Using Embeddings for Causal Estimation of Peer Influence in Social Networks

Irina Cristali[1] Victor Veitch[1,2]

[1]*Department of Statistics, The University of Chicago*
[2]*Google Research*

Abstract

We address the problem of using observational data to estimate peer contagion effects, the influence of treatments applied to individuals in a network on the outcomes of their neighbors. A main challenge to such estimation is that *homophily*—the tendency of connected units to share similar latent traits—acts as an unobserved confounder for contagion effects. Informally, it's hard to tell whether your friends have similar outcomes because they were influenced by your treatment, or whether it's due to some common trait that caused you to be friends in the first place. Because these common causes are not usually directly observed, they cannot be simply adjusted for. We describe an approach to perform the required adjustment using node embeddings learned from the network itself. The main aim is to perform this adjustment nonparametrically, without functional form assumptions on either the process that generated the network or the treatment assignment and outcome processes. The key contributions are to nonparametrically formalize the causal effect in a way that accounts for homophily, and to show how embedding methods can be used to identify and estimate this effect. Code is available at https://github.com/IrinaCristali/Peer-Contagion-on-Networks.

1 Introduction

We are interested in estimating peer effects, the causal influence of individuals on their neighbors.

Example. We want to infer the effect that social pressure has on vaccination. Suppose we observe data from a population where each unit i is a person in an interconnected social network, and for each unit we know whether they were vaccinated at the beginning of the study period, T_i, and whether they were vaccinated at the end of the study period, Y_i. We are interested in estimating the effects of the treatment T_i of person i on the outcome Y_j of person j. In addition to their vaccination status, each unit has attributes C_i that act as (proxies for) causes of both the particular network ties they form, and their vaccination behavior. For instance, C_i may include age, race, education status, income level, political affiliation, and so forth.

The core challenge here is that we want to estimate a causal effect (e.g., what would happen if we intervened by vaccinating popular people?), but the variables C_i, C_j act as confounders between the treatment T_j and outcome Y_i. The reason is that when defining a contagion effect from j to i we must condition on the presence of an edge between (i, j). The edge is causally influenced by both C_i and C_j, so the conditioning creates a dependency between these variables (it acts as a collider). For example, if we learn that Alice and Bob are friends then we can infer that they likely live in the same city. Accordingly, the association between T_j and Y_i may be either due to *contagion*: the causal influence of T_j on Y_i (Bob got vaccinated because Alice did), or due to *homophily*: the tendency of similar people to be connected in a network (e.g., Alice and Bob got vaccinated because they both

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

live in a major city, and it's only due to chance that Alice got vaccinated before Bob). In general, homophily is confounded with contagion [ST11].

Now, if we observed the attributes C, we could formalize the contagion effect using standard causal tools and then identify the effect from observational data by adjusting for C [EKB16; EB17; BDF20]. However, often such detailed knowledge of C is unavailable. In this case straightforward causal identification and estimation is not possible.

In such situations, we might hope to make use of the following intuitive observation. The pattern of network ties itself carries information about each C_i. Indeed, this is the root problem that causes the confounding issue. Accordingly, we might hope to use the network itself to obtain estimates $\hat{C}_i$ for the attributes of each node. Then, we could adjust for the estimated $\hat{C}_i$ in some suitable causal estimation procedure. The aim of this paper is to clarify this type of procedure.

Shalizi and McFowland III [SM16] provide an estimation strategy of this kind. They first assume that the network is generated by either a stochastic block model [HLL83; LW19] or a latent space model [HRH02; RFR16]. They assume the attributes C correspond to the latent community identities or latent space positions. They then further assume that the outcome of each unit is defined by a particular linear structural equation, which includes a term for both the average treatment of that node's neighbors, and a term for the effect of the attributes C_i. Their procedure is to estimate C using the assumed network model, and then use the estimated $\hat{C}$ in a linear regression to determine the coefficient of the average-neighbor-treatment term.

This approach, however, critically relies on the assumed parametric form of both the network model and the outcome model. Indeed, even the target of estimation is defined as a parameter of this assumed model. Accordingly, when the parametric network model or linear outcome model is misspecified, the meaning of both the estimand and estimation procedure becomes unclear.

The goal of this paper is to develop a non-parametric procedure in this vein. That is, the aim is causal estimation of contagion effects by adjusting for network-inferred attributes without relying on detailed parametric assumptions. Towards this end, the paper follows three main steps:

1. Formalize the target causal effect non-parametrically. The main challenge is that the estimand must depend on the network we are working with (because contagion requires knowing who is friends with whom) and the network must itself be modeled as a random variable which is a function of the unobserved confounders C (to accommodate homophily).

2. Derive sufficient conditions for the estimated attributes to yield causal identification. The idea is that it is not necessary to exactly reconstruct C, but only extract the minimum information that will identify the causal effect—this turns out to be (plausibly) a much easier task.

3. Give a concrete method for contagion estimation using node embedding techniques to extract the information from the network that is relevant for predicting peer influence, and illustrate the practical performance of this technique.

2 Setup

Consider a network G_n of n individuals, where connections between people are encoded through undirected edges between nodes. We define the degree of a node as the number of connections it has. The neighbors of node i are the nodes with which i has ties. Each such link is captured by the network adjacency matrix A, where $A_{ij} = \mathbb{1}_{\{\text{i and j share a tie}\}}$. We take $A_{ii} = 0$ for all i. We also consider a vector of variables associated with each node: (Y_i, C_i, T_i). Here, Y_i is the observed outcome, T_i is the treatment, and C_i are (unobserved) attributes that may causally influence Y, T and A.

The data is generated according to the following structural equation model, adapted from Ogburn et al. [Ogb+17] to model the network as a random variable.[1]

[1] We simplify by removing the influence of neighbors' covariates on outcomes and treatments.

$$
\begin{aligned}
C_i &\leftarrow f_C[\varepsilon_{C_i}]; \\
A_{ij} &\leftarrow f_A[\{C_i\}_i, \varepsilon_{ij}]; \\
T_i &\leftarrow f_T[C_i, \varepsilon_{T_i}]; \\
Y_i &\leftarrow f_Y[S_Y(\{T_j : A_{ij} = 1\}), C_i, \varepsilon_{Y_i}]
\end{aligned} \tag{2.1}
$$

The ε variables represent exogenous noise, which we take to be identically distributed and independent of the network and of each other. The function S_Y summarizes the neighbors' treatment—e.g., S_Y could be the average function, or the logical OR function. This means that the treatment assignment for each node i depends only on its attributes C_i, the outcome assignment depends on the node attributes C_i and the treatments of all its neighbors, and the network structure depends on all C. The structural functions f are fixed but unknown.

3 Formalizing the causal estimand

Consider setting $T_i \leftarrow t_i^*$ for each unit i. Our aim is to formalize the idea of "average influence of a node's neighbors' treatments on its outcome".

An intuitive choice is

$$
\psi_{t^*}^{\text{full info}} := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i | \text{do}(T = t^*), \{C_i\}_i, G_n]. \tag{3.1}
$$

Here, do is Pearl's do notation, and indicates that the treatment (vector) of all nodes is intervened on and set to t^*. Note that, following the structural model in equation 2.1, nodes' outcomes are only affected by their neighbor's treatments. Accordingly, the estimand is the average outcome we would have seen had the treatment assignment been set to $T \leftarrow t^*$, keeping both the node attributes $\{C_i\}$ and the network G_n fixed. The interpretation of this effect is: the average outcome under the hypothetical treatment, applied to the *same set of people* connected by the *same link structure*.

