# 1 Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks

Based on (Forastiere et al., 2021).

## 1.1 Background and motivation

- Interference: in experimental and observational studies, when a treatment assigned to one unit has an effect on others.
- Spillover effects: the effects of interference.
- Problem and goal: given a known network where the assignment mechanism of the treatment is unknown, estimate (1) the causal effect of individual treatment and (2) the spillover effect from treatments of others.
- Contributions of paper:
  1. A general formulization for the problem of interference in networks under the potential outcome framework.
  2. Derivation of the bias for estimators of the treatment effect when SUTVA is wrongly assumed.
  3. A joint propensity score (probability of assignment to particular individual and neighborhood treatment given observed covariates) with balancing properties, and a joint propensity score-based estimator.

## 1.2 Interference based on exposure to neighbourhood treatment

Notation:

- Undirected network $G = (\mathcal{N}, \mathbb{E})$ where $\mathcal{N}$ is a set of $N$ nodes and $\mathbb{E}$ is a set of edges $(i, j) = (j, i)$.
- Define partition $(i, \mathcal{N}_i, \mathcal{N}_{-i})$ around node $i$ where $\mathcal{N}_i$ is set of $N_i$ nodes (neighbourhood) that contains all nodes $j$ connected to $i$ and $\mathcal{N}_{-i}$ is set of all other nodes not $i$ and not in $\mathcal{N}_i$.
- $Z_i \in \{0, 1\}$ treatment assignment to unit $i$, $\mathbf{Z}$ treatment vector for population $\mathcal{N}$, and $(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}})$ partitions for $(i, \mathcal{N}_i, \mathcal{N}_{-i})$.
- $Y_i \in \mathcal{Y}$ observed outcome of unit $i$, $\mathbf{Y}$ outcome vector for population $\mathcal{N}$, and $(Y_i, \mathbf{Y}_{\mathcal{N}_i}, \mathbf{Y}_{\mathcal{N}_{-i}})$ partitions for $(i, \mathcal{N}_i, \mathcal{N}_{-i})$.
- $\mathbf{X}_i \in \mathcal{X}$ vector of covariates for unit $i$ and decomposes into $\mathbf{X}_i^{\text{ind}} \in \mathcal{X}^{\text{ind}}$ (individual-level characteristics) and $\mathbf{X}_i^{\text{neigh}} \in \mathcal{X}^{\text{neigh}}$ (neighbourhood-level characteristics and aggregates of individual-level covariates).

Potential outcomes and neighbourhood interference:

- Under presence of interference, the observable outcome at node $i$ is a function of the treatment assignment vector and can be written as $Y_i(\mathbf{Z})$. This is well-defined only if Assumption 1 holds.
- Assumption 1: if $\mathbf{Z} = \mathbf{z}$ then $Y_i = Y_i(\mathbf{z})$. (The outcome only depends on the treatment assignments and not the mechanism used to assign treatments.)
- Stable unit treatment value assumption (SUTVA): Assumption 1 and no interference between individuals (i.e., $Y_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = Y_i(Z_i, \mathbf{Z}'_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_{-i}})$ for all $\mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}'_{\mathcal{N}_{-i}}$).

- (Relaxing SUTVA Assumption 2 to allow neighbourhood interference)

  Assumption 2: for a $g_i : \{0,1\}^{N_i} \to \mathcal{G}_i$ s.t. $g_i(\mathbf{Z}_{\mathcal{N}_i}) = g_i(\mathbf{Z}'_{\mathcal{N}_i})$ for all $\mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_i}$, for all $\mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}'_{\mathcal{N}_{-i}}$ we have

  $$Y_i(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i}}) = Y_i(Z_i, \mathbf{Z}'_{\mathcal{N}_i}, \mathbf{Z}'_{\mathcal{N}_{-i}})$$

  (The outcome of node $i$ only depends on its treatment assignment and some summary of the treatment assignments of its neighbours, e.g., proportion.)

  Define $G_i = g_i(\mathbf{Z}_{\mathcal{N}_i})$ and assume $g_i$ known and well-specified.

- Stable unit treatment on neighbourhood value assumption (SUTNVA): Assumption 1 and Assumption 2.

Individual and neighbourhood treatments:

- Node $i$ is assigned to treatment if $Z_i = 1$ (control otherwise) and is exposed to neighbourhood treatment $G_i = g$ if $g_i(\mathbf{Z}_{\mathcal{N}_i}) = g$.
- Let $\mathbf{G}$ be the vector of neighbourhood treatments to which units are exposed to, $\mathbf{X}$ the covariate matrix collecting all vectors $\mathbf{X}_i$, and $\mathbf{Y}(z,g)$ the collection of potential outcomes $Y_i(z,g)$ for all units.
- The assignment mechanism can be written as

  $$\mathbb{P}(\mathbf{Z}, \mathbf{G}|\mathbf{X}, \{\mathbf{Y}(z,g), z \in \{0,1\}, g \in \mathcal{G}\}) = \begin{cases} \mathbb{P}(\mathbf{Z}|\mathbf{X}, \{\mathbf{Y}(z,g), z \in \{0,1\}, g \in \mathcal{G}\}) & \text{if } \mathbf{G} = \mathbf{g}(\mathbf{Z}) \\ 0 & \text{otherwise} \end{cases}$$

  where $\mathbf{g}(\mathbf{Z}) = [g_1(\mathbf{Z}_{\mathcal{N}_1}, \ldots, g_N(\mathbf{Z}_{\mathcal{N}_N}))]$.

