ad hoc manner, which risks failing to capture the full range of interference effects. In this paper, we apply graph convolutional autoencoders (GCAs), which are a specific type of graph neural networks (GNNs), to learn exposure mappings in a data-driven way based on network embeddings. By exploiting network dependencies and relations, GCAs allow us to capture complex patterns of interference that may be missed by simple, a priori defined mappings. As our main contribution, we introduce a machine learning-based testing framework for the validity of exposure mappings. Our approach uses a learned, complex exposure mapping as an instrument to assess whether a given simpler, researcher-defined exposure mapping sufficiently captures all interference. Specifically, if the simpler mapping accurately captures all interference effects such that any interference operates through them, then the learned mapping should be statistically independent of the individual's outcome, conditional on the less complex mapping. By examining violations of this condition, we can assess whether less complex, researcher-defined mappings fail to capture some of the underlying interference. Although developed in the context of interference, our testing framework is related to conditional independence testing approaches in de Luna and Johansson (2014) and Huber and Kueck (2023), who investigate tests for joint satisfaction of conditional treatment exogeneity and instrumental variable (IV) assumptions. Analogously, we test whether the user-defined exposure mapping is both exogenous and sufficient to capture all interference effects. If the exposure mapping is correctly specified, then the treatment assignments of others affect an individual's outcome only through this mapping, implying an IV type exclusion restriction. In this case, the learned embedding should be conditionally independent of the outcome given the user-defined exposure mapping, which forms the basis of our test.

We base the estimation of the direct effect of the treatment on the outcome, as well as the proposed testing procedure, on the double machine learning (DML) framework of Chernozhukov et al. (2018), which builds on doubly robust score functions; see Robins et al. (1994) and Robins and Rotnitzky (1995). The DML framework satisfies the Neyman (1959)-orthogonality condition, which implies that the resulting estimators and tests are relatively insensitive to modest approximation errors in the estimation of the exposure mapping and of the treatment or outcome models. Consequently, the estimators and tests satisfy asymptotic normality under specific regularity conditions, in particular if the machine learning methods used to estimate these models, such as GNNs, converge at a rate of $o(n^{-1/4})$ to the respective true model. Within this framework, we focus on identifying average direct effects and, correspondingly, on testing conditional mean (rather than full) independence.

Our study contributes to the growing literature that seeks to construct exposure mappings by exploiting information contained in the data. For instance, Bargagli-Stoffi et al. (2023) propose a tree-based method to assess heterogeneity in direct treatment and interference effects with respect to individual, neighborhood, and network characteristics. In their framework, it is assumed that interference occurs only within, but not across, predefined clusters (e.g., geographic regions) - a setting known as partial interference; see, e.g., Sobel (2006), Hong and Raudenbush (2006), and Hudgens and Halloran (2008). However, interference may still vary depending on the network structure within clusters. The proposed method therefore aims at detecting network structures and classifying clusters that exhibit similar interference effects based on tree-based algorithms. In contrast, our method is not confined to learning exposure mappings within clusters. Instead, we learn

exposure mappings as embeddings in a GCA, which constitutes a highly flexible approach capable of capturing complex network dependencies. Furthermore, we complement this estimation strategy with a novel statistical testing procedure to validate exposure mappings. Sävje (2024) emphasizes that exposure mappings are often used both to define causal estimands and to impose identifying restrictions, and argues that these roles should be separated; under additional weak-dependence conditions, standard estimators can remain consistent for misspecification-robust exposure effects even when the mapping does not capture the full interference structure. Our contribution is complementary: taking the usual mapping-based identification perspective as a starting point, we provide a formal test of whether a simple, researcher-defined mapping is sufficient for identification in the observed data, and we learn a flexible alternative mapping when it is rejected. Ma and Tresp (2021) also propose a GNN-based approach to learn interference from the data, focusing on the problem of learning optimal treatment policies across subgroups (see, e.g., Manski (2004), Hirano and Porter (2009), Kitagawa and Tetenov (2018), Athey and Wager (2021)) in the presence of interference. In contrast, the focus of our study is not on optimal policy learning but rather testing whether the exposure mapping is correctly specified which allows the identification of the direct treatment effect. Veitch et al. (2019) consider the estimation of direct treatment effects while controlling for network-related confounders by conditioning on embeddings learned from networks using DML. To this end, they adapt the DML assumptions in Chernozhukov et al. (2018) to account for learned embeddings and present high-level conditions under which estimation of the direct treatment effect is $\sqrt{n}$-consistent. Also Ma et al. (2022) discuss direct effect estimation when accounting for confounding induced by network interference.

More closely related to our setting, Leung and Loupos (2025) propose using GNNs to control for network-induced confounding, with the goal of estimating both direct and interference effects and conducting statistical inference. To this end, they rely on the assumption of approximate neighborhood interference (ANI) introduced by Leung (2022), which is conceptually related to approximate sparsity as considered in lasso regression by Belloni et al. (2014). ANI posits that interference decays sufficiently fast with the distance between individuals in the network, thereby addressing the problem of potentially high-dimensional confounding induced by network structure. Leung and Loupos (2025) show that, under ANI and additional regularity conditions, the estimation of propensity score and outcome models can achieve convergence at a rate of $o(n^{-1/4})$, implying that direct and interference effect estimators based on the given exposure mappings can be $\sqrt{n}$-consistent and asymptotically normal when implemented within the DML framework. Importantly, the exposure mapping is not learned but prespecified by the researcher and the GNNs are only used to estimate the nuisance functions. These results demonstrate that some restriction on the complexity of interference is necessary to obtain well-behaved estimators. Specifically, the depth of the relevant interference network must not be too large, which allows GNNs with a limited number of layers to approximate the interference structure. Baharan Khatami et al. (2025) also propose a graph-based doubly robust estimator that uses graph neural networks to flexibly learn network confounding and to estimate both direct and interference effects using a prespecified exposure mapping. The issue of estimating exposure mappings is also related to the framework of DML estimation with generated (rather than directly observed) regressors, as considered, for instance, in models with estimated control functions; see e.g. Pan and Zhang (2024) and Escanciano and Pérez-Izquierdo (2023). Our study also applies the DML methodology in the context of interference and GNNs, but extends this line of research by proposing a statistical test to assess the validity of the exposure mappings.

2. Causal effects and identification

This section introduces the concept of exposure mappings, as well as the definition and identification of direct, interference and total effects under specific assumptions. Let D_i denote the treatment of individual i in a population of interest, and let $\mathcal{D}_{-i}$ represent the vector of treatments assigned to all other individuals (excluding individual i) in that population. Furthermore, let Y_i denote the observed outcome. Throughout, we use uppercase letters to denote random variables and lowercase letters for specific realizations. Using the potential outcomes framework, as proposed by Neyman (1923) and advocated by Rubin (1974), the potential outcome of individual i under specific treatment assignments $D_i = d$ and $\mathcal{D}_{-i} = \mathbf{d}$ is written as $Y_i(d, \mathbf{d})$. This contrasts with the standard assumption in most treatment evaluations, which impose the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980; Cox, 1958). SUTVA rules out interference effects, implying that the potential outcome depends only on individual i's own treatment: $Y_i(d)$. For the subsequent discussion, we assume a binary treatment, such that $d \in \{0, 1\}$.