Unfortunately, this estimand will not usually be identifiable. The reason is that the node attributes C are unobserved; e.g., we don't observe city of residency, and we cannot perfectly reconstruct it from the network structure. To circumvent this, we instead define the version of this estimand that we can (hope to) identify from the graph alone:

$$
\psi_{t^*} := \frac{1}{n} \sum_{i} \mathbb{E}[Y_i \mid \text{do}(T = t^*), G_n]. \tag{3.2}
$$

This new estimand can be understood as taking the estimand in (3.1) and marginalizing out the information about C that cannot be inferred from G_n. It is the average outcome under the hypothetical treatment, applied to the *same link structure* and to *a set of people consistent with the link structure*. In other words, it represents the peer contagion effect on the same graph with node properties only fixed to be consistent with the graph structure. Although this is somewhat less natural, we will see that it makes identification plausible.

There is an apparent significant drawback of the formalizations presented in Equations (3.1) and (3.2): both estimands are fundamentally tied to the particular sample available in our study, since both involve conditioning on the observed network G_n. We may be nervous that the formal quantity is tied to idiosyncracies of the particular observed network. In Theorem 1, we show that kind of sample dependency is not a serious issue. The approach is to show that both the causal estimands $\psi_{t^*}^{\text{full info}}$ and ψ_{t^*} introduced in Equations (3.1) and (3.2) converge in probability to a fixed quantity as the size of the network G_n increases. See Appendix A for the proof.

Theorem 1 (Generalizing the sample-based causal estimands)**.** *Consider an observed social network G_n, and let Y_i be the labels associated with each node i. Assume the following*

(i) There exists $M > 0$ such that Var$(Y_i) \leq M$, for all i;

(ii) For any pair (i, j) of nodes selected uniformly at random, $\mathbb{P}(\{$ i and j share a neighbor$\}) \to 0$, as $n \to \infty$.

Then, there exists a fixed real number ψ such that $\psi_{t^*}^{\text{full info}} \to \psi$ and $\psi_{t^*} \to \psi$, in probability.

This theorem establishes that, provided the graph is sparse in a suitable sense, then both causal estimands converge to the same fixed quantity which does not depend on the particular network G_n. This decouples the interpretation of the causal estimand from the particular network under consideration. Additionally, it shows that the distinction between $\psi_{t^*}^{\text{full info}}$ and ψ_{t^*} is not so important for practical applications—for large networks, the two values will anyway be very close. Note that this result is about proving that the *estimand* is well-defined, not about establishing convergence of some estimator.

4 Causal inference of peer effects using node embedding methods

Having formalized and motivated the target causal effect of interest, we now turn to sufficient conditions for causally identifying it.

As previously discussed, if we could exactly reconstruct the latent C_i then identification would be straightforward. However, usually, this reconstruction is impossible. Happily, it is not necessary: an imprecise proxy for C_i might still suffice for causal identification. Our goal now is to find sufficient conditions for an inferred proxy to enable identification. The next theorem gives such a condition. Informally, the condition is that we need a proxy λ_i that carries enough information about C_i so that the graph itself carries no further information about Y_i after conditioning on λ_i. As we shall discuss in Section 5, node embedding methods can be viewed as a tool for constructing such λ_i.

We first define the following object, which plays a substantial role in the remainder of the paper.

Definition 1. *Let $V_i := S_Y(\{T_j : A_{ij} = 1\})$[2] be the aggregated treatment at node i, and v_i^* its value under the hypothetical treatment intervention $T = t^*$. For a graph G_n, the* vertex conditional outcome model *is the function m given by $m_{G_n}(v_i^*, \lambda_i) := \mathbb{E}[Y_i \mid v_i^*, \lambda_i]$.*

Theorem 2 (Sufficient conditions for identification)**.** *Let G_n be the network, A its adjacency matrix, and Y_i the node-associated labels. Suppose that for each node i we have a proxy for the latent attributes, $\lambda_i \in \mathbb{R}^k$, such that*

i $Y_i \perp\!\!\!\perp A_{ij} | (\lambda_i, v_i^*)$, *for all i and j.*
ii $P(V_i = v^* \mid \lambda_i) > 0$ *for all v^*;*
iii λ_i *is C_i-measurable.*

Then $\mathbb{E}[Y_i | \text{do}(T = t^*), G_n, \lambda_i] = m_{G_n}(v_i^*, \lambda_i)$.

Proof. We sketch the proof and provide the full mathematical justification in Appendix B. Consider Figure 1 illustrating the core identification issue. The node-level causal effect $\mathbb{E}[Y_i | \text{do}(T = t^*), G_n, \lambda_i]$ relies on conditioning on the network structure, and, in particular, on the edges A_{ij}. Doing so opens a backdoor path between Y_i and T_j (hence, also V_i) going through the unobserved C_i. The main idea of the theorem is that conditioning on λ_i removes the effect of conditioning on the A_{ij}, and thus prevents the backdoor path from being opened.

□

In Theorem 2, condition (i) gives a notion of "sufficient amount of information about C" for causal adjustment, and is the main point of the theorem. Condition (ii) is the standard positivity assumption required for causal identification. Finally, condition (iii) is the condition that λ_i is a partial reconstruction

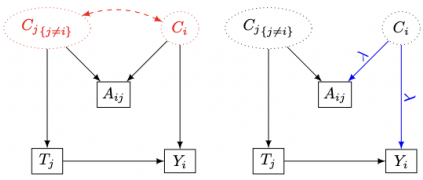

Figure 1: Identification of causal peer effects using embeddings.

[2]In general, the function S_Y is set by the analyst for practical estimation purposes; this paper assumes S_Y is the mean function.

of the node-specific attributes C_i. This condition is necessary to ensure λ_i is not itself a collider for the latent node attributes.

This result readily extends to a condition for the identification of the peer contagion estimand $\psi_{t^*} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i|\mathrm{do}(T=t^*), G_n]$ introduced in Section 3. This is the content of the next result; see Appendix C for a proof.

Corollary 3 (Identification of the causal estimand ψ_{t^*})**.** *Consider the same set-up and notation as that of Theorem 2. Then, the causal estimand ψ_{t^*} can be identified as follows*

$$\psi_{t^*} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[m_{G_n}(v_i^*, \lambda_i)|G_n]. \tag{4.1}$$

Corollary 3 gives, in principle, a strategy for identifying the causal effect using the observed λ. Namely, learn $m_{G_n}(v_i^*, \lambda_i)$ and $\mathrm{P}(\lambda_i \mid G_n)$ and plug in to (4.1). This result is exact, and applies to any network. However, learning $\mathrm{P}(\lambda_i \mid G_n)$ seems difficult—it's unclear how to model this distribution. To circumvent the need to learn the unit-wise uncertainty about λ_i, we now give another result that again relies on the validity of the law of large numbers derived in Theorem 1. Intuitively, when the law of large numbers holds, we can replace averaging over the per-node uncertainty by averaging over all of the nodes (that is, the mean converges to the unit-level expectation).

We now state the main identification result as Theorem 4 below which is derived from both Theorem 2 and Theorem 1. We include its proof in Appendix D.

