Unconfoundedness with Network Interference

Michael P. Leung [1] Pantelis Loupos [2]

[1]UC Santa Cruz

[2]UC Davis

Neighborhood Interference

- Let $\boldsymbol{D} = (D_i)_{i=1}^n$ be the vector of assigned binary treatments for the n units connected through a network $\boldsymbol{A}$, represented as a binary, symmetric matrix with ijth entry A_{ij}.
- $Y_i(\boldsymbol{d})$ is the potential outcome of unit i under the counterfactual that the network is assigned treatment vector $\boldsymbol{d} \in \{0,1\}^n$.
- Most of the literature studies the neighborhood interference model that represents potential outcomes as

$$Y_i(\boldsymbol{D}) = \tilde{Y}_i(T_i) \quad \text{for} \quad T_i = f_n(i, \boldsymbol{D}, \boldsymbol{A}),$$

where T_i is the *effective treatment* or *exposure mapping*.

Neighborhood Interference

- $T_i \in \mathcal{T}$ only depends on treatments assigned to i's *K-neighbors*, those of path distance at most K from i.
- Under this model, the typical estimand of interest is the average exposure effect (AEE)

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[\tilde{Y}_i(t) - \tilde{Y}_i(t') \mid \boldsymbol{A}] \quad \text{for} \quad t, t' \in \mathcal{T}.$$

- E.g. $T_i = (D_i, \sum_{j=1}^{n} A_{ij} D_j)$. Then for $t = (1, 0)$ and $t' = (0, 0)$, the AEE is the average treatment effect for units with no treated neighbors.
- Inference well understood since neighborhood interference generates a convenient degree of independence between units.

Neighborhood Interference

- Neighborhood interference is a widespread assumption, but it is restrictive. It rules out endogenous peer effects, and T_i likely is misspecified (Sävje, 2023).
- We assume the more general condition of approximate neighborhood interference (ANI) (Leung, 2022).
- ANI says that the dependence of Y_i on D_j decays with path distance between i, j, which has been shown to allow for endogenous peer effects.
- Most of the literature studies RCTs. We will study observational settings satisfying an unconfoundedness condition.

Unconfoundedness

- Forastiere et al. (2021) and Ogburn et al. (2022) are key contributions studying neighborhood interference under the unconfoundedness condition

$$\tilde{Y}_i(\cdot) \perp\!\!\!\perp T_i \mid W_i. \tag{1}$$

- Vector of controls W_i may include unit-level covariates X_i and "network controls" such as $\sum_{j=1}^{n} A_{ij}X_j / \sum_{j=1}^{n} A_{ij}$ or other centrality measures.
- Just as T_i may be misspecified, so may be W_i. We can choose W_i to be any function of $(\boldsymbol{X}, \boldsymbol{A})$, and it may be hard to justify a particular choice.

Unconfoundedness

- We study a nonparametric behavioral model that allows for ANI in both the outcome and treatment selection stages. Under this model, (Y_i, D_i) depends on the entirety of $(\boldsymbol{X}, \boldsymbol{A})$.
- We therefore consider the unconfoundedness condition

$$\{Y_i(\cdot)\}_{i=1}^n \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{X}, \boldsymbol{A}. \tag{2}$$

- We do not assume the existence of a low-dimensional function T_i of $(\boldsymbol{D}, \boldsymbol{A})$ that summarizes interference or W_i of controls $(\boldsymbol{X}, \boldsymbol{A})$ that summarizes confounding.
- But (2) is challenging to utilize because the doubly robust estimator depends on $\mathbf{P}(T_i = t \mid \boldsymbol{X}, \boldsymbol{A})$.

Graph Neural Networks

- The nonparametric nuisance functions must also satisfy permutation-invariance.
- We therefore propose to estimate the functions with graph neural networks (GNNs).
- The key parameter of a GNN is its depth L (number of layers) because a GNN predicts Y_i not using the entirety of $(\boldsymbol{X}, \boldsymbol{A})$ but rather only i's L-neighborhood $(\boldsymbol{X}_{\mathcal{N}(i,L)}, \boldsymbol{A}_{\mathcal{N}(i,L)})$.
- GNNs have been found to perform best when shallow.
- We show that ANI implies a network analog of approximate sparsity, which provides low-dimensional structure justifying the use of shallow architectures in our setting.