When SUTVA is violated, one pathway to identifying direct, interference, and total effects of an individual's own treatment relies on exposure mappings, see Aronow and Samii (2017). Such mappings impose structure on how the treatment assignments of other individuals, $\mathcal{D}_{-i}$, influence the outcome of individual i. It is assumed that interference effects operate through an individual's social network, denoted by $\mathcal{A}$, which is observed - for example, as an adjacency matrix indicating which individuals interact with one another.[1] Exposure mappings can be viewed as sufficient statistics for capturing any interference effects within a network. More formally, the exposure for individual i, denoted by Z_i, defines strengths of interference as a function (or mapping) $\mathcal{F}$ of i's network $\mathcal{A}$, the treatment assignments of other individuals, $\mathcal{D}_{-i}$, and the covariates of other individuals $\mathcal{X}_{-i}$:

$$Z_i = \mathcal{F}(\mathcal{A}, \mathcal{D}_{-i}, \mathcal{X}_{-i}). \tag{1}$$

The complexity of interference captured by the function $\mathcal{F}$ determines the number of possible values Z_i can take. A common choice in the literature defines $\mathcal{F}$ as the number of individuals who are both treated according to $\mathcal{D}_{-i}$ and neighbors of individual i in the network $\mathcal{A}$, in which case Z_i takes values $z \in \{0, 1, 2, \ldots\}$. A simpler alternative specifies $\mathcal{F}$ as a binary indicator, where $Z_i = 1$ if at least one treated individual in $\mathcal{D}_{-i}$ is a neighbor of individual i in network $\mathcal{A}$, and $Z_i = 0$ otherwise. This results in a binary exposure: $z \in \{0, 1\}$. It is worth noting that we also allow the exposure mapping to depend on the covariates of other individuals.

Under a correctly specified exposure mapping in Equation (1), the potential outcome $Y_i(d, \mathbf{d})$ simplifies to $Y_i(d, z)$. This permits defining the average direct effect, interference effect, and total effect, denoted by $\gamma(z)$, $\delta(d, z, z')$, and $\Delta(z, z')$, respectively, which are functions of an individual's own treatment and the exposure:

$$\begin{aligned} \gamma(z) &= E[Y_i(1, z) - Y_i(0, z)], \\ \delta(d, z, z') &= E[Y_i(d, z) - Y_i(d, z')], \\ \Delta(z, z') &= E[Y_i(1, z) - Y_i(0, z')], \end{aligned} \tag{2}$$

where z and z' are two distinct exposures (e.g., 1 and 0 in the binary case). Next, we provide the formal assumptions on the causal structure that ensure identification of the causal parameters

1. The network $\mathcal{A}$ is typically assumed to be fixed. However, Li and Wager (2022) consider the network structure in a population as a random draw and propose an asymptotic framework for constructing confidence intervals for direct and interference effects under this assumption.

defined in Equation (2). We assume that we observe data $W_i = (Y_i, D_i, X_i)$ for a fixed network $\mathcal{A}$, $i = 1, \ldots, n$, where (X_i, D_i) is $i.i.d.$, but Y_i may depend on $\mathcal{D}_{-i}$ and $\mathcal{X}_{-i}$. Our first assumption concerns the exposure mapping and the treatment assignment:

Assumption 1 (Identification - Independence of D_i and Z_i)

$$Z_i = \mathcal{F}(\mathcal{A}, \mathcal{D}_{-i}, \mathcal{X}_{-i}), \text{ and } D_i = D_i(X_i).$$

First, Assumption 1 states that the true exposure $Z_i = \mathcal{F}(\mathcal{A}, \mathcal{D}_{-i}, \mathcal{X}_{-i})$ is a function only of $\mathcal{A}$, $\mathcal{D}_{-i}$, and $\mathcal{X}_{-i}$. Second, the treatment assignment of individual i may depend only on its own characteristics X_i, that is, $D_i = D_i(X_i)$. This implies that individual i's treatment assignment D_i is conditionally independent of its own exposure Z_i given the covariates $\mathcal{X} = (X_i, \mathcal{X}_{-i})$ and the network structure $\mathcal{A}$.

As discussed in Aronow and Samii (2017), the propensity score under inference is defined as the joint conditional probability of the treatment and the exposure, given the network structure $\mathcal{A}$ and the observed covariates $\mathcal{X} = (X_i, \mathcal{X}_{-i})$. Under Assumption 1, the propensity score is given by

$$p_i(d, z) = \Pr(D_i = d, Z_i = z | \mathcal{X}, \mathcal{A}) = \Pr(D_i = d | X_i) \cdot \Pr(Z_i = z | \mathcal{X}_{-i}, \mathcal{A}). \tag{3}$$

Using a directed acyclic graph (DAG) (see, e.g. Pearl (2000)), Figure 1 represents a causal structure where Assumption 1 is satisfied. In the graph, nodes represent variables, and arrows indicate causal associations between those variables. The treatment assignments of other individuals $\mathcal{D}_{-i}$, the network structure $\mathcal{A}$ and their covariates $\mathcal{X}_{-i}$ jointly determine the exposure Z_i, which influences the outcome Y_i in the presence of interference. The confounder X_i influences individual treatment assignment D_i and the outcome Y_i, whereas $\mathcal{X}_{-i}$ influences the treatment assignments of other individuals $\mathcal{D}_{-i}$. In this structure, the direct effect refers to the effect of an individual's own treatment D_i on their outcome Y_i, i.e., $D_i \to Y_i$. The interference effect corresponds to the impact of other individuals' treatment assignments $\mathcal{D}_{-i}$ on individual i's outcome Y_i through the exposure Z_i, i.e., $\mathcal{D}_{-i} \to Z_i \to Y_i$. We assume that both the network structure $\mathcal{A}$ and the covariates of other individuals $\mathcal{X}_{-i}$ do not directly affect the individual i's outcome Y_i, but only through Z_i.

Consistent with the causal structure in Figure 1, identification of the direct treatment effect of D_i requires that all backdoor paths from D_i to Y_i are blocked by the individual i's observed covariates X_i and the exposure Z_i. This motivates the following conditional exogeneity assumption:

Assumption 2 (Identification - Conditional exogeneity of D_i and Z_i)

$$Y_i(d, z) \perp\!\!\!\perp (D_i, Z_i) | \mathcal{X}, \mathcal{A} \quad \forall d \in \{0, 1\}, \; z \in \mathcal{Z}.$$

where $\mathcal{Z}$ denotes the support of Z_i.

Assumption 2 imposes that there are no unobserved variables that jointly affect Y_i and the treatment assignment D_i, or Y_i and the true exposure Z_i conditional on $\mathcal{X}$ and $\mathcal{A}$. Notably, under our assumed causal structure, there is no direct effect of $\mathcal{A}$ or $\mathcal{X}_{-i}$ on the outcome Y_i. As a result, conditioning on X_i alone would be sufficient in Assumption 2, since all paths from $\mathcal{X}_{-i}$ and $\mathcal{A}$ to Y_i are blocked by Z_i. Our identifying assumptions (1 and 2) rule out endogenous treatment choices driven by peers' behavior, strategic treatment interactions, or unobserved treatment-outcome confounders across the network. They appear plausible in environments where treatment assignment is individually determined or externally administered - such as randomized or quasi-randomized

interventions, eligibility rules based solely on individual covariates, or settings where agents do not take their peers' treatment into account at treatment assignment.