Theorem 4. *Consider the same set-up as that of Theorem 2. Furthermore, assume that the two conditions of Theorem 1 are also satisfied. Then $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} m_{G_n}(v_i^*, \lambda_i) = \psi$, in probability, where ψ is the fixed real number such that $\lim_{n\to\infty}\psi_{t^*}^{\text{full info}}$ and $\lim_{n\to\infty}\psi_{t^*} = \psi$, in probability.*

5 Estimation of causal peer influence effects

We have now established a formal causal estimand ψ_{t^*}, and conditions on node-attribute proxies that enable identification.The basis for our estimation procedure will be Theorem 4, which says that—for suitable λ and sufficiently sparse G_n—the statistical estimand

$$\widetilde{\psi}_{t^*} := \frac{1}{n}\sum_{i=1}^{n} m_{G_n}(v_i^*, \lambda_i) \tag{5.1}$$

approximates the causal estimand ψ_{t^*}, with the approximation becoming exact as $n \to \infty$. Recall $m_{G_n}(v_i^*, \lambda_i) := \mathbb{E}[Y_i \mid v_i^*, \lambda_i]$. Our task is to estimate $\widetilde{\psi}_{t^*}$ from the observed data using the observed treatments and outcome $(T_i, Y_i)_{i=1:n}$ and the link structure G_n.

To achieve this, we use embedding-based semi-supervised prediction models to learn $\hat{\lambda} \in \mathbb{R}^k$ and $\hat{m}_{G_n}(v^*, \hat{\lambda}_i) \approx \mathbb{E}[Y_i \mid v_i^*, \hat{\lambda}_i]$. Then, we plug these estimates into (5.1) as m and λ. For concreteness, we describe a particular approach based on Veitch et al. [Vei+19] and Veitch et al. [VWB19]. The estimation procedure follows three main steps:

Step 1. We train a model using relational empirical risk minimization [Vei+19] to jointly learn the embeddings $\hat{\lambda}_i$ and $\hat{m}_{G_n}(v_i, \hat{\lambda}_i)$. To do this, we first compute $v_i = S_Y(\{t_j : A_{ij} = 1\})$ at each vertex i—i.e., we create a new aggregate treatment feature and attach it to each vertex.

Relational ERM works by minimizing a risk defined as an expectation over randomly sampled subgraphs of the larger network. Let $\texttt{Sample}(G_n, k)$ be a sampling algorithm that returns a random subgraph of size k from G_n (e.g., the subgraph induced by a random walk of length k). Let v_{G_k} be the set of vertices of the subgraph G_k. We then define the loss function on the G_k to be

$$L(G_k, \lambda, \gamma) = q \cdot \sum_{i \in v_{G_k}} (y_i - m(v_i, \lambda_i; \gamma))^2 + \sum_{i,j \in v_{G_k} \times v_{G_k}} \mathrm{CrossEntropy}(A_{ij}, \sigma(\lambda_i^T \lambda_j)), \tag{5.2}$$

where $q \in [0, 1]$ and σ is the sigmoid function. The second term is a network reconstruction term that extracts node-level information from the network by requiring the embeddings to be predictive

of the edge structure. The first term learns a predictor m, parameterized by γ (global parameters shared across the network), that predicts y_i from the aggregated treatment v_i and the embedding λ_i. In our experiments, m will be a linear predictor, and γ the (parameters of the) linear map. When y is binary, the binary cross-entropy loss is used instead of the MSE loss. Although the map from the embeddings is linear, because the embeddings are unconstrained, the combined embedding-linear map can flexibly represent complex relationships between the network structure and outcome/treatment. In Section 6 we see that the model ably captures a relationship between Y and a latent attribute C that is non-linear in C, where the relationship between the network and C is complex.

Then, we train the model by fitting

$$\hat{\lambda}, \hat{\gamma} = \underset{\lambda,\gamma}{\operatorname{argmin}}\, \mathbb{E}_{G_k = \texttt{Sample}(G_n,k)}[L(G_k, \lambda, \gamma) | G_n]. \tag{5.3}$$

Step 2. We define $\hat{m}_{G_n}(v_i^*, \lambda_i) = m(v_i^*, \lambda_i; \hat{\gamma})$, where v_i^* is the value of v_i attained under the interventional treatment $T = t^*$. Then, for each unit vertex i in G_n we compute $\hat{m}_{G_n}(v_i^*, \hat{\lambda}_i)$.

Step 3. Finally, the peer influence estimate is the average over the estimated vertex conditional outcomes $\hat{m}_{G_n}(v_i^*, \hat{\lambda}_i)$:

$$\hat{\psi}_{t^*}(G_n) = \frac{1}{n} \sum_i \hat{m}_{G_n}(v_i^*, \hat{\lambda}_i).$$

5.1 Validity of the estimation procedure

In this section, we informally discuss the validity of the estimation procedure described in the previous section. More specifically, recall that, in Theorem 1, we showed that both causal peer influence estimands proposed, $\psi_{t^*}^{\text{full info}}$ and ψ_{t^*}, converge, in probability, to a fixed real number ψ, as $n \to \infty$. The question that remains is whether the estimator proposed, $\hat{\psi}_{t^*} = 1/n \sum_{i=1}^n \hat{m}_{G_n}(v_i^*, \hat{\lambda}_i)$, also converges to ψ, in probability, as $n \to \infty$. This will clearly hold as long as $\hat{m}_{G_n}(v_i^*, \hat{\lambda}_i) \to m_{G_n}(v_i^*, \lambda_i)$ for some λ satisfying the conditions of Theorem 2.

Why should one expect the above consistency property? First, consider $\hat{m}(v, \lambda) \to \mathbb{E}[Y_i \mid v, \lambda]$. The map m is defined as the minimizer of the squared error empirical risk. The minimizer of the true risk is the required conditional expectation. Accordingly, this part of the learning is, essentially, the same as what occurs with non-network data and is plausible.

The trickier judgement is why can one expect that $\hat{\lambda}_i \approx \lambda_i$ for some λ satisfying the conditions of Theorem 2? That is, is it plausible that the learned embeddings behave as proxies for the node-level attributes (i.e., are C-measurable), and carry sufficient information to adjust for the homophily effect (i.e., $Y \perp\!\!\!\perp A_{ij} \mid (\lambda_i, v_i^*)$)?

The C_i-measurability is just the requirement that each node-level embedding captures unobserved information specific to that node. This is exactly the aim of embedding methods, to leverage network structure in order to capture the latent properties of each node [HYL17]. Accordingly, it is reasonable to assume that the embedding-based method described in Section 5 satisfies this property (at least, asymptotically). With respect to whether the embeddings carry sufficient information to adjust for the homophily effect when estimating peer influence, we note the following: in order to satisfy this condition, it suffices that the embeddings are predictive of the outcome or of the edge structure. To see why this is plausible, recall the expression of (5.3) used to fit the embeddings. The cross-entropy term encourages the embeddings to be sufficient (at the graph level) for predicting the edge structure. The squared error term encourages the embeddings to be sufficient for predicting the labels Y_i. Hence, the objective function is aimed at exactly the requisite information.

Unfortunately, despite the intuitive justification above, there is no firm guarantee that any given embedding method actually achieves its conceptual goal. In effect, by working with embedding methods one is trading off the precise guarantees enabled by parametric models (exploited in [SM16]) in favor of techniques that have good empirical performance even in the (typical) case that the networks have structure that is inconsistent with parametric specification.