Contributions

- Methodology: propose to use doubly robust estimator with GNNs.
- Modeling: provide nonparametric behavioral model allowing for a general form of interference in outcomes and treatment assignment.
- Identification: provide conditions under which the Sävje (2023) estimand has a causal interpretation.
- Large-network asymptotics: establish consistency and asymptotically normality for the doubly robust estimator and propose a new bandwidth for the network HAC estimator.
- It isn't currently possible to verify GNN rate conditions, but we provide useful intermediate results including primitive conditions for network approximate sparsity.

Model

- Each unit i is endowed with unobservables $(\varepsilon_i, \nu_i) \in \mathbb{R}^{d_\varepsilon} \times \mathbb{R}^{d_\nu}$ and observed covariates $X_i \in \mathbb{R}^{d_x}$.
- For $\boldsymbol{X} = (X_i)_{i=1}^n$ and $\boldsymbol{\varepsilon}$, $\boldsymbol{\nu}$ similarly defined,

$$Y_i = g_n(i, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\varepsilon}) \quad \text{and} \quad D_i = h_n(i, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\nu}).$$

- Potential outcome: $Y_i(\boldsymbol{d}) = g_n(i, \boldsymbol{d}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\varepsilon})$.
- Unconfoundedness: $\boldsymbol{\varepsilon} \perp\!\!\!\perp \boldsymbol{\nu} \mid \boldsymbol{X}, \boldsymbol{A}$.
- $(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\varepsilon}, \boldsymbol{\nu})$ is random. We observe $(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{A})$.

Interference

Assumption 1 (ANI)

There exists a sequence of nonrandom functions $\{(\gamma_n(\cdot), \eta_n(\cdot))\}_{n\in\mathbb{N}}$ *with* $\gamma_n, \eta_n \colon \mathbb{R}_+ \to \mathbb{R}_+$ *such that* $\sup_{n\in\mathbb{N}} \max\{\gamma_n(s), \eta_n(s)\} \to 0$ *as* $s \to \infty$ *and, for any* $n \in \mathbb{N}$,

$$\max_{i\in\mathcal{N}_n} \mathbf{E}\big[|g_n(i, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{A}, \varepsilon) - g_{n(i,s)}(i, \boldsymbol{D}_{\mathcal{N}(i,s)}, \boldsymbol{X}_{\mathcal{N}(i,s)}, \boldsymbol{A}_{\mathcal{N}(i,s)}, \varepsilon_{\mathcal{N}(i,s)})| \mid \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{A}\big] \leq \gamma_n(s) \quad a.s.$$

and

$$\max_{i\in\mathcal{N}_n} \mathbf{E}\big[|h_n(i, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\nu}) - h_{n(i,s)}(i, \boldsymbol{X}_{\mathcal{N}(i,s)}, \boldsymbol{A}_{\mathcal{N}(i,s)}, \boldsymbol{\nu}_{\mathcal{N}(i,s)})| \mid \boldsymbol{X}, \boldsymbol{A}\big] \leq \eta_n(s) \quad a.s.$$

Examples

Network version of the Manski (1993) linear-in-means model (Bramoullé et al., 2009):

$$Y_i = \alpha + \beta \frac{\sum_j A_{ij} Y_j}{\sum_j A_{ij}} + Z_i' \gamma + \varepsilon_i$$

for $Z_i = (D_i, X_i)$. This has a reduced-form (if $\boldsymbol{A}$ is connected)

$$\boldsymbol{Y} = \frac{\alpha}{1-\beta} \mathbf{1} + \boldsymbol{Z}\gamma + \gamma\beta \sum_{k=0}^{\infty} \beta^k \tilde{\boldsymbol{A}}^{k+1} \boldsymbol{Z} + \sum_{k=0}^{\infty} \beta^k \tilde{\boldsymbol{A}}^k \boldsymbol{\varepsilon},$$

which is a function $g_n(i, \boldsymbol{D}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\varepsilon})$. Proposition 1 of Leung (2022) shows that this satisfies ANI with $\gamma_n(s) = C|\beta|^s$.

