Causal Estimation of Peer Influence Effects

Edward Kao, Panos Toulis, Edoardo Airoldi, Donald Rubin
Department of Statistics, Harvard University

Abstract

Causal inference on peer influence effects is of rising interest, but presents new challenges such as dependencies from network structure, complex response functions and network uncertainty. Adopting the potential outcomes framework, we introduce novel causal estimands and explore two estimation procedures. In the randomization based procedure, we characterize a bias-information trade-off: getting more causal information may lead to more biased estimates and vice versa. In the model-based procedure, we perform Bayesian inference through a hierarchical linear model, which achieves higher precision under true linear response function and also accounts for network uncertainty. On the other hand, the randomization-based procedure requires knowledge of the network but can operate under more complicated response functions with non-linearities. This complementarity gives insight for practicioners estimating peer influence effects.

Problem Statement

Treatments in a network not only affect the treated individuals, but also their neighbors through peer influence. This breaks the classical independence assumption between individual responses to treatments on others. Estimating the causal effects of network treatments remains an open problem. We investigate:

1. What quantities (i.e. estimands) capture the relevant causal effects?
2. How do we design an experiment and perform inference to estimate these quantities?
3. What is the performance of the resulting estimators, relating to the network structure and treatment response function?

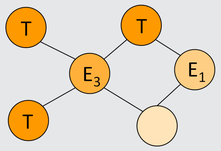

Figure 1: An example social network where three individuals (labeled with T) are given campaign messages. They in turn may share these messages with friends via Facebook. Although not having received the messages directly, node E_3 may be exposed to significant peer influence, and similarly node E_1 to a lesser degree.

Background

The prevalence of the social web in recent years has provided new and challenging causal questions. Related literature includes:

- Effects of product endorsement by Facebook friends (Bakshy, 2012)
- Click-pricing policy effects on total revenue of Yahoo! ad auctions in the presence of interacting, self-interested agents (Ostrovsky, 2010)
- Housing mobility studies and effects of government subsidies for family neighborhood relocation (Sobel, 2006)
- Studies on statistical causal inference under interference with an application on vaccination (Hudgens & Halloran, 2008)
- Spread of (mis) information in social networks (Acemoglu, 2010)
- Day-by-day adoption of technology products and distinguising effects from homophily and social diffusion (Aral, 2009; Manski, 1995)
- Modeling of rumor spreading in a network and estimation of rumor source (Shah, 2011)

Our contributions include:

1. Propose a novel causal estimand for peer influence effects
2. Characterize the bias and validity issues of randomization-based estimation and propose novel assignment strategy
3. Propose a linear model-based estimation procedure and compute optimal assignment through Fisher information analysis
4. Contrast randomization and model-based approaches through experimentation

Overview of Method

- Perform causal inference using the Neyman-Rubin potential outcome framework (Rubin, 1974, 1990)
- Define novel causal estimands for peer influence effects from k neighbors
- Estimate causal estimands through randomization-based and model-based approaches
- Evaluate theoretical performance of each approach in terms of manipulability, bias, and variance. Derive optimal assignment
- Contrast randomization-based with model-based approach through empirical results

Notations

- Network: $G = (V, E)$, with V and E vertex and edge sets. $|V|$ = N. Node $i \in V$ has neighbor set $\mathcal{N}_i$, excluding i. $|\mathcal{N}_i| = n_i$. Define V_k as the set of nodes with at least k neighbors. Define $\mathcal{M}$ as the set of nodes (i.e. shared neighbors) that neighbor with more than one node in V_k, $\mathcal{M}_i = \mathcal{M} \cap \mathcal{N}_i$, $|\mathcal{M}_i| = m_i$.
- Treatment vector: $\mathbf{Z}$ with $Z_i = 1$ if node i is treated and $Z_i = 0$ if in control. Define $\mathcal{D} = \{0, 1\}$, $\mathbf{Z} \in \mathcal{D}^{|\mathcal{N}_i|}$. $\mathbf{Z}_{\mathcal{N}_i}$ is a treatment subvector on the neighbors of i.
- K-level exposure: $\mathbf{Z}(\mathcal{N}_i; k)$ the set of all assignments on $\mathcal{N}_i$ in which exactly k neighbors of i get treated. Define $\mathbf{Z}_1(\mathcal{N}_i; k) \in \mathbf{Z}(\mathcal{N}_i; k)$ where at least one of the $\mathcal{M}_i$ is treated. Reversely, $\mathbf{Z}_0(\mathcal{N}_i; k) \in \mathbf{Z}(\mathcal{N}_i; k)$ where none of $\mathcal{M}_i$ is treated.
- Treatment response: $Y_i(\mathbf{Z}) \equiv Y_i(Z_i, \mathbf{Z}_{-i})$ of node i with treatment vector $\mathbf{Z}$. The second form denotes the treatment on i separately from the rest.