In fact, for the specific case of relational ERM we can say something further. Davison and Austern [DA21] study the asymptotics of the embedding learning procedure under the assumption that G_n is

Table 1: The embedding-based estimator $\hat{\psi}_{t^*}$ effectively adjusts for confounding and recovers the true treatment effect. The ground truth value of peer contagion is 1. Zero, low, and high confounding levels correspond to $\beta_1 = 0, 1$, and 10, respectively. For $\hat{\psi}_{t^*}$, the reported values represent the mean over 100 different global random seeds. The seed for the simulated treatment and outcome data is kept constant. For the Unadjusted and Parametric estimators, the reported values represent the respective estimated regression coefficients for the aggregated treatment used when predicting Y.

	district			**age**			**join_date**		
Conf. level	**Zero**	**Low**	**High**	**Zero**	**Low**	**High**	**Zero**	**Low**	**High**
Unadjusted	0.99	1.64	7.40	1.00	1.39	4.90	0.99	1.38	4.81
Parametric	0.99	1.41	5.28	1.00	1.33	4.20	0.98	1.28	4.00
$\hat{\psi}_{t^*}$	0.84	0.96	1.17	0.94	0.94	1.11	1.01	1.03	1.10

generated by an exchangeable random graph, or a graphon (a broad family of models that includes, e.g., stochastic block models and latent space models [Bor+16; CCB16; CF17; CD18; Jan18]). Informally, Davison and Austern [DA21] show that, asymptotically, the embeddings are functions of node-specific attributes C_i alone and that these embeddings capture the limiting distribution of the network (see Davison and Austern [DA21] Thm. 7). Accordingly, in this case, the learned embeddings satisfy the conditions of Theorem 2.

6 Experiments

To study the practical performance of the estimation procedure described in Section 5, we conduct experiments using semi-synthetic data. We use a subset of data from the Slovakian social media website *Pokec* [TZ12; LK14]. The data includes both links between users and limited node-level covariates. Our basic strategy will be to use these node level covariates as the latent C_i's. To do this, we simulate treatment and outcomes for each unit in a manner that depends on C_i. At estimation time, we hide the C_i variables, and check how well we can recover the true causal effect using only the graph data and observed treatments and outcomes.

We extend the setup of Veitch et al. [VWB19] who applied a similar relational empirical risk minimization based technique in order to adjust for the network unobserved confounding effect when performing average treatment effect (ATE) estimation of T_i on Y_i. That work assumes pure homophily, whereas this paper examines peer-based contagion effect estimation.

Following Veitch et al. [VWB19] we analyze a sub-network of approximately 70000 users connected by roughly 1.3 million links, representing a subset of the original Pokec data restricted to the districts Žilina, Cadca, and Namestovo, from the same region. The analysis consists of two parts:

1. estimating the peer contagion effect of binary treatments on continuous response variables, varying both the source of confounding and its strength;
2. estimating the peer contagion effect of binary treatments assigned at time t_1, the beginning of a study period, on treatments assigned at a subsequent time t_2, the end of that study period.

See supplement (Appendix F) for additional experimental details.

6.1 Continuous outcome

For the first set of experiments, the treatment T and outcome Y values are generated from three variables (one at a time), taken as the unobserved confounders C: district, age, and Pokec join date. For each of these three confounders, we standardize, then bin each of them into three possible values: $-1, 0$, and 1. We then apply a function g which transforms them into probabilities: $g(C) = 0.5 + 0.35 \cdot C$, so that $g(C) \in \{0.15, 0.5, 0.85\}$. Then, we simulate according to:

$$\begin{aligned} T_i &= \text{Bern}(g(C_i)); \\ V_i &= \text{Average}(T_j : A_{ij} = 1); \\ Y_i &= \beta_0 \cdot V_i + \beta_1 \cdot g(C_i) + \varepsilon_i,\ \varepsilon_i \sim N(0,1). \end{aligned} \tag{6.1}$$

The parameter β_0 controls the amount of peer influence, while β_1 measures the amount of network confounding. In this simulation setup, the causal peer contagion effect of interest, $\hat{\psi}_{t^*}$ is precisely equal to β_0. We fix $\beta_0 = 1$ as the *ground truth* value of confounding, and let β_1 vary in $\{0, 1, 10\}$ corresponding to no, low, and high contagion, respectively.

Since we only observe T_i, V_i and Y_i we test whether $\hat{\psi}_{t^*}$, the estimator proposed in Section 5, adjusts for the unobserved confounding due to C, and accurately estimates $\beta_0 = 1$. For each combination of C, and strength level β_1, we obtain simulated V_i and Y_i values, then we train the embedding-based relational empirical risk minimization model described in Section 5, using a random walk sampler with negative sampling, and the default settings of Veitch et al. [VWB19]. Finally, we use the trained model to predict two sets of adjusted Y values:

1. Y_i's resulting from the aggregated treatment v_i^*, when all T_i are set to 0, i.e. $\hat{m}(\hat{\lambda}_i, v^*_{\{i,t_i^*=0\}})$;
2. Y_i's resulting from the aggregated treatment v_i^*, when all T_i are set to 1, i.e. $\hat{m}(\hat{\lambda}_i, v^*_{\{i,t_i^*=1\}})$.

The causal effect estimator of interest is then $1/n \sum_{i=1}^n \hat{m}(\hat{\lambda}_i, v^*_{\{i,t_i^*=1\}}) - 1/n \sum_{i=1}^n \hat{m}(\hat{\lambda}_i, v_{\{i,t_i^*=0\}})$.

We compare the results of this procedure against those of two baseline estimators. The first one is the naive estimator which does not adjust for confounding, namely $1/n \sum_i \mathbb{E}[Y_i | v^*_{\{i,t_i^*=1\}}] - 1/n \sum_i \mathbb{E}[Y_i | v^*_{\{i,t_i^*=0\}}]$, where expectations are the estimated values obtained via linear regression. The second baseline seeks to mimic the parametric peer contagion estimation procedure of Shalizi and McFowland III [SM16], by fitting a mixed-membership stochastic block model to the Pokec data, inferring 128 latent communities (128 is equal to the dimension of the embeddings), then predicting Y via a linear regression on the aggregated treatment and inferred latent community values.

Table 1 summarizes the results of the embedding method compared to those of the above baselines. Whenever confounding was present, $\hat{\psi}_{t^*}$ yielded the most accurate results out of all three estimators, by a significant margin. This confirms that the nonparametric method can leverage the network structure and obtain reliable peer contagion estimates. With no confounding present, all three estimators produced valid results, yet, when district and age were used as unobserved confounders, $\hat{\psi}_{t^*}$ was slightly less accurate compared to the baselines. Indeed, if there is no unobserved confounding, then no network-based adjustment is necessary; the variation due to embedding-model fitting causes errors that simpler methods do not suffer from. In Appendix E, Table 3, we provide error bars for the average estimated peer effects obtained in Table 1, which aim to illustrate the consistency of our results. The errors reported for $\hat{\psi}_{t^*}$ represent the standard errors computed over 100 different global random seeds, while the seed for the simulated treatment and outcome data is kept constant. For the Unadjusted and Parametric estimators, the reported error bars represent the standard errors of the estimated regression coefficients for the aggregated treatment used when predicting Y. The main caveat is that due to the relational, non-i.i.d. structure of our data, these errors do not represent proper confidence bands, and should be interpreted with caution. To obtain valid confidence intervals, proper asymptotic normality results for the studied estimators are necessary.

6.2 Binary outcome: a model for peer influence and vaccination

Recall the motivating example in Section 1: some proportion of the observed population gets vaccinated at time t_1, while others get vaccinated at a subsequent time t_2. We now illustrate how to use node embeddings to study the peer influence of the former group on the latter. We use the following data setup:

1. Simulate the treatment by the same procedure as in Section 6.1, i.e. as a function of the unobserved confounders;
2. Randomly select 50% of the vertices, and set $Y_i = T_i$, but then delete T_i for those respective nodes. The uncensored treatment values T_i correspond to individuals who were vaccinated at t_1, while the Y_i values represent individuals vaccinated at t_2;
3. Compute the causal effect of V_i on Y_i only for the vertices where Y_i is defined.