Examples

Peer effects in treatments can be modeled as a binary game

$$D_i = \mathbf{1}\left\{\alpha + \beta \frac{\sum_{j=1}^n A_{ij} D_j}{\sum_{j=1}^n A_{ij}} + \frac{\sum_{j=1}^n A_{ij} X_j}{\sum_{j=1}^n A_{ij}} \gamma + X_j' \delta + \nu_i > 0\right\},$$

which may have multiple equilibria. The equilibrium selection mechanism is a mapping from primitives to treatments, which has the form $h_n(i, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\nu})$.

Proposition 2 of Leung (2022) provides conditions under which ANI holds with $\eta_n(s)$ decaying at an exponential rate with s uniformly in n.

Estimand

- We consider an analog of the estimand proposed by Sävje (2023) appropriate for unconfoundedness:

$$\tau(t, t') = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{E}[Y_i \mid T_i = t, \boldsymbol{X}, \boldsymbol{A}] - \mathbf{E}[Y_i \mid T_i = t', \boldsymbol{X}, \boldsymbol{A}]). \tag{3}$$

- This simply compares average outcomes under different exposure mappings.
- We provide conditions under which (3) has a causal interpretation.
- We make the case that (3) is estimable despite the high-dimensional controls.

Estimator

Abbreviate $\mathbf{1}_i(t) = \mathbf{1}\{T_i = t\}$, and define

$$p_t(i, \boldsymbol{X}, \boldsymbol{A}) = \mathbf{E}[\mathbf{1}_i(t) \mid \boldsymbol{X}, \boldsymbol{A}], \quad \mu_t(i, \boldsymbol{X}, \boldsymbol{A}) = \mathbf{E}[Y_i \mid T_i = t, \boldsymbol{X}, \boldsymbol{A}].$$

Let $\hat{p}_t(i, \boldsymbol{X}, \boldsymbol{A})$ and $\hat{\mu}_t(i, \boldsymbol{X}, \boldsymbol{A})$ denote their respective GNN estimators, and define the doubly robust estimator

$$\hat{\tau}(t, t') = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i(t, t'), \quad \text{where}$$

$$\hat{\tau}_i(t, t') = \frac{\mathbf{1}_i(t)(Y_i - \hat{\mu}_t(i, \boldsymbol{X}, \boldsymbol{A}))}{\hat{p}_t(i, \boldsymbol{X}, \boldsymbol{A})} + \hat{\mu}_t(i, \boldsymbol{X}, \boldsymbol{A}) - \frac{\mathbf{1}_i(t')(Y_i - \hat{\mu}_{t'}(i, \boldsymbol{X}, \boldsymbol{A}))}{\hat{p}_{t'}(i, \boldsymbol{X}, \boldsymbol{A})} - \hat{\mu}_{t'}(i, \boldsymbol{X}, \boldsymbol{A}).$$

Estimator

- To estimate the asymptotic variance, we use a network HAC estimator with uniform kernel (Kojevnikov et al., 2021)

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\tau}_i(t,t') - \hat{\tau}(t,t')) \times (\hat{\tau}_i(t,t') - \hat{\tau}(t,t'))\mathbf{1}\{\ell_{\boldsymbol{A}}(i,j) \leq b_n\},$$

  where $\ell_{\boldsymbol{A}}(i,j)$ is the path distance between (i,j) and b_n is the bandwidth.
- We propose a bandwidth b_n that modifies the formula proposed by Leung (2022) to account for first-stage estimates.