Furthermore, we require that the propensity score for any combination of treatment and exposure is strictly positive. This implies that the network structure does not deterministically determine the exposure. Thus, there exists variation in exposures, conditional on the network and the covariates, that can be leveraged to assess their effects. This leads to the following common support assumption:

Assumption 3 (Identification - Common support)

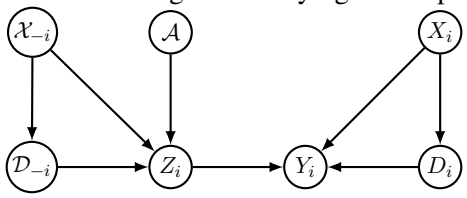
$$p_i(d,z) > 0 \quad \forall d \in \{0,1\}, z \in \mathcal{Z}.$$

Under Assumptions 1 - 3, the causal effects defined in Equations (2) are identified through the propensity score. As discussed in Aronow, Eckles, Samii, and Zonszein (2020), inverse probability weighting (IPW) (Horvitz and Thompson, 1952) can be applied, reweighting observations by the inverse of the propensity score to recover the mean potential outcomes and effects:

$$\begin{aligned}
E[Y_i(d,z)] &= E\left[\frac{Y_i \cdot I\{D_i = d, Z_i = z\}}{p_i(d,z)}\right], \\
\gamma(z) &= E\left[\frac{Y_i \cdot D_i \cdot I\{Z_i = z\}}{p_i(1,z)} - \frac{Y_i \cdot (1-D_i) \cdot I\{Z_i = z\}}{p_i(0,z)}\right], \\
\delta(d,z,z') &= E\left[\frac{Y_i \cdot I\{D_i = d, Z_i = z\}}{p_i(d,z)} - \frac{Y_i \cdot I\{D_i = d, Z_i = z'\}}{p_i(d,z')}\right], \\
\Delta(z,z') &= E\left[\frac{Y_i \cdot D_i \cdot I\{Z_i = z\}}{p_i(1,z)} - \frac{Y_i (1-D_i) \cdot I\{Z_i = z'\}}{p_i(0,z')}\right],
\end{aligned} \tag{4}$$

where $I\{\cdot\}$ is the indicator function, which is one if its argument is satisfied and zero otherwise.

Figure 1: Causal diagram satisfying Assumption 1.



3. Learning exposure mappings

While most studies evaluating interference effects rely on researcher-specified mappings, in real-world applications the function $\mathcal{F}$, and thus the true definition of the exposure mapping, is unknown. For this reason, we aim to approximate the true exposure, Z_i, using a graph convolutional autoencoder (GCA) related to the graph autoencoder (GAE) of Kipf and Welling (2016)[2]. In our supervised learning approach, $\tilde{Z}_i$ is learned as the embedding from an intermediate hidden layer of

2. Although inspired by the GAE of Kipf and Welling (2016), our GCA differs in that it is trained to predict Y_i in a supervised setting, rather than constructing the adjacency matrix.

a GCA trained to predict the outcome Y_i from the network $\mathcal{A}$, the treatment assignments of the individuals $\mathcal{D} = (D_i, \mathcal{D}_{-i})$, and the observed covariates $\mathcal{X} = (X_i, \mathcal{X}_{-i})$. A graph autoencoder consists of an encoder that maps the input graph and node features into a low-dimensional latent representation (embedding) capturing its most relevant information, and a decoder that predicts target values from this latent representation. In our setting, the GCA combines an encoder that consists of graph convolutional network (GCN) layers (Kipf and Welling, 2017) and a regression-based decoder. This architecture allows the model to automatically learn dependencies and relationships between nodes based on the network structure. The resulting learned exposures, denoted by $\tilde{Z}_i$, correspond to the embeddings produced by the encoder and are designed to capture how treatment assignments across the network affect individual outcomes.

We consider an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where the nodes are individuals $v_i \in \mathcal{V}$ with $|\mathcal{V}| = n$ and the edges are given by pairs $(v_i, v_j) \in \mathcal{E}$. The adjacency matrix $\mathcal{A} \in \mathbb{R}^{n \times n}$ represents the connections between the individuals and is defined as

$$\mathcal{A} = \begin{bmatrix} 0 & a_{12} & \dots & a_{1n} \\ a_{21} & 0 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & 0 \end{bmatrix}. \tag{5}$$

In the adjacency matrix, $a_{ij} = 1$ indicates that node i (i.e., individual i) is connected to node j, while $a_{ij} = 0$ indicates that there is no edge, i.e., no social connection, between nodes i and j. Each node has its own features and the input feature vector consists of the treatment assignments and the covariates of all individuals. Denote this matrix by M with dimensions $n \times 2$, where d_i is the individual assignment and x_i is the covariate of individual i:

$$M_{n \times 2} = \begin{bmatrix} d_1 & x_1 \\ d_2 & x_2 \\ \vdots & \vdots \\ d_n & x_n \end{bmatrix}. \tag{6}$$

Each layer in the encoder is indexed by $k \in \{0, \dots, K-1\}$, where $k = 0$ corresponds to the raw input, $k = 1$ to the first layer and $K - 1$ to the final layer of the encoder. The representation of the raw input (layer $k = 0$) is $H^{(0)} = M$. The graph convolutional layers $k \in \{1, \dots, K-1\}$ in the encoder follow the layer-wise propagation rule

$$H^{(k+1)} = \sigma\big(T^{-\frac{1}{2}} \mathcal{A} T^{-\frac{1}{2}} H^{(k)} \tilde{W}^{(k)}\big), \tag{7}$$

where T is the degree matrix. The matrix T contains zeros everywhere except on the diagonal, where $t_{ii} = \sum_j a_{ij}, \forall i \in \{1, \dots, n\}$. $\tilde{W}^{(k)}$ is a trainable and layer-specific weight matrix, and $\sigma(\cdot)$ is an activation function. $H^{(k+1)} \in \mathbb{R}^{n \times x_k}$ contains the node representations with each row corresponding to one node, where x_k is the feature dimension of that layer. The final encoder layer $H^{(K)}$, with $k = K - 1$, has feature dimension $x_k = 1$, as its output serves as the learned exposure $\tilde{Z}_i$, which we model as a one-dimensional embedding, that is, $H^{(k+1)} \in \mathbb{R}^{n \times 1}$ for $k = K - 1$. This layer-wise update in the graph convolutional layers corresponds to a standard message-passing mechanism, in which each node aggregates transformed information from its neighbors according to the graph structure in $\mathcal{A}$. To obtain an outcome prediction $\hat{\mathcal{Y}} \in \mathbb{R}^{n \times 1}$, the embeddings from the

encoder are passed into a regression-based decoder implemented as a linear layer, which maps the learned exposure representation into a predicted outcome:

$$\hat{\mathcal{Y}} = \tilde{W}_{dec} H^{(K)} + b_{dec} \tag{8}$$

where $\tilde{W}_{dec} \in \mathbb{R}^{1\times 1}$ is the trainable weight in the decoder layer, $H^{(K)}$ is the vector of learned exposures $\tilde{Z}_i$, $i \in \{1\dots,n\}$, and $b_{dec} \in \mathbb{R}$ is the bias term. The architecture of the GCA is illustrated in Appendix A Figure 3.

The GCA is trained by minimizing the mean squared error loss function $\mathcal{L}(Y_i, \hat{Y}_i) = \frac{1}{n}\sum_i (Y_i - \hat{Y}_i)^2$, which is appropriate, because the outcome variable Y_i is continuous.

It is important to note that the adjacency matrix $\mathcal{A}$ contains no self-loops, i.e., $a_{ii} = 0, \forall i \in \{1,\dots,n\}$. As a result, each node aggregates information exclusively from its neighbors. Thus, the representation $H^{(k+1)}$ at any encoder layer does not use the node's own features (D_i, X_i), but only those of other individuals $(\mathcal{D}_{-i}, \mathcal{X}_{-i})$. This is consistent with the interpretation of the exposure $\tilde{Z}_i$, which is intended to summarize how the treatments and characteristics of others affect individual i's outcome.