The above design matches the following scenario: whether or not an individual gets vaccinated depends on latent attributes such as their age, or location. Since $Y_i = T_i$, there is no causal peer

Table 2: The embedding-based estimator $\hat{\psi}_{t^*}$ accurately recovers the true peer contagion effect of binary treatments on other subsequent binary treatments. The ground truth peer contagion value is 0. For $\hat{\psi}_{t^*}$, the reported values represent the mean over 100 different global random seeds. For the Unadjusted and Parametric estimators, the reported values represent the respective estimated regression coefficients for the aggregated treatment used when predicting Y.

Peer influence on vaccination	**district**	**age**	**join_date**
Unadjusted	2.03	0.12	0.68
Parametric	1.30	1.03	0.98
$\hat{\psi}_{t^*}$	0.09	0.11	0.22

influence effect of the neighbors j on node i. Despite the lack of a causal relationship, an association is clearly present due to the dependence on latent confounders. The question is whether the proposed method can adjust for the unobserved causes and render the true causal effect.

Using the same baseline estimators, and an analysis similar to that in Section 6.1, we obtain estimates for the peer contagion effect, summarized in Table 2. We note that, for all three hidden confounders, the embedding-based method produces the estimates closest to the ground truth value of 0. For the "age" covariate, the naive, unadjusted estimate is very close to the embedding-based one, showing that age may not be the optimal predictor for network tie formation, preventing the non-parametric method from appropriately adjusting for the network structure, and yielding our method similar to the naive one, in this situation. We also note that, when registration date is used as a hidden confounder, all three estimates are less accurate, suggesting that "join_date" may also not be the best predictor of the network edges. Nonetheless, in this scenario, the nonparametric estimator still has superior performance over the two baselines. In Appendix E, Table 4, we provide error bars for the variation induced by the choice of random seed, obtained in a similar way to those corresponding to the continuous outcome scenario.

In Appendix G we perform additional analyses further varying the values of β_1 and noise level ε, as well as illustrate the performance of our method on the Wikipedia hyperlink network [Yin+17].

7 Related work

There is an active literature on estimating causal effects from networked data [ZA21].

Randomized controlled trials. To estimate causal effects under interference and network confounding, several recent works rely on designing randomized experiments. Toulis and Kao [TK13] use potential outcomes and a sequential randomization design to estimate peer influence on observed networks. Eckles et al. [EKB16] develop an experimental design to study the impact of peer feedback on the behavior of Facebook users. Fatemi and Zheleva [FZ20b] propose an experimental design for estimating the effect of the treatment alone on the unit's outcome, isolating it from peer effects, while Fatemi and Zheleva [FZ20a] discuss an approach of minimizing selection bias when performing A/B testing on networks.

Causal network inference from observational data. Within observational settings, van der Laan [van14], Ogburn et al. [Ogb+17], Tchetgen Tchetgen et al. [TFS17], and Tran and Zheleva [TZ22] tackle the issue of causally estimating social contagion from observed networked data, yet they assume there are no latent sources of network confounding which affect both the treatment and the outcome, by considering all node-related features as known. This is a significant limitation which this paper seeks to address.

Other works which account for latent confounding assume that the network (and, implicitly, the unobserved confounders) can be represented using parametric methods. Shalizi and McFowland III [SM16] and Shalizi and Thomas [ST11] use stochastic block models and latent space models in order to obtain consistent estimators of peer contagion. Building upon this work, Sridhar et al. [SBB22] adjust for the unobserved confounders in peer influence estimation by using Poisson Influence Factorization (PIF) models.

Some of the recent studies which address nonparametric peer influence estimation from observational data propose different approaches which complement the node embedding based method proposed in this paper. Eckles and Bakshy [EB17] estimate peer effects on hundreds of millions of observations from Facebook data by using high-dimensional adjustments for covariates via propensity score models. One limitation of this work is that covariates are readily available, whereas this paper seeks to cover the scenario in which they are unobserved. Egami and Tchetgen Tchetgen [ETT21], on the other hand, tackle the situation of unobserved confounding, by using negative control outcomes and exposure variables to estimate contagion effects. This approach complements the methods of this paper. While not directly related to peer influence estimation, Guo et al. [GLL20] also propose a nonparametric technique of adjustment for the unobserved features that affect network treatment and outcome. They propose using "graph attentional layers" in order to map the network information to a partial representation of the hidden confounders and use it to perform counterfactual evaluation.

8 Discussion

This paper studies the problem of causally estimating peer influence from observational data in the presence of unobserved confounding. The main contributions of this work are

1. formalizing and justifying two causal estimands for contagion while accounting for homophily (Section 3);
2. giving sufficient conditions for network embeddings to enable causal identification (Theorem 2);
3. illustrating how embeddings can be used to estimate peer contagion and showing their practical performance is overall better than that of standard naive and parametric estimation techniques (Section 5 and Section 6).

There are several avenues for further work. First, one current methodological limitation is the lack of an asymptotic normality result for the peer influence estimator $\hat{\psi}_{t^*}$ defined in Section 5. Characterizing the asymptotic normality of the main estimator would allow one to construct valid confidence intervals around the estimated values, at a desired significance level. Secondly, this paper only focused on network embeddings learned via subsampling as a proposed nonparametric method for peer contagion estimation. An important future work direction is exploring other nonparametric techniques (e.g. graph neural networks [Zho+18; Wu+19; ZCZ22; Ma+22; Hus+21]) which leverage the homophily present in social networks and are able to learn meaningful node representations to be used downstream for peer influence estimation.

9 Acknowledgements

This work was partially supported by Open Philanthropy. We thank Steve Yadlowsky for helpful comments and feedback. We would also like to thank the Research Computing Center (RCC) at The University of Chicago for computing resources.

References

[Bor+16] C. Borgs, J. Chayes, H. Cohn, and N. Holden. "Sparse exchangeable graphs and their limits via graphon processes". In: *Journal of Machine Learning Research* (2016).

[BDF20] Y. Bramoulle, H. Djebbari, and B. Fortin. "Peer effects in networks: a survey". In: *Annual Review of Economics* (2020).

[CCB16] D. Cai, T. Campbell, and T. Broderick. "Edge-exchangeable graphs and sparsity (nips 2016)". In: *arXiv: Machine Learning* (2016).

[CF17] F. Caron and E. Fox. "Sparse graphs using exchangeable random measures". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2017).

[CD18] H. Crane and W. Dempsey. "Edge exchangeable models for interaction networks". In: *Journal of the American Statistical Association* (2018).

[DA21] A. Davison and M. Austern. "Asymptotics of network embeddings learned via subsampling". In: *arXiv:stat.ML* (2021).

[EB17] D. Eckles and E. Bakshy. "Bias and high-dimensional adjustment in observational studies of peer effects". In: *Journal of the American Statistical Association* (2017).

[EKB16] D. Eckles, R. Kizilcec, and E. Bakshy. "Estimating peer effects in networks with peer encouragement designs". In: *Proceedings of the National Academy of Sciences* (2016).

[ETT21] N. Egami and E. Tchetgen Tchetgen. "Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding". In: *arXiv: stat.ME* (2021).