Causal Estimands

Definition 1. [Main estimand for peer influence effects] Define as δ_k the causal estimand of k-level effects, as follows:

$$\delta_k \equiv \frac{1}{|V_k|} \sum_{i \in V_k} \left[\binom{n_i}{k}^{-1} \sum_{\mathbf{z} \in \mathbf{Z}(\mathcal{N}_i; k)} Y_i(0, \mathbf{z}) - Y_i(\mathbf{0}) \right] \quad (1)$$

Definition 2. [Estimand for primary effects] Define as ξ the causal estimand of primary effects as follows.

$$\xi \equiv \frac{1}{N} \sum_i Y_i(1, \mathbf{z} = \mathbf{0}) - Y_i(\mathbf{0}) \quad (2)$$

Randomization-Based Estimation

Observe a subset of $Y_i(0, \mathbf{z})$ and $Y_i(\mathbf{0})$ using an assignment strategy, which then give the Neyman estimate through averaging.

Algorithm 1 Simple Sequential Randomization (SSR)

$\mathbf{Z} \leftarrow$ NA
while $i \leftarrow$ `sample`$\{i : i \in V_k$ & $sum(\mathbf{Z}_{\mathcal{N}_i}) \leq k\}$ **do**
  $\mathbf{w} \leftarrow$ `sample`$\{\mathbf{z} : \mathbf{z} \in \mathcal{D}^{|\mathcal{N}_i|}$ & $(sum(\mathbf{z}) - k) \cdot sum(\mathbf{z}) = 0\}$
  $\mathbf{Z}_{\mathcal{N}_i} \leftarrow \mathbf{w}$
  $Z_i \leftarrow 0$
end while

Network structure renders many of the potential outcomes unobservable. INR^x addresses this by putting $100x\%$ of the shared neighbors in control.

Algorithm 2 Insulated Neighbors Randomization (INR^x)

$\mathbf{Z} \leftarrow$ NA
$\mathbf{w} \leftarrow$ `sample`$\{$`n`$=x \cdot |\mathcal{M}|, \mathcal{M}\}$
$\mathbf{Z}_{\mathbf{w}} \leftarrow \mathbf{0}$
$\mathbf{Z} \leftarrow$ `SSR`$($`G`$,\mathbf{Z})$

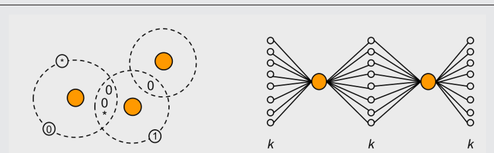

Figure 2: (Orange nodes are in V_k, '*' means unassigned) Left: INR^x puts in control proportion of the common neighbors. RIGHT: Candy network. To observe a causal estimate, the middle nodes need to be put in control. Only 2 out of $\binom{2k}{k}^2$ estimates are causal.

Model-Based Estimation

Assuming additivity of treatment effects (Manski, 1993), model individual response using a linear model, where τ: primary effect, γ: peer effect (unit exposure), μ: baseline response, and $\boldsymbol{A}$: directed, weighted network adjacency matrix:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (3) \qquad \boldsymbol{\beta} = \begin{bmatrix} \tau & \gamma & \mu \end{bmatrix}' \quad (4)$$

$$\boldsymbol{X} = \begin{bmatrix} \mathbf{Z} & \boldsymbol{A}^{\boldsymbol{T}}\mathbf{Z} & \mathbf{1} \end{bmatrix} \quad (5) \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathbb{N}(0, \sigma^2) \quad (6)$$

Estimating the causal estimands, under this model, simplifies to estimating τ and γ:

$$\begin{aligned} \xi &= \frac{1}{N} \sum_i \Big[Y_i(1, \mathbf{z} = \mathbf{0}) - Y_i(\mathbf{0}) \Big] = \frac{1}{N} \sum_i \tau = \tau \quad (7) \\ \delta_k &= \frac{1}{|V_k|} \sum_{i \in V_k} \left[\binom{n_i}{k}^{-1} \sum_{\mathbf{z} \in \mathbf{Z}(\mathcal{N}_i; k)} (Y_i(0, \mathbf{z}) - Y_i(\mathbf{0})) \right] \\ &= \frac{1}{|V_k|} \sum_{i \in V_k} \left[\binom{n_i}{k}^{-1} \sum_{\mathbf{z} \in \mathbf{Z}(\mathcal{N}_i; k)} S_i(\mathbf{z})\gamma \right] = \frac{k\gamma}{|V_k|} \sum_{i \in V_k} W_i \ \propto \ \gamma \quad (8) \end{aligned}$$

where W_i is the average incoming edge weight to node i.