4. Testing the validity of exposure mappings

In this section, we discuss the testability of whether a researcher-defined exposure mapping is correctly specified for capturing all interference effects. Our approach uses the learned exposures $\tilde{Z}_i$, $i = 1,\dots,n$, from Section 3 as an instrument to assess whether the (simpler) exposures $\dot{Z}_i$ sufficiently capture all interference. Specifically, if the simpler mapping accurately captures all interference effects, then the exposure $\tilde{Z}_i$ is independent of the individual's outcome, conditional on $\dot{Z}_i$. By examining violations of this condition, we can assess whether the researcher-defined mapping fails to capture some of the underlying interference. Notably, the true exposure $Z_i = \mathcal{F}(\mathcal{A}, \mathcal{D}_{-i}, \mathcal{X}_{-i})$, for $i = 1,\dots,n$, is i.i.d. by assumption. We therefore retain the index i in the learned and researcher-defined exposures $\tilde{Z}_i$ and $\dot{Z}_i$ for notational convenience. Next, we introduce the assumptions that underlie our testing approach for validating the exposure mapping. These assumptions formalize the structural restrictions on how the exposure variables may depend on one another and on the outcome. We further assume that any statistical independencies correspond to d-separation in the underlying causal model, a condition known as causal faithfulness and formally stated below.

Assumption 4 (Testing method - causal structure and faithfulness) *We assume that*

$$\dot{Z}_i(y) = \dot{Z}_i, \text{ and } \tilde{Z}_i(\dot{z}, y) = \tilde{Z}_i \quad \forall \dot{z} \in \dot{\mathcal{Z}} \text{ and } y \in \mathcal{Y},$$

where $\dot{\mathcal{Z}}$ and $\mathcal{Y}$ denote the corresponding support of $\dot{Z}_i$ and Y and that only variables which are d-separated in some causal model are statistically independent.

The first part of Assumption 4 rules out reverse causal effects of outcome Y on $\dot{Z}_i$ and $\tilde{Z}_i$ as well as any causal effect of $\dot{Z}_i$ on $\tilde{Z}_i$. The latter reflects the fact that $\tilde{Z}_i$ is a causal parent of $\dot{Z}_i$, which is consistent with the interpretation of $\tilde{Z}_i$ as a more complex mapping from the network structure and neighbor treatments, while $\dot{Z}_i$ represents a simpler transformation thereof. The following assumption requires that every possible combination of $\tilde{Z}_i$ and $\dot{Z}_i$ occurs with positive probability.

Assumption 5 (Testing method - common support)

$$Pr(\dot{Z}_i = \dot{z}, \tilde{Z}_i = \tilde{z}) > 0 \quad \forall \dot{z} \in \dot{\mathcal{Z}}, \tilde{z} \in \tilde{\mathcal{Z}}$$

where $\dot{\mathcal{Z}}$ and $\tilde{\mathcal{Z}}$ denote the corresponding support of $\dot{Z}_i$ and $\tilde{Z}_i$.

In Assumption 6, we impose that the simpler exposure $\dot{Z}_i$ and the learned exposure $\tilde{Z}_i$ are statistically dependent.

Assumption 6 (Testing method - dependence between $\dot{Z}_i$ and $\tilde{Z}_i$)

$$\dot{Z}_i \not\perp\!\!\!\perp \tilde{Z}_i.$$

Together with Assumption 4, which rules out any effect of $\dot{Z}_i$ on $\tilde{Z}_i$, Assumption 6 ensures that $\tilde{Z}_i$ causally affects $\dot{Z}_i$. This corresponds either to a first-stage relationship in the IV literature or to the presence of (potentially unobserved) characteristics that jointly influence both $\dot{Z}_i$ and $\tilde{Z}_i$. This assumption is satisfied in Figure 2, where $\tilde{Z}_i$ serves as a causal parent of $\dot{Z}_i$. Assumptions 4 and 6, allow us to apply Theorem 1 of Huber and Kueck (2023) in order to construct a test for validating the exposure mapping. It is worth noting that we still rely on the causal structure shown in Figure 1 and Figure 2. Most importantly, we assume that there are no unobserved confounder jointly affecting D_i and $\tilde{Z}_i$ or D_i and $\dot{Z}_i$.

Theorem 1 *Conditional on Assumptions 4 and 6, it holds that*

$$Y_i(d, \dot{z}) \perp\!\!\!\perp \dot{Z}_i, \quad Y_i(d, \dot{z}) \perp\!\!\!\perp \tilde{Z}_i \iff Y_i \perp\!\!\!\perp \tilde{Z}_i | \dot{Z}_i = \dot{z}, \quad \forall \dot{z} \in \dot{\mathcal{Z}}, d \in \{0, 1\}. \tag{9}$$

Conditional on Assumptions 4 and 6, the testable implication $Y_i \perp\!\!\!\perp \tilde{Z}_i | \dot{Z}_i = \dot{z}$ is necessary and sufficient for the joint satisfaction of $Y_i(d, \dot{z}) \perp\!\!\!\perp \dot{Z}_i$ and $Y_i(d, \dot{z}) \perp\!\!\!\perp \tilde{Z}_i$ when considering potential outcomes $Y_i(d, \dot{z})$ matching $\dot{Z}_i = \dot{z}$ and $D_i = d$.

This statement follows directly from Theorem 1 of Huber and Kueck (2023), considering $\dot{Z}_i$ as the treatment variable and $\tilde{Z}_i$ as the suspected instrument. The first two conditions of Theorem 1 state two independence conditions, which correspond to selection-on-observables for the exposure $\dot{Z}_i$ and instrument validity for the learned exposure $\tilde{Z}_i$, respectively. The first condition imposes that the potential outcome $Y_i(d, \dot{z})$ is independent of $\dot{Z}_i$ for all possible exposure levels $\dot{z} \in \dot{\mathcal{Z}}$ and $d \in \{0, 1\}$, i.e., $Y_i(d, \dot{z}) \perp\!\!\!\perp \dot{Z}_i$. This rules out unobserved confounding between $\dot{Z}_i$ and Y_i. In Figure 2, this corresponds to the absence of dotted arrows from the unobserved confounder V_i to both $\dot{Z}_i$ and Y_i. The second independence condition of Theorem 1 imposes instrument validity for the learned exposure $\tilde{Z}_i$. It requires that $Y_i(d, \dot{z})$ is independent of $\tilde{Z}_i$, i.e., $Y_i(d, \dot{z}) \perp\!\!\!\perp \tilde{Z}_i$. This means that the learned mapping does not directly affect the potential outcome and that there are no unobserved confounders jointly affecting $\tilde{Z}_i$ and $Y_i(d, \dot{z})$. In Figure 2, this corresponds to the absence of the dotted arrows from the unobserved confounder U_i to both $\tilde{Z}_i$ and Y_i. Therefore, the second condition of Theorem 1 ensures that $\tilde{Z}_i$ satisfies the same type of independence requirements as a valid instrument in the IV literature. Following Theorem 1 of Huber and Kueck (2023), the testable conditional independence $Y_i \perp\!\!\!\perp \tilde{Z}_i | \dot{Z}_i = \dot{z}$ in the third condition of Equation (9) is necessary and sufficient for the joint satisfaction of the first two independence conditions. In our setting, this provides the following interpretation of the testable implication: we assess whether the