[FZ20a] Z. Fatemi and E. Zheleva. "Minimizing interference and selection bias in network experiment design". In: *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)*. 2020.

[FZ20b] Z. Fatemi and E. Zheleva. "Network experiment design for estimating direct treatment effects". In: *KDD International Workshop on Mining and Learning with Graphs (MLG 2020)*. 2020.

[GLL20] R. Guo, J. Li, and H. Liu. "Counterfactual evaluation of treatment assignment functions with networked observational data". In: 2020.

[HYL17] W. Hamilton, R. Ying, and J. Leskovec. "Representation learning on graphs: methods and applications". In: *IEEE Data Engineering Bulletin* (2017).

[HRH02] P. D. Hoff, A. Raftery, and M. Handcock. "Latent space approaches to social network analysis". In: *Journal of the American Statistical Association* (2002).

[HLL83] P. Holland, K. B. Laskey, and S. Leinhardt. "Stochastic blockmodels: first steps". In: *Social Networks* (1983).

[Hus+21] H. Hussain, T. Duricic, E. Lex, D. Helic, and R. Kern. "The interplay between communities and homophily in semi-supervised classification using graph neural networks". In: *Applied Network Science* (2021).

[Jan18] S. Janson. "On edge exchangeable random graphs". In: *Journal of Statistical Physics* (2018).

[LW19] C. Lee and D. Wilkinson. "A review of stochastic block models and extensions for graph clustering". In: *Applied Network Science* (2019).

[LK14] J. Leskovec and A. Krevl. "{Snap datasets}: {stanford} large network dataset collection". In: 2014.

[Ma+22] Y. Ma, X. Liu, N. Shah, and J. Tang. "Is homophily a necessity for graph neural networks?" In: *International Conference on Learning Representations*. 2022.

[Ogb+17] E. L. Ogburn, O. Sofrygin, I. Diaz, and M. J. van der Laan. "Causal inference for social network data". In: *arXiv e-prints* (2017).

[RFR16] R. Rastelli, N. Friel, and A. Raftery. "Properties of latent variable network models". In: *Network Science* (2016).

[SM16] C. Shalizi and E. McFowland III. "Estimating causal peer influence in homophilous social networks by inferring latent locations". In: *arXiv e-prints* (2016).

[ST11] C. R. Shalizi and A. C. Thomas. "Homophily and contagion are generically confounded in observational social network studies". In: *Sociol. Methods Res.* 2 (2011).

[SBB22] D. Sridhar, C. D. Bacco, and D. Blei. "Estimating social influence from observational data". In: *First Conference on Causal Learning and Reasoning*. 2022.

[TZ12] L. Takac and M. Zábovský. "Data analysis in public social networks". In: *International Scientific Conference and International Workshop Present Day Trends of Innovations* (2012).

[TFS17] E. J. Tchetgen Tchetgen, I. Fulcher, and I. Shpitser. "Auto-g-computation of causal effects on a network". In: *arXiv e-prints* (2017).

[TK13] P. Toulis and E. Kao. "Estimation of causal peer influence effects". In: *Proceedings of the 30th International Conference on Machine Learning*. 3. 2013.

[TZ22] C. Tran and E. Zheleva. "Network experiment design for estimating direct treatment effects". In: *Proceedings of the 2022 AAAI Conference on Artificial Intelligence (AAAI 2020)*. 2022.

[van14] M. van der Laan. "Causal inference for a population of causally connected units". In: *Journal of Causal Inference* (2014).

[Vei+19] V. Veitch, M. Austern, W. Zhou, D. M. Blei, and P. Orbanz. "Empirical risk minimization and stochastic gradient descent for relational data". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. 2019.

[VWB19] V. Veitch, Y. Wang, and D. M. Blei. "Using embeddings to correct for unobserved confounding in networks". In: *Advances in Neural Information Processing Systems*. 2019.

[Wu+19] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. "A comprehensive survey on graph neural networks". In: *IEEE Transactions on Neural Networks and Learning Systems* (2019).

[Yin+17] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich. "Local higher-order graph clustering". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017.

[ZCZ22] Z. Zhang, P. Cui, and W. Zhu. "Deep learning on graphs: a survey". In: *IEEE Transactions on Knowledge & Data Engineering* 01 (2022).

[ZA21] E. Zheleva and D. Arbour. "Causal inference from network data". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. 2021.

[Zho+18] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun. *Graph Neural Networks: A Review of Methods and Applications*. cite arxiv:1812.08434. 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Section 5.1 which discusses the limitations of the proposed method, in light of the required assumptions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This paper mainly focuses on proposing a new causal estimation procedure for peer influence, and illustrating the performance of this procedure on social network data; this is a purely methodological paper which does not have any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The full model assumptions are detailed in Section 2. All theoretical results (i.e. Theorem 1, Theorem 2, Corollary 3, and Theorem 4) include full assumptions in their statements.
   (b) Did you include complete proofs of all theoretical results? [Yes] Full proofs of all theoretical results are presented in Appendix A, Appendix B, Appendix C, and Appendix D which can be found in the "Appendix.pdf" file within the folder "SUPPLEMENTARY.zip"
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The full code and data are included in the supplemental material within the folder "SUPPLEMENTARY.zip". This folder also includes the instructions needed to reproduce the main experimental results, which can be found in the file "instructions.md".
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see Section 6, Appendix F, as well as the code attached in the supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We present error bars for all our results in Appendix E together with a full description of how they were obtained.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix F.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] All assets used are publicly available online.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The full code and data used are attached in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data used in experiments is publicly available online at https://snap.stanford.edu/data/soc-Pokec.html.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data used does not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of Theorem 1

Below we include the proof of Theorem 1, which generalizes the two measures of peer contagion, $\psi_{t^*}^{\text{full info}}$ and ψ_{t^*}. It suffices to prove that the first estimand introduced in Section 3

$$\psi_{t^*}^{\text{full info}} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i|\text{do}(T=t^*), \{C_i\}_i, G_n]$$

converges to a fixed real value ψ, in probability. Then, since ψ is bounded, by the Dominated Convergence Theorem, it will immediately follow that the second estimand introduced in Section 3

$$\psi_{t^*} := \mathbb{E}[\psi_{t^*}^{\text{full info}}|\text{do}(T=t^*), G_n]$$

also converges, in probability, to the same real value ψ.

We now proceed to prove our main result for $\psi_{t^*}^{\text{full info}}$. We want to show that $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i|\text{do}(T = t^*), \{C_i\}_i, G_n] \to 0$, as $n \to \infty$, in probability. To simplify notation, we define the n-tuple of random variables $(X_1, \ldots, X_n)$, where each $X_i = \mathbb{E}[Y_i|\text{do}(T = t^*), \{C_i\}_i, G_n]$, respectively. Furthermore, let $S_n = \sum_{i=1}^{n} X_i$, the nth partial sum of the random variables $\{X_i\}_i$. We want to show that

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| > \varepsilon\right] = 0.$$

To prove this, we first note that, by Chebyshev's inequality

$$\mathbb{P}\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| > \varepsilon\right] \leq \frac{\text{Var}\left(\frac{S_n}{n}\right)}{\varepsilon^2}. \tag{A.1}$$