To account for network uncertainty, model each weighted edge as a Poisson random variable with known rate. S_i, the amount of peer effect exposure to node i, is the sufficient network statistic:

$$A_{ij} \sim \text{Pois}(\lambda_{ij}) \quad (9) \qquad S_i = \boldsymbol{A}_{\cdot i}^{T}\mathbf{Z} \sim \text{Pois}(\kappa_i) \quad (10)$$

$$\boldsymbol{\kappa} = \boldsymbol{\Lambda}\mathbf{Z} \quad (11)$$

The above Poisson random network linear treatment model:

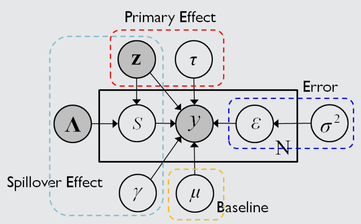

Figure 3: Model plate diagram. $\boldsymbol{y}$, $\boldsymbol{\Lambda}$, and $\mathbf{z}$ are known or observed.

Infer τ and γ through Markov Chain Monte Carlo:

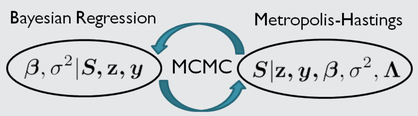

Figure 4: Joint inference of $\tau, \gamma, \mu, \sigma^2$, and $\boldsymbol{S}$ through MCMC.

Theoretical Results

As discussed in "randomization-based estimation", network structure may cause many potential outcomes to be unobservable. Here we analyze how much bias is introduced by this constraint. First, we break the peer effects estimand into two parts:

Definition 3. [Additional peer influence effects estimands]

("insulated neighbors")

$$\delta_{k,0} = \frac{1}{N} \sum_i \binom{n_i - m_i}{k}^{-1} \cdot \sum_{\mathbf{z} \in \mathbf{Z}_0(\mathcal{N}_i; k)} Y_i(0, \mathbf{z}) - Y_i(\mathbf{0}) \quad (12)$$

("non-insulated neighbors"):

$$\delta_{k,1} = \frac{1}{N} \sum_i \left(\binom{n_i}{k} - \binom{n_i - m_i}{k} \right)^{-1} \cdot \sum_{\mathbf{z} \in \mathbf{Z}_1(\mathcal{N}_i; k)} Y_i(0, \mathbf{z}) - Y_i(\mathbf{0}) \quad (13)$$

Definition 4. [Sharing index] For a given network, the sharing index $\alpha \in [0, 1]$ is defined by:

$$\alpha = \frac{1}{|V_k|} \sum_{i \in V_k} \frac{m_i}{n_i} \quad (14)$$

Theorem 1. If $\forall i$, $\binom{n_i - m_i}{k} / \binom{n_i}{k} = 1 - \alpha \leq 1$, then it holds:

$$E[\hat{\delta}_{k,INR}] = \delta_k + \alpha \cdot (\delta_{k,0} - \delta_{k,1}) \quad (15)$$

Theoretical Results (Cont.)

Corollary 1. *For an ego-centric network with no commonly shared nodes($\alpha = 0$), the estimate from INR is unbiased. Furthermore, if peer influence effects are invariant to permutations of node ids (and so $\delta_{k,0} = \delta_{k,1}$) the estimate from INR is unbiased.*

Optimal Assignment: For model-based estimation, Fisher information computation reveals optimal treatment assignments that lead to minimal variance estimates. For $\hat{\tau}$, optimal assignments consist of isolated treatments. For $\hat{\gamma}$, optimal assignments maximize peer effect exposure of selected nodes.

