Using Embeddings to Correct for Unobserved Confounding in Networks

Victor Veitch, Yixin Wang, David M. Blei

Setup

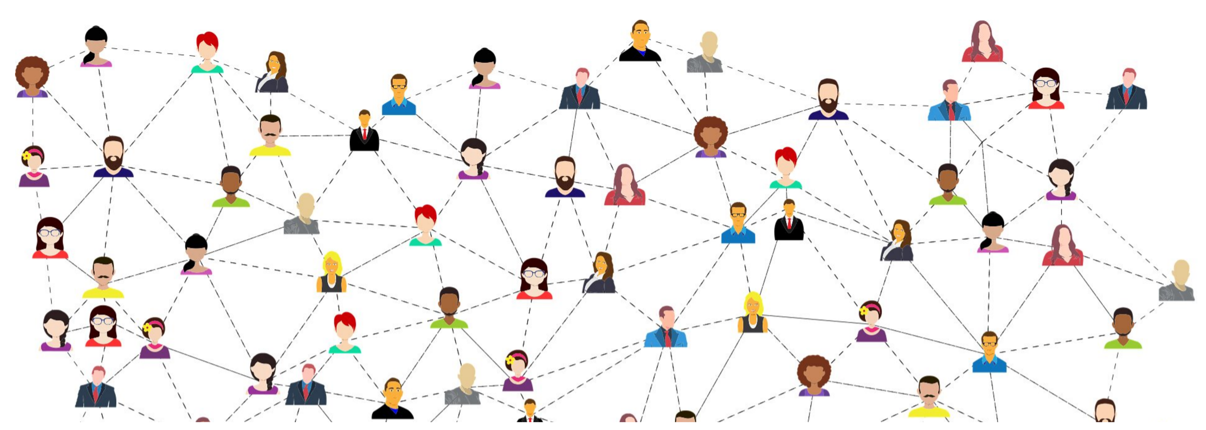
- Goal: infer (causal) average treatment effect ψ of treatment T on outcome Y

$$\psi = \mathbb{E}[Y \mid \mathrm{do}(T=1)] - \mathbb{E}[Y \mid \mathrm{do}(T=0)]$$

- Problem: unobserved confounder Z might affect T and Y
- Salvation: we observe a surrogate X for Z

The surrogate is nasty

- Network (non-iid) structure
  $\implies$ standard non-parametrics = :(
- no plausible generative model for $(X, \{Z_i\}, \{T_i\}, \{Y_i\})$:
  $\implies$ standard parametrics = :(



Example: Social Network

- Have treatments and outcomes of people in a social network.
- Maybe latent confounders; age, sex, affinity for punk rock.
- These are associated with the social network itself.

If life was easy, we'd observe $\{Z\}$

- estimate $\hat{Q}_n(t,z) \approx Q(t,z) = \mathbb{E}[Y \mid t, z]$
- estimate $\hat{g}_n(z) \approx g(z) = \mathrm{P}(T=1 \mid z)$
- plug-in to (non-parametrically efficient, robust) estimator

Or at least have a good generative model

- estimate $\{\hat{z}_{n,i}\}$.
- plug-in as though z observed.
- works if $\hat{z}_{n,i}$ is really good estimate.
- but generative models for networks suck $\implies$:(

[1] V. Veitch, M. Austern, W. Zhou, D.M. Blei, and P. Orbanz. Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data. 2018.

[2] L. Takac and M. Zabovsky. Data Analysis in Public Social Networks. 2012

Overview

- A network connecting units can be a surrogate for unobserved confounding
- Insight: identification only requires treatment and outcome predictions
- Causal adjustment achieved by adapting black-box node embeddings
- IF [pure homophily] and [black-box predictor works]
  THEN correct estimation at parametric rate

So What?

- Doesn't require knowledge of relationship between network and confounding
- Stop worrying and love the black box

Method

- Assign embedding $\lambda_i \in \mathbb{R}^p$ for each unit i
- Not justified by generative model, but kickass for prediction
- Train by *jointly* minimizing (we use [1] default)
  1. graph reconstruction on subgraphs
  2. predictor $\hat{g}^T(\lambda)$ for T from embedding (minimize CrossEnt)
  3. predictor $\hat{Q}^Y(t,\lambda)$ for Y from treatment t and λ (minimize MSE)
- Plug into modified A-IPTW estimator

$$\hat{\psi}_n^{\mathrm{A}} = \frac{1}{n}\sum_i \hat{Q}_n(1,z_i) - \hat{Q}_n(0,z_i) + \frac{1}{n}\sum_i \left(\frac{I[t_i=1]}{\hat{g}_n(z_i)} - \frac{I[t_i=0]}{1-\hat{g}_n(z_i)}\right)(y_i - \hat{Q}_n(t_i,z_i))$$

Assumptions and Validity

- IF Many technical conditions
- IF Pure homophily and observing Z would make identifiable
- IF The model learns something, eventually. For $t \in \{0,1\}$:

$$\|\mathbb{E}[Y \mid T=t, Z] - \tilde{Q}(t, \hat{\lambda}_{n,i}; \hat{\gamma}_n^Q)\|_2 \|\mathrm{P}(T=1 \mid Z) - \tilde{g}(\hat{\lambda}_{n,i}; \hat{\gamma}_n^g)\|_2 = o(\frac{1}{\sqrt{n}})$$

- Asymptotically, the embeddings are independent-ish. For all bounded continuous f with mean 0, $\mathbb{E}[f(\hat{\lambda}_i)f(\hat{\lambda}_j)] = o(\frac{1}{n})$

THEN

$\hat{\psi}_n^{\mathrm{A}}$ converges to the ATE ψ at $\sqrt{n}$-rate.

Example: Pokec

Semi-synthetic data using a large social network (Pokec) [2].

	age		district		join date	
Conf.	Low	High	Low	High	Low	High
Unadjusted	1.32 ± 0.02	4.34 ± 0.05	1.34 ± 0.03	4.51 ± 0.05	1.29 ± 0.03	4.03 ± 0.06
Parametric	1.30 ± 0.00	4.06 ± 0.01	1.21 ± 0.00	3.22 ± 0.01	1.26 ± 0.00	3.73 ± 0.01
Two-stage	1.33 ± 0.02	4.55 ± 0.05	1.34 ± 0.02	4.55 ± 0.05	1.30 ± 0.03	4.16 ± 0.06
$\hat{\psi}_n^{\mathrm{A}}$	1.24 ± 0.04	3.40 ± 0.04	1.09 ± 0.02	2.03 ± 0.07	1.21 ± 0.05	3.26 ± 0.09

Outcome and propensity simulated using labeled attribute as hidden confounder. Ground truth is 1. Network adjustment always helps relative to baseline.