We further upper bound the term Var $\left(\frac{S_n}{n}\right)$, as follows

$$\begin{aligned}
\text{Var}\left(\frac{S_n}{n}\right) &= \frac{\text{Var}(S_n)}{n^2} \\
&= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j)}{n^2} \\
&= \mathbb{E}[\text{Cov}(X_i, X_j)],
\end{aligned}$$

where the expectation is taken over nodes i, j sampled uniformly at random. We now distinguish two possible cases

i. Nodes i and j do not share a common neighbor. In that case, based on the assumed structural equation model (2.1), $\mathbb{E}[Y_i|\text{do}(T = t^*), \{C_i\}_i, G_n]$ is independent from $\mathbb{E}[Y_j|\text{do}(T = t^*), \{C_j\}_j, G_n]$, therefore X_i is independent from X_j, and their covariance vanishes.

ii. Nodes i and j share common neighbors. In that case, note that, by the Cauchy-Schwarz inequality

$$\text{Cov}(X_i, X_j) \leq \sqrt{\text{Var}(X_i)\text{Var}(X_j)} \leq \max\{\text{Var}(X_i), \text{Var}(X_j)\}.$$

Furthermore, notice that, by assumption (i) of Theorem 1, $\text{Var}(Y_i) \leq M$, $\forall\ i$. By Jensen's inequality, and noting that the variance is a convex function, it follows that $\text{Var}(\mathbb{E}[Y_j|\text{do}(T = t^*), \{C_i\}_i, G_n]) = \text{Var}(X_i) \leq M$ as well. Hence, when i and j share common neighbors, we have that

$$\text{Cov}(X_i, X_j) \leq M.$$

From the two possible cases above, it follows that

$$\text{Var}\left(\frac{S_n}{n}\right) = \mathbb{E}[\text{Cov}(X_i, X_j)] \leq \mathbb{P}(\{\text{i and j share a neighbor}\}) \cdot M.$$

Then, by assumption (ii) of Theorem 1 it follows that the right-hand side of (A.1), $\text{Var}\left(\frac{S_n}{n}\right)/\varepsilon^2 \to 0$, as $n \to \infty$. This shows that

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{S_n}{n} - \mathbb{E}\left[\frac{S_n}{n}\right]\right| > \varepsilon\right] = 0,$$

as desired.

B Proof of Theorem 2

Consider Figure 1 which illustrates how conditioning on A_{ij} opens a backdoor path between T_j and Y_i, and how the embedding λ_i can be used to remove the effect of this conditioning.

More precisely,

$$\begin{aligned}
\mathbb{E}[Y_i|\text{do}(T = t^*), G_n, \lambda_i] &= \mathbb{E}[Y_i|\text{do}(V = v_i^*), G_n, \lambda_i] \\
&\quad \text{(the treatment function } V_i \text{ fully determines } Y_i \text{ by (2.1))} \\
&= \mathbb{E}[Y_i|\text{do}(V = v_i^*), \lambda_i] \\
&\quad \text{(invoke condition (i))} \\
&= \mathbb{E}[Y_i|V = v_i^*, \lambda_i] \\
&\quad \text{(no open backdoor paths after removing conditioning on } G_n\text{)},
\end{aligned}$$

as desired.

C Proof of Corollary 3

Starting from the definition of ψ_{t^*}, we have the following chain of equalities

$$\begin{aligned}
\psi_{t^*} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i \mid \text{do}(T = t^*), G_n] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[Y_i \mid \lambda_i, \text{do}(T = t^*), G_n]|\text{do}(T = t^*), G_n] \text{ (expand and marginalize over } \lambda_i\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[m_{G_n}(v_i^*, \lambda_i)|\text{do}(T = t^*), G_n] \text{ (invoke Theorem 2)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[m_{G_n}(v_i^*, \lambda_i)|G_n] \text{ (do}(T = t^*) \text{ does not affect } \lambda_i\text{)},
\end{aligned}$$

as desired.

D Proof of Theorem 4

Recall, from Theorem 1, that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y|\text{do}(T = t^*), C_i, G_n] = \psi$. This corollary assumes the conditions of Theorem 1, therefore, in order to show

$$\lim_{n\to\infty} \sum_{i=1}^{n} m_{G_n}(v_i^*, \lambda_i) = \psi,$$

it suffices to prove that $\mathbb{E}[\mathbb{E}[Y|\text{do}(T=t^*), C_i, G_n]] = \mathbb{E}[m_{G_n}(v_i^*, \lambda_i)]$, for each i, then invoke the version of the Law of Large Numbers derived from Chebyshev's inequality in Appendix A.

We first establish the necessary equality of expectations as follows

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y_i|\text{do}(T=t^*), C_i, G_n]] &= \mathbb{E}[\mathbb{E}[Y_i|\text{do}(V=v^*), C_i, G_n]] \quad (V_i \text{ determines } Y_i) \\
&= \mathbb{E}[\mathbb{E}[Y_i|\text{do}(V=v^*), \lambda_i, G_n]] \quad (\text{tower property; } \lambda_i \text{ is } C_i\text{-measurable}) \\
&= \mathbb{E}[m_{G_n}(v_i^*, \lambda_i)] \quad (\text{invoking Theorem 2})
\end{aligned}$$

Now, by the argument in the proof of Theorem 1, the objects $\sum_{i=1}^n \mathbb{E}[Y_i|\text{do}(T=t^*), C_i, G_n]$ and $\sum_{i=1}^n m_{G_n}(v_i^*, \lambda_i)$, depend, in the limit, only on the expected values of the summands. Since these expected values are equal, the limits are also equal.

E Simulation-variation in simulation results for the Pokec data

We expand the results presented in Table 1 and Table 2 to include error bars for each of the reported average causal peer effect values due to the variation induced by the choice of random seed. Please see Table 3 and Table 4. The tight error bands indicate the consistency of our simulation results, however, one main drawback is that due to the relational, non-i.i.d. structure of our data, these errors do not represent proper confidence bands, and hence should be interpreted with caution. In order to obtain valid confidence intervals, and thereby be able to conduct valid statistical inference in finite samples, one would require proper asymptotic normality results for the studied estimators which would allow constructing confidence intervals at a desired significance level. This limitation is a direction for future work.

Table 3: The embedding-based estimator $\hat{\psi}_{t^*}$ effectively adjusts for confounding and recovers the true treatment effect. The ground truth value of peer contagion is 1. Zero, low, and high confounding levels correspond to $\beta_1 = 0, 1$, and 10, respectively. For $\hat{\psi}_{t^*}$, the reported values represent the mean and standard error over 100 different global random seeds, while the seed for the simulated treatment and outcome data is kept constant. For the Unadjusted and Parametric estimators, the reported values represent the mean and standard error of the respective regression coefficients of the aggregated treatment used when predicting Y.

	district			**age**			**join_date**		
Conf. level	**Zero**	**Low**	**High**	**Zero**	**Low**	**High**	**Zero**	**Low**	**High**
Unadjusted	0.99± 0.00	1.64± 0.00	7.40± 0.02	1.00± 0.00	1.39± 0.00	4.90± 0.03	0.99± 0.00	1.38± 0.00	4.81± 0.03
Parametric	0.99± 0.01	1.41± 0.01	5.28± 0.03	1.00± 0.01	1.33± 0.01	4.20± 0.04	0.98± 0.01	1.28± 0.01	4.00± 0.04
$\hat{\psi}_{t^*}$	0.84± 0.01	0.96± 0.01	1.17± 0.01	0.94± 0.01	0.94± 0.01	1.11± 0.01	1.01± 0.01	1.03± 0.01	1.10± 0.01

F Additional experimental details

In this section we describe in more detail how the relational empirical risk minimization models were trained in order to obtain the estimated values $\hat{\psi}_{t^*}$ for peer contagion reported in Table 1 and Table 2. Both embedding-based models in Section 6.1 and Section 6.2 follow similar architectures. First, for both models, we sample subgraphs of size 800 from the full Pokec network. The subsampling scheme used is "biased-walk" which consists of a skipgram-based random-walk with unigram negative sampling. We simulate the treatments and the outcomes according to the model equations presented in Section 6.1 and Section 6.2, respectively, using a seed value of 0 in both cases. Then, for both of the embedding-based models in our two simulation studies (i.e., continuous outcomes—Section 6.1 and binary outcomes—Section 6.1), we jointly learn the embeddings $\hat{\lambda}$ and the vertex

Table 4: The embedding-based estimator $\hat{\psi}_{t^*}$ accurately recovers the true peer contagion effect of binary treatments on other subsequent binary treatments. The ground truth peer contagion value is 0. For $\hat{\psi}_{t^*}$, the reported values represent the mean and standard error over 100 different global random seeds, while the seed for the simulated treatment and outcome data is kept constant. For the Unadjusted and Parametric estimators, the reported values represent the mean and standard error of the respective regression coefficients of the aggregated treatment used when predicting Y.