$$\begin{aligned} \mathbf{I}(\boldsymbol{\beta}, \sigma^{\boldsymbol{2}}) &= -E\left(\frac{\partial^2 l_{\mathbf{y}}(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{\kappa}, \mathbf{z})}{\partial[\tau, \gamma, \mu, \sigma^2] \partial[\tau, \gamma, \mu, \sigma^2]} \right) \\ &= \begin{bmatrix} \sum_i \frac{z_i}{\phi_i} & \sum_i \frac{z_i \kappa_i}{\phi_i} & \sum_i \frac{z_i}{\phi_i} & 0 \\ \cdot & \sum_i \left[\frac{\kappa_i^2}{\phi_i} - \frac{2\gamma^2 \kappa_i^2}{\phi_i^2} \right] & \sum_i \frac{\kappa_i}{\phi_i} & 0 \\ \cdot & \cdot & \sum_i \frac{1}{\phi_i} & 0 \\ \cdot & \cdot & \cdot & \sum_i \frac{1}{2\phi_i^2} \end{bmatrix} \quad (16) \end{aligned}$$

where $\phi_i = \sigma^2 + \gamma^2 \kappa_i$

Empirical Results

Compare the performance of the 2 randomization-based estimators (SSR and INR) and the 2 model-based estimators with random assignments (LMR) and optimal assignments (LMO), on simulated data using different random network models. Results show that when the additivity assumption is correct (Table 1), model-based estimators outperform the randomization-based estimators. However, when the response function is nonlinear (Table 2), randomization-based estimators are preferred.

Table 1: Linear response function. Reported are estimator mean(σ).

Graphs			Estimation strategy			
Type	Parameters	Truth: δ_4	*SSR*	*INR*$^{0.6}$	*LMR*	*LMO*
Small world	$p = 0.05$	11.18	10.49(2.04)	9.15(2.36)	10.85(1.09)	9.7(0.63)
	$p = 0.5$	11.45	11.45(1.82)	10.60(2.60)	11.31(1.08)	11.66(0.76)
	$p = 0.9$	12.97	12.54(1.35)	12.17(1.57)	12.82(1.54)	12.42(0.72)
4-community block model	diag(0.9)/off(0.1)	3.19	3.47(0.59)	3.34(0.61)	3.15(0.56)	3.37(0.26)
	diag(0.25)/off(0.75)	3.28	3.56(0.74)	3.77(0.82)	3.28(0.27)	3.72(0.23)
	Beta(0.1,0.1)	3.19	3.2(0.49)	3.51(0.57)	3.25(0.36)	3.28(0.21)
	Beta(1,1)	3.27	3.41(0.57)	3.6(0.65)	3.30(0.33)	3.40(0.22)
Chung-Lu	-	2.88	2.9(0.61)	2.96(0.5)	2.99(0.31)	2.93(0.21)

Table 2: Quadratic response function. Reported are estimator mean(σ)

Graphs			Estimation strategy			
Type	Parameters	Truth: δ_4	*SSR*	*INR*$^{0.6}$	*LMR*	*LMO*
Small world	$p = 0.05$	42.46	36.38(10.7)	32.13(13.96)	25.37(3.96)	27.32(1.49)
	$p = 0.5$	44.68	45.73(10.01)	39.52(14.24)	27.27(4.15)	35.45(2.05)
	$p = 0.9$	50.52	47.54(7.00)	46.87(9.12)	33.39(4.38)	36.00(1.80)
4-community block model	diag(0.9)/off(0.1)	5.48	5.97(1.25)	5.77(1.45)	6.64(0.9)	9.28(0.66)
	diag(0.25)/off(0.75)	5.63	6.01(1.26)	7.05(1.81)	8.96(1.20)	11.25(0.61)
	Beta(0.1,0.1)	5.42	5.63(0.94)	6.0(1.37)	8.40(1.19)	9.24(0.52)
	Beta(1,1)	5.59	6.07(1.08)	6.52(1.43)	6.53(0.75)	9.01(0.46)
Chung-Lu	-	4.65	4.83(0.75)	4.9(0.87)	6.87(0.74)	6.69(0.34)

References

Acemoglu, Daron, Ozdaglar, Asuman, and ParandehGheibi, Ali. Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010.

Sobel, M.E. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks *PNAS*, 10621544–21549, 2009.

Bakshy, Eytan, Eckles, Dean, Yan, Rong, and Rosenn, Itamar. Social influence in social advertising: Evidence from field experiments. *CoRR*, abs/1206.4327, 2012.

Hudgens, M.G. and Halloran, M.E. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

Manski, C.F. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993.

Manski, C.F. *Identification problems in the social sciences*. Harvard Univ Pr, 1995.

Ostrovsky, M. and Schwarz, M. Reserve prices in internet advertising auctions: A field experiment. *Available at SSRN 1573947*, (2054), 2010.

Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology; Journal of Educational Psychology*, 66(5):688, 1974.

Rubin, D.B. [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

Shah, Devavrat and Zaman, Tauhid. Rumors in a network: Who's the culprit? *IEEE Transaction on Information Theory*, 57(8):5163–5181, 2011.

Sobel, M.E. What do randomized studies of housing mobility demonstrate? *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.