Using Embeddings for Causal Estimation of Peer Influence in Social Networks

Irina Cristali and Victor Veitch

Department of Statistics, The University of Chicago

Overview

Challenge: Homophily acts as an unobserved confounder for peer contagion effects on networks.

- Insight: Use the **network itself** as a **proxy** for unobserved confounding.
- **Nonparametrically** formalize the causal peer influence effect.
- Use **black-box embedding methods** to identify and estimate this effect.

A Motivating Example

What effect does peer pressure have on vaccination?

- *If I get vaccinated, will my friends also get vaccinated?* (peer influence).
- *What if we all got vaccinated because we go to the same school, and that's why we're friends in the first place?* (homophily).

Challenge: How to isolate homophily from true peer influence effect?

Setup

- Outcomes Y_i—e.g. vaccination status of person i at end of study period. This is affected by:
- $\{T_j\}$—the treatment status of i's peers j.
- $\{C_i\}$—a vector of individual covariates.
- $\{A_{ij}\}$—the edge between i and j.

Problem: The covariates $\{C_i\}$ are unobserved. Can't escape conditioning on the edges A_{ij}. Turns out, doing so creates a collider bias between the unobserved confounders.

Solution: Infer a surrogate λ that captures the info. about the unobserved $\{C_i\}$.

Creating the Surrogate

Problems:

- No realistic generative model for $(\lambda_i, C_i, T_i, Y_i) \rightarrow$ **standard parametric methods fail.**
- Non-i.i.d. network structure $\rightarrow$ **standard nonparametric methods fail too!**

Solutions:

- Use **structural equation models** (flexible & general).
- Assign **embedding vectors** $\lambda_i \in \mathbb{R}^p$ for each node i.
- Embeddings $\rightarrow$ good at learning the **local network structure** and **unobserved info.**

Model Equations

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

ϵ = exogeneous noise, S_Y = summary function.

Formalizing Peer Influence

Let $T_i \leftarrow t^*$ = treatment intervention. Two possible ways to define **peer influence estimands**:

- $\psi^{\text{full info}}_{t^*} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i | \mathrm{do}(T = t^*), \{C_i\}_i, G_n] \rightarrow$ avg. outcome under t^*, for **same set of people** connected by **same link structure**.
- $\psi_{t^*} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i \mathrm{do}(T = t^*), G_n] \rightarrow$ avg. outcome under t^*, for **same link structure** and **set of people consistent with the link structure**.

First estimand **not identifiable**, but **second is**. *However*, IF [suitable graph sparsity conditions] THEN **both estimands** converge to same $\psi \in \mathbb{R}$.

Using Embeddings for Identification

- Causal estimand depends on graph structure and unobserved confounders.
- **Insight:** Don't need full info, only a proxy λ sufficient for identification.
- Let $V_i = S_Y(\{T_j : A_{ij} = 1\})$ = agg. treatment at node i; v_i^* = value under $T = t^*$.
- **Sufficient condition:** $Y_i \perp\!\!\!\perp A_{ij} | (\lambda_i, v_i^*), \forall\ i, j$.
- **Proof intuition:** See Fig. 1.

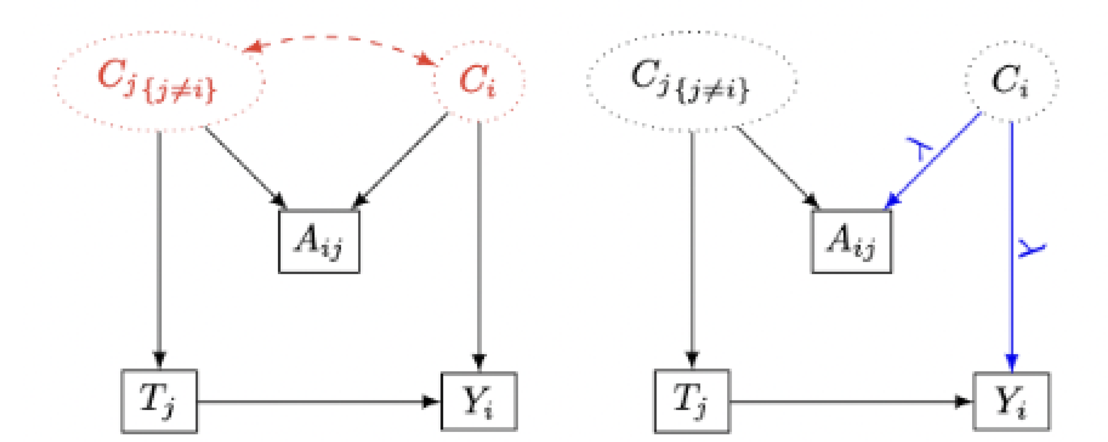

Figure 1: Identification of causal peer effects using embeddings. Adjusting for λ suffices to block the backdoor path between T_j and Y_i that is opened by conditioning on A_{ij}.

Estimation Method

Let $m_{G_n}(v_i^*, \lambda_i) = \mathbb{E}[Y_i | v_i^*, \lambda_i]$. Three-step empirical risk minimization estimation method:

1. Assign embeddings $\lambda_i \in \mathbb{R}^p$ for each unit i.
2. Jointly learn $\hat{\lambda}$ and $\hat{m}_{G_n}(v_i^*, \lambda_i)$ by minimizing a loss function with a **CrossEnt term** (for λ) and a **MSE term** (for $m_{G_n}(v_i^*, \lambda_i)$).
3. Plug in estimated values in $\hat{m}_{G_n}(v_i^*, \hat{\lambda}_i)$, and take average.

Causal peer influence effect of interest is then:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{m}(\hat{\lambda}_i, v^*_{\{i,t_i^*=1\}}) - \frac{1}{n}\sum_{i=1}^{n} \hat{m}(\hat{\lambda}_i, v^*_{\{i,t_i^*=0\}}).$$

Experiments: Social Network Pokec

- Analyze a sub-network of 70000 users connected by 1.3 mil. links (*Pokec* online social network).
- Take **district, age** and **registration date** as hidden confounders.
- Simulate treatments and outcomes as functions of hidden confounders.

Results: Continuous Outcome

Outcome and treatment simulated using labeled attribute as hidden confounder. Ground truth peer influence = 1. Our method beats baselines:

	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

Results: Demo for Vaccination

50% of individuals get vaccinated at time t_1, and 50% get vaccinated t_2. Real influence effect of first group on second is 0, yet there's still association via hidden confounders. Embeddings show great improvement:

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

References

[1] Ogburn, E. L., Sofrygin, O., Diaz, I. and van der Laan, M. J. Causal Inference for Social Network Data. arXiv 2017.

[2] Veitch, V., Austern, M., Zhou, W., Blei, D. and Orbanz, P. Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data. AISTATS 2019.

[3] Veitch, V., Wang, Y. and Blei, D. Using Embeddings to Correct for Unobserved Confounding in Networks. NeurIPS 2019.