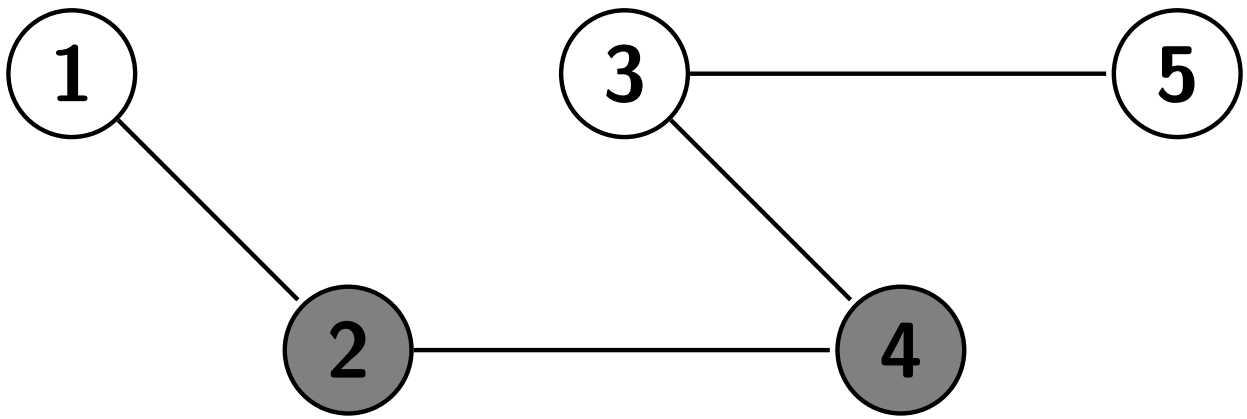
Invariance

- The p-score can't vary arbitrarily across i if we want a consistent estimator.
- Literature assumes $\mathbf{P}(T_i = t \mid \boldsymbol{X}, \boldsymbol{A}) = p(W_i)$, i.e. two units with the same W_i have the same p-score.
- We instead impose the much weaker condition of (permutation)-*invariance*: two units with isomorphic labeled network positions have equal p-scores. This is a consequence of a weak form of exchangeability.

Assumption 2 (Invariance)

For any $n \in \mathbb{N}$, *permutation* π, $i \in \mathcal{N}_n$, *and* $t \in \mathcal{T}$,
$p_t(i, \boldsymbol{X}, \boldsymbol{A}) = p_t(\pi(i), \pi(\boldsymbol{X}), \pi(\boldsymbol{A}))$ *and*
$\mu_t(i, \boldsymbol{X}, \boldsymbol{A}) = \mu_t(\pi(i), \pi(\boldsymbol{X}), \pi(\boldsymbol{A}))$.