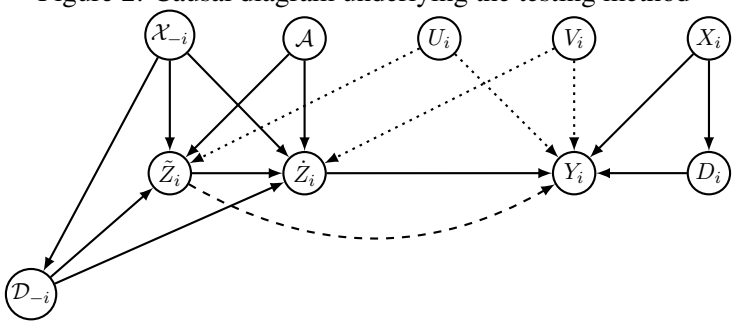
learned exposure $\tilde{Z}_i$ contains any residual association with the outcome Y_i after conditioning on the predefined exposure $\dot{Z}_i$. If such an association remains, the mapping $\dot{Z}_i$ fails to capture all relevant interference, and the learned mapping offers additional, causally meaningful information. In turn, if the conditional independence holds, the simpler exposure mapping is sufficient for capturing interference effects. In Figure 2 the dashed line (i.e., $\tilde{Z}_i \dashrightarrow Y_i$) indicates a violation of the testable condition, as $\dot{Z}_i$ does not fully capture the interference effects of $\tilde{Z}_i$ on Y_i due to its definition being too simple to account for all forms of interference.

In this paper, we focus on mean conditional independence between the outcome variable and the learned exposure mapping rather than on the full outcome distribution. This is sufficient for identifying average effects, see Theorem 2 of Huber and Kueck (2023), within our framework of average direct and interference effects defined in Equation (2). Modifying the testable implication in Equation (9) to hold in expectation yields the following testable implication:

$$E[Y_i|\dot{Z}_i = \dot{z}_i, \tilde{Z}_i = \tilde{z}_i] = E[Y_i|\dot{Z}_i = \dot{z}_i] \quad \forall \dot{z} \in \dot{\mathcal{Z}}, \tilde{z} \in \tilde{\mathcal{Z}}. \tag{10}$$

Equation (10) requires that all combinations of $\dot{Z}_i$ and $\tilde{Z}_i$ occurring in the conditioning sets are observed with positive probability. This is ensured by Assumption 5. Following Huber and Kueck (2023) and Apfel et al. (2024), testing whether the difference in expected outcomes in Equation (10) equals zero requires checking this condition for all values of $\dot{Z}$ and $\tilde{Z}$ in their respective supports, which can imply infinitely many testable implications. The solution is to test Equation (10) globally, i.e., across all values of $\dot{Z}$ and $\tilde{Z}$, using an aggregated L_2-type measure proposed in equation (3.4) in Apfel et al. (2024). Formally, we test $H_0 : \theta_0 = 0$ where $\theta_0 = (E[Y_i|\dot{Z}_i = \dot{z}_i, \tilde{Z}_i = \tilde{z}_i] - E[Y_i|\dot{Z}_i = \dot{z}_i])^2 + (E[Y_i|\dot{Z}_i = \dot{z}_i, \tilde{Z}_i = \tilde{z}_i] - E[Y_i|\dot{Z}_i = \dot{z}_i])$ which evaluates the testable implication in Equation (10). Relying on the DML approach, Apfel et al. (2024) derive an estimator $\hat{\theta}_0$ that is asymptotically normal and $\sqrt{n}$-consistent under suitable regularity conditions. We use this estimator to test whether the researcher-defined exposure is correctly specified, i.e., whether Equation (10) holds true. The technical details of the testing approach are given in Appendix B.

Figure 2: Causal diagram underlying the testing method



5. Estimation of causal effects

In this section, we outline the estimation of the causal effects defined in Section 2. In the following the exposure mapping Z_i is either the researcher-defined exposure mapping, $\dot{Z}_i$, or the learned exposure mapping, $\tilde{Z}_i$, depending on the result of the testing method introduced in Section 4. If H_0 is rejected, the learned exposure mapping is used for the estimation of the causal effect, i.e.,

$Z_i = \tilde{Z}_i$. If H_0 is not rejected, the researcher-defined exposure mapping is used instead, i.e., $Z_i = \dot{Z}_i$. When $Z_i = \tilde{Z}_i$, the exposure mapping may be difficult to interpret[3]. Hence, our focus is on the identification and estimation of the direct effect averaged over all exposure levels, $\gamma = E[\gamma(Z)]$. In applied work, researchers often prefer simple exposure mappings because they are easy to interpret and communicate. However, such simple mappings may fail to capture relevant interference effects and thus lead to biased effect estimates. The purpose of the proposed test is therefore to assess whether a simple, interpretable mapping is sufficient or whether a more complex mapping is required. If the simpler mapping is rejected, the learned exposure is primarily used to improve identification of the direct effect rather than to provide mechanistic insight.

To estimate these causal effects, we build on the identification assumptions introduced in Section 2. While IPW identification is valid under correct specification of the propensity score, outcome-regression-based identification using a model for the conditional mean outcome is valid under correct specification of that model. Both methods can be combined in a doubly robust (DR) approach, which remains consistent if either the propensity score or the outcome model is correctly specified (Robins et al., 1994; Robins and Rotnitzky, 1995). Moreover, the DR approach is first-order insensitive to small deviations of both the propensity score and the conditional mean outcome from their respective true models, a property known as Neyman-orthogonality (Neyman, 1959). This property is key for the application of the DML framework of Chernozhukov et al. (2018), in which propensity scores and outcome models are estimated with machine learning methods that may be prone to approximation errors. If the propensity score and outcome models are estimated with convergence rate $o(n^{-1/4})$, DML estimation of average direct, interference and total effects can attain $\sqrt{n}$-consistency under certain regularity conditions. Denoting by $\mu_i(d, z, x) = E[Y_i|D_i = d, Z_i = z, X_i = x]$ the conditional mean outcome, the DR expression for the mean potential outcome is given by

$$E[Y_i(d, z)] = E\Bigg[\underbrace{\mu_i(d, z, x) + \frac{I\{D_i = d, Z_i = z\}}{p_i(d, z)}(Y_i - \mu_i(d, z, x))}_{:=\phi(W_i)}\Bigg], \tag{11}$$

where $\psi(W_i) = \phi(W_i) - E[Y_i(d, z)]$ is the efficient score function.

The outcome regression $\mu_i(d, z, x)$ can be estimated using any machine learning method that satisfies the required convergence rate. The propensity score $p_i(d, z)$ defined in Equation (3) consists of two components. The first component $\Pr(D_i = d|X_i)$ can be estimated using, for example, logistic regression when the treatment is binary. The second component $\Pr(Z_i = z|\mathcal{X}_{-i}, \mathcal{A})$ is conditional on the network $\mathcal{A}$, which is why we suggest estimating it using the GNN-based propensity score estimator introduced in Leung and Loupos (2025). They use a graph isomorphism network (GIN) to estimate the propensity score, that is, the conditional distribution of exposure given the covariates of the other individuals $\mathcal{X}_{-i}$ and the network structure $\mathcal{A}$. Since the exposure can take multiple values, the GIN produces a probability for each possible exposure level through a multi-class output layer. The target variable is a one-hot-encoding of the observed exposure level. The model is trained using a logistic loss, $exp(\hat{m}_z)/\sum_{z' \in \mathcal{Z}} exp(\hat{m}_{z'})$, to estimate the propensity score,

3. Although the learned exposure may not be directly interpretable, it can still be related to standard network summaries in post-analysis, for example by regressing the learned embedding on treated-neighbor counts, two-hop exposure, centrality measures, or community-level treatment shares, by clustering nodes in embedding space into a small number of exposure types, or by conducting local perturbation analyses to assess which network features drive changes in the learned exposure.