Peer influence on vaccination	**district**	**age**	**join_date**
Unadjusted	2.03 ± 0.02	0.12 ± 0.02	0.68 ± 0.02
Parametric	1.30 ± 0.04	1.03 ± 0.03	0.98 ± 0.03
$\hat{\psi}_{t^*}$	0.09 ± 0.00	0.11 ± 0.00	0.22 ± 0.00

conditional outcomes $\hat{m}_{G_n}(v_i^*, \lambda_i)$. We do this by training two simple Keras models consisting of: 1. an embedding layer and a linear layer applied to the tensor of concatenated embeddings and aggregated treatment values for the continuous outcome model; and 2. an embedding layer and a dense layer with a sigmoid activation function applied to the tensor of concatenated embeddings and aggregated treatment values for the binary outcome model. For both models, we use embeddings of dimension 128. We optimize for the loss function in (5.2), using an "outcome-specific" weight q value of 0.005 for both models. This particular value of q is set to match that used in previous work [Vei+19; VWB19]. Based on experimental performance, for the continuous outcome model we use SGD as the optimizer (with learning rates ranging from 0.1 to 0.6 depending on the given combination of hidden confounder and confounding level), while for the binary outcome model we use Adam (with a learning rate of 0.001 for all different simulation scenarios - i.e. all three possible hidden confounders). The values for the optimizer learning rates correspond to the lowest validation loss values obtained when performing a 50/50 train/test split of the network data, fitting the models on the train data, and evaluating their performance on the held-out data.

The models were trained using only CPUs, as training each individual model is relatively fast (approximately 15 minutes on a CPU). Each experiment was run across 100 different global random seeds (with values ranging from 1 to 100), with nine different networks trained for the experiments in Section 6.1 (one for each possible combination of hidden confounder variable $\{$district, age, join_date$\}$ and confounding level $\{$zero, low, high$\}$), and three different networks trained for the experiments in Section 6.2 (for each possible hidden confounder variable $\{$district, age, join_date$\}$). A total of 1200 networks were thus fit across all random seeds and experimental setups.

G Supplementary experiments

G.1 Ablation studies

As discussed in Section 6.1, the continuous outcome results in Table 1 indicate that, when no unobserved confounding is present, in some situations $\hat{\psi}_{t^*}$ performs slightly worse compared to the baselines. This section investigates the performance of our method when the confounding level β_1 is strictly positive, yet low, and smaller than the peer contagion effect $\beta_0 = 1$. We conduct similar experiments to those in Table 1, letting β_1 vary in $\{0.25, 0.5, 0.75\}$. The results presented in Table 5 interestingly show that even for small, non-zero confounding level values, the embedding-based method still achieves the best overall performance. This indicates that the slight lack of accuracy of our proposed method in the no unobserved confounding case could potentially be an edge case.

Furthermore, we also study the impact of the noise level ε added to the response variable Y in (6.1). Recall that for all experiments in Section 6, $\varepsilon \sim N(0, 1)$. To investigate the performance of the embedding technique over various signal to noise ratios (SNR), we fix $\beta_0 = \beta_1 = 1$, take *district* as the unobserved confounder, and perform experiments similar to those in Table 1, letting the standard deviation of ε vary in $\{0.25, 2.5, 5\}$. These choices ensure signal to noise ratios which range from both high to low values. As shown in Table 6, we notice that the embedding based estimator $\hat{\psi}_{t^*}$ is

Table 5: The embedding-based estimator $\hat{\psi}_{t^*}$ outperforms baselines when the level of unobserved confounding β_0 is strictly positive, yet very small.

	district			**age**			**join_date**		
Conf. level	**0.25**	**0.5**	**0.75**	**0.25**	**0.5**	**0.75**	**0.25**	**0.5**	**0.75**
Unadjusted	1.20	1.48	1.20	1.27	1.34	1.62	1.28	1.28	1.57
Parametric	1.11	1.19	1.31	1.11	1.16	1.23	1.12	1.12	1.26
$\hat{\psi}_{t^*}$	1.05	1.09	1.13	0.94	0.93	0.93	0.94	0.94	0.95

Table 6: The embedding-based estimator $\hat{\psi}_{t^*}$ outperforms baselines even under high levels of noise added to the response variable Y.

	district			**age**			**join_date**		
Noise level	**0.25**	**2.5**	**5**	**0.25**	**2.5**	**5**	**0.25**	**2.5**	**5**
Unadjusted	1.69	1.87	1.76	1.71	1.95	2.21	1.60	1.91	2.24
Parametric	1.43	1.36	1.56	1.32	1.35	1.40	1.31	1.34	1.34
$\hat{\psi}_{t^*}$	0.95	0.94	0.93	0.93	0.93	0.93	0.95	0.95	0.95

Table 7: Experiments for Wikipedia network data. The embedding-based estimator $\hat{\psi}_{t^*}$ accurately estimates the ground truth peer contagion value is 1.

	category tag		
Conf. level	**Zero**	**Low**	**High**
Unadjusted	0.99	0.88	-0.05
$\hat{\psi}_{t^*}$	0.95	0.94	0.91

able to most accurately learn the true signal β_0 for both low and high values of noises. This is not the case for the other two baselines. This further confirms the validity of the proposed method.

G.2 Experiments on Wikipedia hyperlink network data

In this subsection, we apply the proposed technique to a new dataset, namely a network of Wikipedia articles (vertices) joined by hyperlinks (edges) [Yin+17]. Each article has a vector of labels which represent its tagged categories. We subset the data and reduce it to a sub-network of 27361 articles connected by 43809 edges. These articles corresponding to three of the most popular categories in the original network (i.e., *Olympic canoeists of Great Britain*, 20th *century Fox films*, and *French astronomers*), such that no two of these categories tag the same article.

For this data, we take the unique category attached to each article as the unobserved confounder C and generate treatment T and outcome Y values in the same way as in (6.1). The ground truth peer contagion effect is again set to $\beta_0 = 1$. Similarly to the experiment in Section 6.1, we let the unobserved confounding level β_1 vary in $\{0, 1, 10\}$, corresponding to zero, low, and high confounding, respectively, and compare the embedding-based estimator $\hat{\psi}_{t^*}$ against the naive unadjusted one. The results in Table 7 show that our method has both superior overall performance over the naive baseline and also accurately estimates the true peer contagion effect. This analysis shows the applicability of the proposed method to new datasets with different network structures from that in the Pokec data.