Invariance



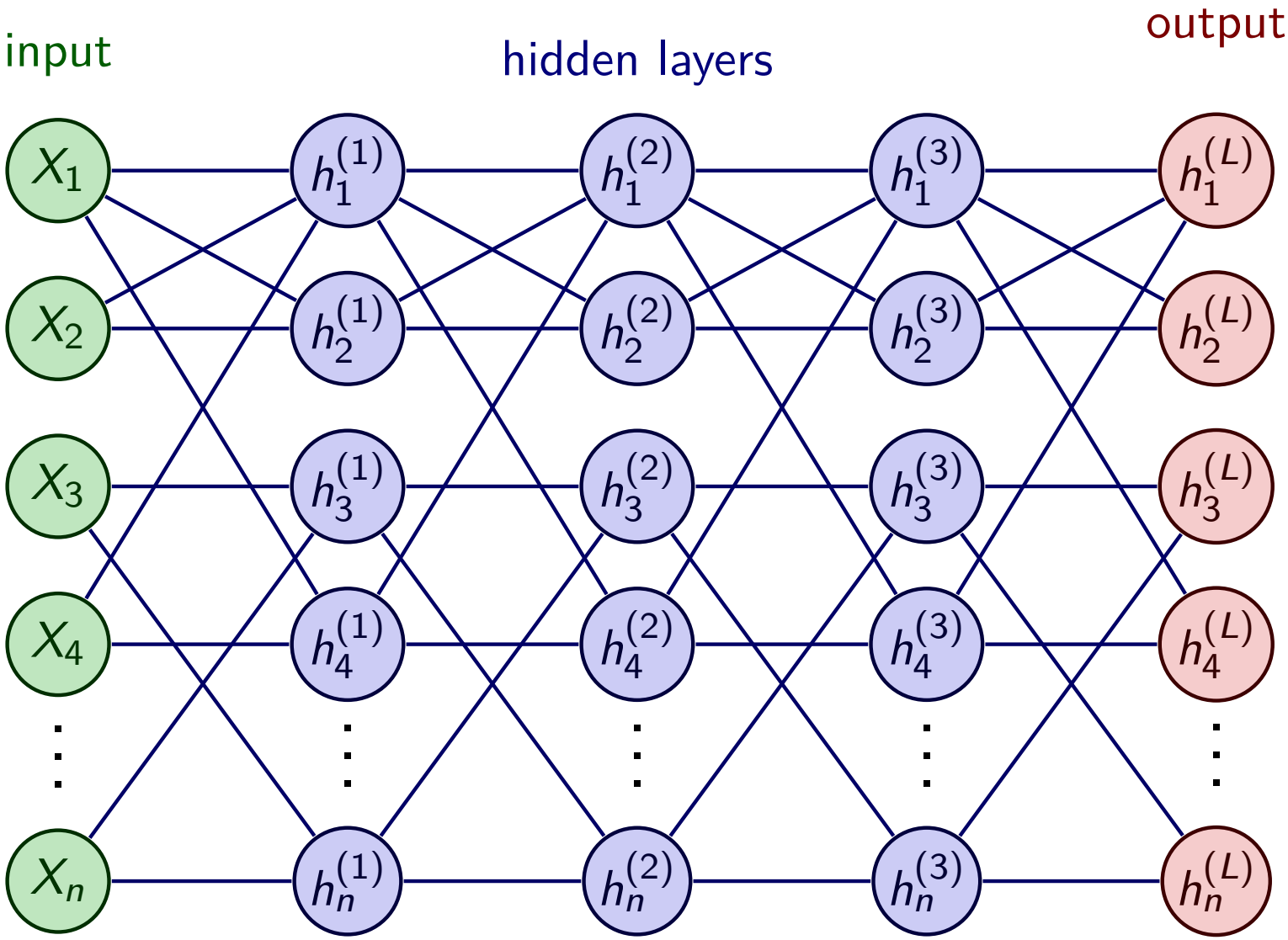
Each unit i has a binary covariate X_i that is an indicator for its color being gray. Let $W_i = (X_i, \sum_{j=1}^{n} A_{ij}, \sum_{j=1}^{n} A_{ij} X_j)$, a common choice of controls in the literature. Then $W_2 = W_4$, but units 2 and 4 are not isomorphic.

Architecture

- Goal is to estimate $\mathbf{P}(T_i = t \mid \boldsymbol{X}, \boldsymbol{A}) = f(i, \boldsymbol{X}, \boldsymbol{A})$. A GNN is a parameterized model for $f(\cdot)$.
- For $l = 1, \ldots, L$ and $i = 1, \ldots, n$, the ith neuron in layer l is

$$h_i^{(l)} = \Phi_{0l}\left(h_i^{(l-1)},\ \Phi_{1l}(h_i^{(l-1)}, \{h_j^{(l-1)} : A_{ij} = 1\})\right),$$

  where $h_i^{(0)} = X_i$ and $\Phi_{0l}(\cdot), \Phi_{1l}(\cdot)$ are parameterized functions. The latter imposes permutation-invariance.
- $\mathcal{F}_{\text{GNN}}(L)$ = set of all GNNs with L layers ranging over functions $\Phi_{0l}(\cdot), \Phi_{1l}(\cdot)$ for $l = 1, \ldots, L$ in some class.
- GNN estimator is the function $\hat{f}_{\text{GNN}} \in \mathcal{F}_{\text{GNN}}(L)$ that minimimzes a loss function.


input
hidden layers
output
X_1
X_2
X_3
X_4
X_n
$h_1^{(1)}$
$h_2^{(1)}$
$h_3^{(1)}$
$h_4^{(1)}$
$h_n^{(1)}$
$h_1^{(2)}$
$h_2^{(2)}$
$h_3^{(2)}$
$h_4^{(2)}$
$h_n^{(2)}$
$h_1^{(3)}$
$h_2^{(3)}$
$h_3^{(3)}$
$h_4^{(3)}$
$h_n^{(3)}$
$h_1^{(L)}$
$h_2^{(L)}$
$h_3^{(L)}$
$h_4^{(L)}$
$h_n^{(L)}$