where $\hat{m}_z$ denotes the GIN output corresponding to exposure level z. They show that, under approximate neighborhood interference (ANI), the propensity score can be estimated with a convergence rate of $o(n^{-1/4})$. This allows valid estimation of the average direct effect, $\gamma = E[\gamma(Z)]$, which we will also show in a simulation study in the next section.[4]

6. Simulation

This section provides a simulation study to investigate the finite sample behavior of the testing approach and direct effect estimation. The data are generated according to the following process:

$$
\begin{aligned}
&Y_i = \alpha + \delta Z_i + \gamma D_i + \xi X_i + \varepsilon_i, \quad \text{where } \varepsilon_i \sim N(0,1), \\
&D_i \sim bernoulli\Big(\frac{1}{1+\exp(-(1+2X_i))}\Big), \quad X_i \sim bernoulli(0.5), \\
&a_{ij} = I\{|\rho_i - \rho_j| \leq r_n\} \text{ with } \rho_i \sim U([0,1])^2 \text{ and } r_n = \Big(\frac{30}{\pi n}\Big)^{(1/2)},
\end{aligned}
$$

where $(\alpha, \delta, \gamma, \xi) = (-1, 5, 1, 1)$. The outcome Y_i is a linear function of the individual treatment D_i, the true exposure Z_i, the covariate X_i and an unobservable ε_i. The covariate X_i is a binary variable, and the binary treatment D_i is a function of X_i. Following Leung and Loupos (2025), the network structure $\mathcal{A}$ is generated from a random geometric graph model. The GCA used for learning the exposure mapping consists of two graph convolutional layers in the encoder and one regression-based decoder. The learning rate is 0.01 and the number of epochs is 200. For estimating the score function, the support of the continuous learned exposure variable is partitioned based on the quartiles of its distribution, i.e., $L = 4$.

We consider three settings to assess the performance of our testing approach, using 200 Monte Carlo replications for sample sizes $n = 500, 1000,$ and 2000. In the first setting, the true exposure mapping is given by the share of treated neighbors weighted by their covariates, i.e., $Z_i^{S1} = \frac{\sum_{j\neq i} a_{ij} D_j X_j}{\sum_{j\neq i} a_{ij}}$, and the researcher-defined exposure mapping coincides with the true exposure, i.e., $\dot{Z}_i^{S1} = \frac{\sum_{j\neq i} a_{ij} D_j X_j}{\sum_{j\neq i} a_{ij}}$. Thus, the researcher-defined exposure mapping is correctly specified and the exposure learned by the GCA should not contain additional information beyond the researcher-defined mapping. In Setting 2, the true exposure mapping depends on both first-order and second-order network neighborhoods[5]. Specifically, the true exposure is given by $Z_i^{S2} = \frac{\sum_{j\neq i} a_{ij} D_j X_j}{\sum_{j\neq i} a_{ij}} + \frac{\sum_{k\neq i} b_{ik} D_k X_k}{\sum_{k\neq i} b_{ik}}$, where $b_{ik} := I\{\sum_{i\neq j, j\neq k} a_{ij} a_{jk} > 0\} \cdot (1 - a_{ik}) \cdot I\{k \neq i\}$. The first term captures again the share of treated neighbors weighted by their covariates. The second term of the true exposure captures second-degree neighbors of individual i, i.e., nodes that are connected to i via a path of length two, excluding i itself and all direct neighbors. In contrast, the researcher-defined exposure is binary and only accounts for direct neighbors, i.e., $\dot{Z}_i^{S2} = I\{\sum_{j\neq i} a_{ij} D_j X_j > 0\}$.

4. While feasible in moderate-sized networks, the joint estimation of learned exposure mapping, nuisance functions and the DML-based testing procedure may become computationally demanding in very large graphs. In practice, recent advances in scalable GNN training (e.g., sampling-based methods (Hamilton et al., 2017; Zeng and Zhou, 2020) and distributed architectures (Zheng et al., 2020) could be integrated into our framework to improve scalability.
5. In Appendix C.2, we additionally consider a variant of Setting 2 in which the researcher-defined exposure mapping includes only the first-order component of the true exposure. This design isolated the role of the second-order term and shows how the performance of the proposed test depends on the extent of misspecification.

Thus, the researcher-defined exposure is misspecified. In the third setting, the researcher-defined exposure is equal to the one in Setting 1, i.e., $\dot{Z}_i^{S1} = \dot{Z}_i^{S3}$. However, the true exposure is a nonlinear transformation of the cumulative treated neighborhood intensity with an explicit threshold and saturation: $Z_i^{S3} = 1 - \exp\Big(-0.5 \ \max\Big\{0, \ \sum_{j \neq i} a_{ij} D_j X_j - 10\Big\}\Big)$. This setting therefore also represents a case of misspecification if researchers were to apply a linear specification to model the conditional mean outcome in Equation (11) based on $\dot{Z}_i^{S3}$.

	Setting 1		Setting 2		Setting 3	
n	rejection rate	mean p-value	rejection rate	mean p-value	rejection rate	mean p-value
500	0.00	0.76	0.32	0.25	0.8	0.11
1000	0.00	0.73	0.625	0.10	0.805	0.09
2000	0.00	0.69	0.895	0.03	0.83	0.09

Notes: 'rejection rate' gives the empirical rejection rate when setting the level of statistical significance to 0.05 (or 5%); R = 200 replications.

Table 1: Simulation results - testing method

Table 1 reports the empirical rejection rates and average p-values of the testing approach across the three settings. In Setting 1, where the researcher-defined exposure mapping is correct, the test never rejects the null hypothesis and the p-values remain high across all sample sizes. In Settings 2 and 3, where the researcher-defined exposure mapping is misspecified, the rejection rate increases in the sample size, while p-values decrease and approach the significance level $\alpha = 0.05$. Overall, these results are consistent with the theoretical implications discussed in Section 4. Although our framework is designed for hypothesis testing rather than quantification of misspecification, the test statistic and p-values are informative about the degree of misspecification. Small deviations between the true and simpler exposure mappings induce only weak conditional dependence in Equation (10), which results in large p-values and low rejection rates. By contrast, larger deviations imply stronger violations of the null hypothesis and smaller p-values. The chosen significance level thus determines the tolerated degree of misspecification. In the second part of the simulation study, we assess the estimation of the direct effect based on the learned exposure $\tilde{Z}_i$; see Appendix C.1 for the results.

7. Conclusion

In this paper, we develop a data-driven approach to learn exposure mappings in the presence of interference, instead of relying on a priori defined mapping. We use a graph convolutional autoencoder to learn exposure mapping that summarize how others' treatment assignments affect an individual's outcome. Since the identification of average direct effects depends crucially on the correct specification of the exposure mapping, we study whether a simple, researcher-defined exposure mapping is sufficient or a more complex, learned mapping is required. To this end, we propose a testing method based on conditional mean independence implications. This test evaluates whether a researcher-defined exposure mapping captures all relevant interference. Violations of the testable implication indicate that the predefined exposure mapping is misspecified and that a learned mapping should be used for estimating the direct effect. Overall, our study provides guidance on how to learn and validate exposure mappings in the presence of interference.[6]

6. For a fully replicable Jupyter notebook illustrating the practical implementation of our method, see GitHub.

References

Nicolas Apfel, Julia Hatamyar, Martin Huber, and Jannis Kueck. Learning control variables and instruments for causal analysis in observational data. arXiv preprint 2407.04448, 2024.

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. The Annals of Applied Statistics, 11:1912–1947, 2017.

Peter M. Aronow, Dean Eckles, Cyrus Samii, and Stephanie Zonszein. Spillover effects in experimental data. arXiv preprint 2001.05444, 2020.

