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 $\text{Var}\left(\frac{S_n}{n}\right)$, 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,

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

as desired.

C Proof of Corollary 3

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

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

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 $\mathbb{E}[\mathbb{E}[Y|\mathrm{do}(T=t^*), C_i, G_n]] = \mathbb{E}[m_{G_n}(v_i^*, \lambda_i)]$, 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

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

Now, by the argument in the proof of Theorem 1, the objects $\sum_{i=1}^n \mathbb{E}[Y_i|\mathrm{do}(T=t^*), C_i, G_n]$ 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*, 20^{th} *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.