Estimating main and spillover effects:

- A potential outcome $Y_i(z,g) = Y_i(Z_i = z, G_i = g)$ is defined only for a subset of nodes $V_g = \{i : g \in \mathcal{G}_i\}$ with cardinality $v_g$. For units with degree zero, denote $V_\emptyset = \{i : N_i = 0\}$ the set of units without neighbours.
- Super-population perspective: potential outcomes of graph $G$ are fixed and expectations are simple averages of these outcomes, i.e., $\mathbb{E}[\bullet|i \in U] = \frac{1}{|U|}\sum_{i \in U}(\bullet)$ and $\mathbb{P}(\bullet|i \in U) = \frac{1}{|U|}\sum_{i \in U} I(\bullet)$ for some subset $U$ of the super-population.
- Denote the marginal mean of the potential outcome $Y_i(z,g)$ in $V \subseteq V_g$ by

  $$\mu(z, g; V) = \mathbb{E}[Y_i(z,g)|i \in V]$$

  which can be viewed as an average dose-response function (ADRF) for subset $V$ depending on the individual and neighbourhood treatments.

- Causal effects are defined as comparisons between the marginal mean of different potential outcomes. The treatment effect (main effect) is defined as

  $$\tau(g) = \mu(1, g; V_g) - \mu(0, g; V_g)$$

  The overall main effect is defined as the average effect of individual treatments over the distribution of the neighbourhood treatment, i.e.,

  $$\tau = \sum_{g \in \mathcal{G}} \tau(g)\mathbb{P}(G_i = g) \qquad (= \mathbb{E}_{\mathcal{G}}[\tau(g)])$$

  where $\mathcal{G} = \bigcup_i \mathcal{G}_i$ and $i$ is a unit sampled from the population.

- The (causal) spillover effect of having the neighbourhood treatment set to level $g$ versus 0 when the unit is under treatment $z$ is
$$\delta(g;z) = \mu(z,g;V_g) - \mu(z,0;V_g)$$
The overall spillover effect is the average spillover effect over the distribution of the neighbourhood treatment, i.e.,
$$\Delta(z) = \sum_{g\in\mathcal{G}} \delta(g;z)\mathbb{P}(G_i = g) \qquad (= \mathbb{E}_{\mathcal{G}}[\delta(g;z)])$$

- The total effect is defined as
$$\begin{aligned}\text{TE} &= \sum_{g\in\mathcal{G}} \mathbb{E}\left[Y_i(1,g) - Y_i(0,0)|i\in V_g\right]\mathbb{P}(G_i=g)\\ &= \sum_{g\in\mathcal{G}} \mathbb{E}\left[Y_i(1,g) - Y_i(0,g) + Y_i(0,g) - Y_i(0,0)|i\in V_g\right]\mathbb{P}(G_i=g)\\ &= \tau + \Delta(0)\end{aligned}$$
i.e., the sum of overal main and spillover effects.

Unconfoundedness of the joint treatment:

- (Re-defined unconfoundedness assumption from SUTVA for SUTNVA)
Assumption 3: for all $z\in\{0,1\}$, $g\in\mathcal{G}_i$ and all $i$, $Y_i(z,g)\perp\!\!\!\perp Z_i, G_i|X_i$. (Given covariates for unit $i$, the potential outcome of unit $i$ given the treatment assignments for the graph is independent of the treatment assignment for unit $i$. Significance: units are not assigned to the treatment depending on the potential outcomes.)

- Theorem 1 (identification of ADRF): under Assumptions 1, 2, and 3,
$$\begin{aligned}\mathbb{E}\left[Y_i(z,g)|i\in V_g\right] &= \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\left[Y_i|Z_i=z,G_i=g,\mathbf{X}_i=\mathbf{x},i\in V_g\right]\mathbb{P}(\mathbf{X}_i=\mathbf{x}|i\in V_g)\\ &:= \bar{Y}^{\text{obs}}_{z,g}\\ &\left(= \mathbb{E}_{\mathcal{X}}\left[\mathbb{E}\left[Y_i|Z_i=z,G_i=g,\mathbf{X}_i=\mathbf{x},i\in V_g\right]|\, i\in V_g\right]\right)\end{aligned}$$
(Average potential outcome for subset $V_g$ is given by the weighted averaged of observed outcomes of units with $Z_i = z$ and $G_i = g$ and with same values of covariates. Significance: this allows estimation of the ADRF and therefore the causal effects of interest. If we only have a sample rather than the population, an unbiased estimator of the ADRF can be obtained from an unbiased estimator of conditional outcome mean $\bar{Y}^{\text{obs}}_{z,g}$.)