Susan Athey and Stefan Wager. Policy learning with observational data. Econometrica, 89:133–161, 2021.

Seyedeh Baharan Khatami, Harsh Parikh, Haowei Chen, Sudeepa Roy, and Babak Salimi. Graph machine learning based doubly robust estimator for network causal effects. In Proceedings of The 28th International Conference on Artificial Intelligence and Statistics, volume 258 of Proceedings of Machine Learning Research, pages 4366–4374. PMLR, 2025.

Falco J. Bargagli-Stoffi, Costanza Tortù, and Laura Forastiere. Heterogeneous treatment and spillover effects under clustered network interference. arXiv preprint 2008.00707, 2023.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81:608–650, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21:C1–C68, 2018.

D. Cox. Planning of Experiments. Wiley, New York, 1958.

Xavier de Luna and Per Johansson. Testing for the unconfoundedness assumption using an instrumental assumption. Journal of Causal Inference, 2:187–199, 2014.

Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. Journal of Causal Inference, 5:20150021, 2017.

Juan Carlos Escanciano and Telmo Pérez-Izquierdo. Automatic locally robust estimation with generated regressors. arXiv preprint 2301.10643, 2023.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. Advances in neural information processing systems, 30, 2017.

K. Hirano and J.R. Porter. Asymptotics for statistical treatment rules. Econometrica, 77:1683–1701, 2009.

Guanglei Hong and Stephen W Raudenbush. Evaluating kindergarten retention policy. Journal of the American Statistical Association, 101:901–910, 2006.

D. Horvitz and D. Thompson. A generalization of sampling without replacement from a finite population. Journal of American Statistical Association, 47:663–685, 1952.

Martin Huber and Jannis Kueck. Testing the identification of causal effects in observational data. arXiv preprint 2203.15890, 2023.

Michael G Hudgens and M. Elizabeth Halloran. Toward causal inference with interference. Journal of the American Statistical Association, 103:832–842, 2008.

Thomas N. Kipf and Max Welling. Variational graph auto-encoders. arXiv preprint 1611.07308, 2016.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint 1609.02907, 2017.

T Kitagawa and A Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. Econometrica, 86:591–616, 2018.

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

Michael P Leung and Pantelis Loupos. Graph neural networks for causal inference under network confounding. arXiv preprint arXiv:2211.07823v4, 2025.

Shuangning Li and Stefan Wager. Random graph asymptotics for treatment effect estimation under network interference. The Annals of Statistics, 50:2334–2358, 2022.

Jing Ma, Mengting Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. Learning causal effects on hypergraphs. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, page 1202–1212, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539299. URL https://doi.org/10.1145/3534678.3539299.

Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In Arindam Banerjee and Kenji Fukumizu, editors, Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, volume 130 of Proceedings of Machine Learning Research, pages 3700–3708. PMLR, 13–15 Apr 2021.

C F Manski. Statistical treatment rules for heterogeneous populations. Econometrica, 72:1221–1246, 2004.

Charles F. Manski. Identification of treatment response with social interactions. The Econometrics Journal, 16:S1–S23, 2013.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. Statistical Science, Reprint, 5:463–480, 1923.

J Neyman. Optimal asymptotic tests of composite statistical hypotheses, pages 416–444. Wiley, 1959.

Zhewen Pan and Yifan Zhang. Locally robust semiparametric estimation of sample selection models without exclusion restrictions. arXiv preprint 2412.01208, 2024.

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

J M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90:122–129, 1995.

J. M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 90:846–866, 1994.

D. Rubin. Comment on 'randomization analysis of experimental data: The fisher randomization test' by d. basu. Journal of American Statistical Association, 75:591–593, 1980.

D B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66:688–701, 1974.

Fredrik Sävje. Causal inference with misspecified exposure mappings: separating definitions and assumptions. Biometrika, 111(1):1–15, 2024.

Michael E Sobel. What do randomized studies of housing mobility demonstrate? Journal of the American Statistical Association, 101:1398–1407, 2006.

Victor Veitch, Yixin Wang, and David Blei. Using embeddings to correct for unobserved confounding in networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/af1c25e88a9e818f809f6b5d18ca02e2-Paper.pdf.

Hanqing Zeng and Hongkuan Zhou. Srivastava a, et al. graphsaint: Graph sampling based inductive learning method [c]. In The 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, pages 1–19, 2020.

Da Zheng, Chao Ma, Minjie Wang, Jinjing Zhou, Qidong Su, Xiang Song, Quan Gan, Zheng Zhang, and George Karypis. Distdgl: Distributed graph neural network training for billion-scale graphs. In 2020 IEEE/ACM 10th Workshop on Irregular Applications: Architectures and Algorithms (IA3), pages 36–44. IEEE, 2020.

Appendix A. Graph convolutional autoencoder

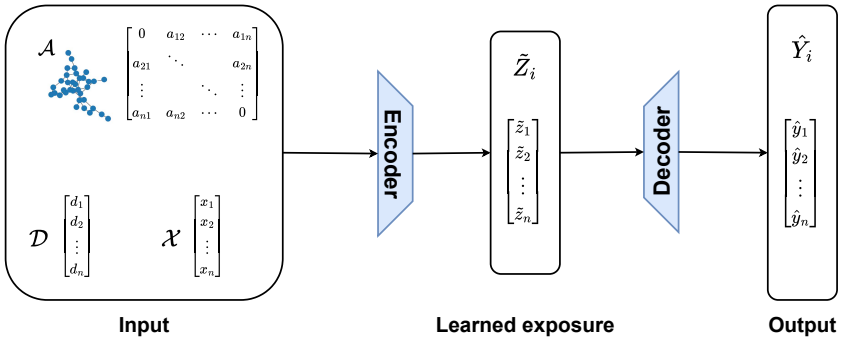


Figure 3: Graph convolutional autoencoder architecture

Appendix B. Details on the testing approach described in Section 4

In this section, we outline the technical details of our testing approach, which builds on the work of Huber and Kueck (2023) and Apfel et al. (2024). We test whether the difference in expected outcomes in Equation (10) equals zero using an aggregated L_2-type measure. Formally, we want to test $H_0 : \theta_0 = 0$ where

$$\theta_0 = (E[Y_i|\dot{Z}_i = \dot{z}_i, \tilde{Z}_i = \tilde{z}_i] - E[Y_i|\dot{Z}_i = \dot{z}_i])^2 + (E[Y_i|\dot{Z}_i = \dot{z}_i, \tilde{Z}_i = \tilde{z}_i] - E[Y_i|\dot{Z}_i = \dot{z}_i]).$$

Given that the learned $\tilde{Z}$ is continuous in the proposed architecture in Section 3, we need to discretize its support. Let $l = 1, \ldots, L$ denote a partition of its support $\tilde{\mathcal{Z}}$ with $\cup_l \tilde{Z}_l = \tilde{\mathcal{Z}}$. Applying the work of Huber and Kueck (2023) to our context, let $\mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l) := E[Y_i|\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l]$, $p_l(\dot{Z}_i) = \Pr(\tilde{Z}_i \in \tilde{Z}_l|\dot{Z}_i)$ and $1(\tilde{Z}_i \in \tilde{Z}_l)$ be an indicator function, which takes the value one if $\tilde{Z}_i$ falls into the partition $\tilde{Z}_l$ (and zero otherwise). Therefore, we test the following null hypothesis H_0:

$$\widetilde{\theta} := E\left[\sum_{l=1}^{L}[(\mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l) - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l))^2 + (\mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l) - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l))]\right] = 0.$$