Example

- The "graph isomorphism network" architecture (Xu et al., 2018) uses sum aggregation:

$$h_i^{(l)} = \phi_{0l}\left(h_i^{(l-1)}, \sum_{j=1}^{n} A_{ij}\phi_{1l}(h_j^{(l-1)})\right),$$

  where $\phi_{0l}(\cdot), \phi_{1l}(\cdot)$ are multilayer perceptrons (MLPs).
- Xu et al. (2018) show that any injective aggregator of a set S can be written as $g(\sum_{s\in S} f(s))$ for some functions f, g when X_i has countable support.
- By using MLPs to approximate f, g, this architecture can approximate a large nonparametric function class (Azizian and Lelarge, 2021).

Network Approximate Sparsity

- We prove under ANI and regularity conditions that a network analog of approximate sparsity holds.
- This converts the problem of estimating the high-dimensional function $\mathbf{P}(T_i = t \mid \boldsymbol{X}, \boldsymbol{A})$ to that of estimating $\mathbf{P}(T_i = t \mid \boldsymbol{X}_{\mathcal{N}(i,L)}, \boldsymbol{A}_{\mathcal{N}(i,L)})$ for $L = O(\log n)$.
- Under typical conditions on MLPs used in the architecture, the number of parameters is $o(\sqrt{n})$, effectively rendering the estimation problem low-dimensional (as in the lasso literature).
- However, the final step of showing that GNNs can estimate $\mathbf{P}(T_i = t \mid \boldsymbol{X}_{\mathcal{N}(i,L)}, \boldsymbol{A}_{\mathcal{N}(i,L)})$ is beyond the scope of the current literature (e.g. no concentration inequality).

Table: Simulation results for random geometric graph

	$L=1$			$L=2$			$L=3$		
n	1000	2000	4000	1000	2000	4000	1000	2000	4000
# treated	567	1137	2277	567	1137	2277	567	1137	2277
H	1	3	5	1	3	5	1	3	5
$\hat{\tau}(1,0)$	0.0783	0.0753	0.0680	0.0937	0.0382	0.0226	0.1288	0.0712	0.0353
CI	0.9316	0.9332	0.9324	0.9318	0.9368	0.9464	0.9360	0.9286	0.9384
SE	0.4279	0.3057	0.2166	0.5134	0.2961	0.2037	0.5745	0.3143	0.2021
Oracle CI	0.9426	0.9434	0.9358	0.9450	0.9498	0.9572	0.9464	0.9420	0.9472
Oracle SE	0.4473	0.3180	0.2190	0.5507	0.3153	0.2116	0.5994	0.3369	0.2094
W $\hat{\tau}(1,0)$	0.1800	0.1701	0.1555	0.1754	0.1650	0.1514	0.1765	0.1652	0.1516
W CI	0.9160	0.9042	0.8906	0.9182	0.9068	0.8932	0.9166	0.9054	0.8940
W SE	0.4338	0.3082	0.2177	0.4335	0.3084	0.2180	0.4328	0.3080	0.2178
IID CI	0.6968	0.6818	0.6862	0.6688	0.6704	0.6926	0.6658	0.6638	0.6822
IID SE	0.2363	0.1667	0.1174	0.2711	0.1567	0.1078	0.3015	0.1656	0.1063

5k simulations. The estimand is $\tau(1,0) = 0$. "# treated" $\approx$ effective sample size for GNN regression estimators. GNN depth is L, and MLP width is H. Rows beginning with "W" use GLMs with hand-selected controls and polynomial sieves of order L in place of GNNs. "CI" rows display the empirical coverage of 95% CIs.

Conclusion

- We study doubly robust estimation of treatment/spillover effects under a general unconfoundedness condition.
- Existing work assumes it suffices to control for a low-dimensional function W_i of $(\boldsymbol{X}, \boldsymbol{A})$.
- We use GNNs to effectively learn this function.
- Results not covered today:
  - Conditions under which $\tau(t, t')$ has a causal interpretation.
  - Estimator properties under large-network asymptotics.
  - Variance estimator.
  - Primitive conditions for network approximate sparsity.
  - Simulations showing GNNs substantially reduce bias relative to hand-selected controls for $L = 1, 2, 3$.