Apfel et al. (2024) propose the following Neyman-orthogonal score for testing

$$
\begin{aligned}
&\psi(W_i, \theta, \eta) \qquad (12)\\
&= \sum_{l=1}^{L} (\mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l) - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l))^2 \\
&+ \sum_{l=1}^{L} 2(\mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l) - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l)) \\
&\left(\frac{(Y_i - \mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l))1(\tilde{Z}_i \in \tilde{Z}_l)}{p_l(\dot{Z}_i)} - \frac{(Y_i - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l))1(\tilde{Z}_i \notin \tilde{Z}_l)}{1 - p_l(\dot{Z}_i)} \right) \\
&+ \sum_{l=1}^{L} (\mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l) - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l)) \\
&+ \sum_{l=1}^{L} \left(\frac{(Y_i - \mu(\dot{Z}_i, \tilde{Z}_i \in \tilde{Z}_l))1(\tilde{Z}_i \in \tilde{Z}_l)}{p_l(\dot{Z}_i)} - \frac{(Y_i - \mu(\dot{Z}_i, \tilde{Z}_i \notin \tilde{Z}_l))1(\tilde{Z}_i \notin \tilde{Z}_l)}{1 - p_l(\dot{Z}_i)} \right) - \theta
\end{aligned}
$$

and show that using this score function within the DML framework (Chernozhukov et al., 2018), leads to an estimator $\hat{\theta}_0$ which is asymptotically normal and $\sqrt{n}$-consistent under suitable regularity conditions. Especially, we need to ensure that the nuisance functions $\mu(\cdot)$ and $p_l(\cdot)$ can be estimated at rate $o(n^{-1/4})$. Hence, we can use this DML estimator to test whether the researcher-defined exposure is correctly specified, i.e., whether Equation (10) holds true.

Appendix C. Simulation details

C.1. Direct effect estimation

In the second part of the simulation study, we asses the estimation of the direct effect based on the learned exposure $\tilde{Z}_i$ obtained from a GCA. The direct effect is estimated using IPW, where the propensity score for the treatment assignment, $Pr(D_i = 1|X_i)$, is estimated via logistic regression, and the exposure propensity $Pr(Z_i|\mathcal{X}_{-i}, \mathcal{A})$ is approximated using an oracle estimator based on the data-generating process. The true exposure that the GCA aims to recover is defined as $Z_i = I\{\sum_{j \neq i} a_{ij} D_j X_j > 2\}$ and we adjust the parameter for generating a_{ij} to $r_n = \left(\frac{5}{\pi n}\right)^{(1/2)}$ in the DGP outlined in the main text.

n	est	std	bias
100	1.178	1.047	0.843
200	1.128	0.684	0.542
500	1.007	0.396	0.314
1000	1.005	0.306	0.241

Notes: Average estimate of γ (est), its standard deviation (std), and the average absolute estimation error (bias) across $R = 200$ replications. The true value of the direct effect is $\gamma = 1$.

Table 2: Simulation results - direct effect estimation

Table 2 reports the simulation results for the estimation of the direct effect γ based on the learned exposure $\tilde{Z}_i$. For small sample sizes, the estimator shows noticeable variability and bias. As the sample size increases, both the bias and the standard deviation decrease. For sample sizes $n = 500$ and $n = 1000$, the average estimated direct effect is close to the true value $\gamma = 1$. This indicates improved precision and convergence toward the true direct effect.

C.2. Additional testing method simulation results

We additionally consider a variant of Setting 2 in which the researcher-defined exposure mapping coincides with the first-order component of the true exposure, while the true exposure also includes a second-order neighborhood term, i.e.,

$$Z_i = \frac{\sum_{j\neq i} a_{ij} D_j X_j}{\sum_{j\neq i} a_{ij}} + \frac{\sum_{k\neq i} b_{ik} D_k X_k}{\sum_{k\neq i} b_{ik}}, \text{ where } b_{ik} := I\{\sum_{i\neq j, j\neq k} a_{ij} a_{jk} > 0\} \cdot (1 - a_{ik}) \cdot I\{k \neq i\}.$$

The researcher-defined exposure represents the share of treated neighbors weighted by their covariates, i.e., $\dot{Z}_i = \frac{\sum_{j\neq i} a_{ij} D_j X_j}{\sum_{j\neq i} a_{ij}}$. Thus, the researcher-defined exposure is misspecified. The DGP of $\mathcal{A}$ is $a_{ij} = I\{|\rho_i - \rho_j| \leq r_n\}$ with $\rho_i \sim U([0,1])^2$ and $r_n = \left(\frac{\kappa}{\pi n}\right)^{(1/2)}$. In the simulation study in Section 6, we set $\kappa = 30$. By varying the network density parameter $\kappa \in \{3, 15, 30\}$, we control the extent and variability of the second-order neighborhood component and thus the deviation between the true exposure mapping Z_i and the simpler researcher-defined mapping $\dot{Z}_i$. In our design, smaller values of κ lead to stronger variability of the second-order component due to a larger set of second-degree neighbors that are not directly connected, while for larger κ many second-degree neighbors also become directly connected, which reduces the additional second-order variation. This variability directly affects the degree of violation of the testable implication in Equation (10) and consequently the performance of the proposed testing procedure, as reflected in the rejection rates and mean p-values reported in Tables 3-5, where the third column in each table reports the standard deviation of the difference between the researcher-defined and learned exposure mappings, which quantifies the variability of the second-order component. Equation (10) requires that the learned exposure mapping $\tilde{Z}$ does not provide additional explanatory power for the outcome once conditioning on the researcher-defined mapping $\dot{Z}$. When the second order-component is weak, this condition is approximately satisfied, which leads to large p-values and low rejection rates. As the variability of the second-order component increases, the conditional independence is increasingly violated. This results in smaller p-values and higher rejection rates.

	Setting A ($\kappa = 30$)		
n	rejection rate	mean p-value	std of $\dot{Z} - \tilde{Z}$
500	0.0	0.66	0.131
1000	0.01	0.58	0.126
2000	0.03	0.48	0.122

Notes: 'rejection rate' gives the empirical rejection rate when setting the level of statistical significance to 0.05 (or 5%); R = 200 replications. 'std of $\dot{Z} - \tilde{Z}$' reports the average (across $R = 200$ replications) of the standard deviation of the difference between the researcher-defined and the learned exposure mapping.

Table 3: Simulation results for Setting A - testing method

	Setting B ($\kappa = 15$)		
n	rejection rate	mean p-value	std of $\dot{Z} - \tilde{Z}$
500	0.13	0.32	0.175
1000	0.30	0.24	0.169
2000	0.55	0.16	0.173

Notes: 'rejection rate' gives the empirical rejection rate when setting the level of statistical significance to 0.05 (or 5%); R = 200 replications. 'std of $\dot{Z} - \tilde{Z}$' reports the average (across $R = 200$ replications) of the standard deviation of the difference between the researcher-defined and the learned exposure mapping.

Table 4: Simulation results for Setting B - testing method

	Setting C ($\kappa = 3$)		
n	rejection rate	mean p-value	std of $\dot{Z} - \tilde{Z}$
500	0.92	0.03	0.400
1000	0.97	0.01	0.400
2000	0.97	0.02	0.396

Notes: 'rejection rate' gives the empirical rejection rate when setting the level of statistical significance to 0.05 (or 5%); R = 200 replications. 'std of $\dot{Z} - \tilde{Z}$' reports the average (across $R = 200$ replications) of the standard deviation of the difference between the researcher-defined and the learned exposure mapping.

Table 5: Simulation results for Setting C - testing method