## 1.3 Bias when SUTVA is wrongly assumed

Naive estimator:

- Under SUTVA, the potential outcome is only indexed by the individual treatment assignment. The average treatment effect is then defined as
$$\tau_{\text{sutva}} = \mathbb{E}\left[Y_i(Z_i=1) - Y_i(Z_i=0)\right]$$

- Covariance-adjusted estimators estimate the quantity
$$\tau^{\text{obs}}_{X^*} = \sum_{\mathbf{x}\in\mathcal{X}^*} \mathbb{E}\left[Y_i|Z_i=1,\mathbf{X}^*_i\mathbf{x}\right] - \mathbb{E}\left[Y_i|Z_i=0,\mathbf{X}^*_i=\mathbf{x}\right]\mathbb{P}(\mathbf{X}^*_i=\mathbf{x})$$
where $\mathbf{X}^*_i\in\mathcal{X}^*$ is a subset of covariates (under SUTVA, neighbourhood covariates would not be considered and so $\mathbf{X}_i = \mathbf{X}^{\text{ind}}_i$). (If the SUTVA unconfoundedness assumption holds, then $\tau_{\text{sutva}} = \tau^{\text{obs}}_{X^*}$ and so an unbiased estimator of $\tau^{\text{obs}}_{X^*}$ is also an unbiased estimator of $\tau_{\text{sutva}}$.)

Bias of naive estimator when unconfoundedness holds:

- Theorem 2A: if Assumption 1 holds, Assumption 2 holds given $g_i$ for each unit $i$, Assumption 3 holds conditional on $\mathbf{X}_i^*$, then

$$\tau_{X^*}^{\text{obs}} = \sum_{\mathbf{x}\in\mathcal{X}^*} \left( \sum_{g\in\mathcal{G}} \mathbb{E}\left[Y_i(1,g)|\mathbf{X}_i^* = \mathbf{x}, i \in V_g\right] \mathbb{P}(G_i = g|Z_i = 1, \mathbf{X}_i^* = \mathbf{x}) \right.$$
$$\left. - \mathbb{E}\left[Y_i(0,g)|\mathbf{X}_i^* = \mathbf{x}, i \in V_g\right) \mathbb{P}(G_i = g|Z_i = 0, \mathbf{X}_i^* = x) \right) \mathbb{P}(\mathbf{X}_i^* = \mathbf{x})$$

- Corollary 1: under the three assumptions of Theorem 2A and the assumption $Z_i \perp\!\!\!\perp G_i|\mathbf{X}_i^*$, then $\tau_{X^*}^{\text{obs}} = \tau$. (If the individual and neighbourhood treatments are conditionally independent on the covariates, then covariate-adjusted estimates that assume SUTVA yield unbiased estimates even if SUTVA does not hold.)

- Corollary 2: under the three assumptions of Theorem 2A, if $Z_i \not\perp\!\!\!\perp G_i|\mathbf{X}_i^*$, an unbiased estimator of $\tau_{X^*}^{\text{obs}}$ would be biased for $\tau$. The bias depends on if the spillover effect at level $g$ versus $g'$ is dependent on the individual treatment assignment. (If there is residual correlation between the individual and neighbourhood treatments after conditioning on the covariates, covariate-adjusted estimates that assume SUTVA are biased.)

Bias of naive estimator when unconfoundedness does not hold:

- Theorem 2B: under Assumptions 1 and 2, if $Y_i(z,g) \not\perp\!\!\!\perp Z_i, G_i|\mathbf{X}_i^*$ but $Y_i(z,G) \perp\!\!\!\perp Z_i, G_i|\mathbf{X}_i^*, \mathbf{U}_i$ for some additional vector of covariates $\mathbf{U}_i \in \mathcal{U}$, then an unbiased estimator of $\tau_{X^*}^{\text{obs}}$ is biased for the $\tau$. The bias depends on if the spillover effect at level $g$ and $g'$ and if the unmeasured confounder $\mathbf{U}_i$ at level $u$ and $u'$ are dependent on the individual treatment.

- Corollary 3: under the three assumptions in Theorem 2B and if $Z_i \perp\!\!\!\perp G_i|X_i^*$, then an unbiased estimator of $\tau_{X^*}^{\text{obs}}$ is biased for the $\tau$ with the bias depending only on the unmeasured confounder $\mathbf{U}_i$.

- Corollary 4: if SUTVA holds and $Y_i(z,G) \perp\!\!\!\perp Z_i, G_i|\mathbf{X}_i^*, \mathbf{U}_i$, then an unbiased estimator of $\tau_{X^*}^{\text{obs}}$ is biased for the $\tau$ with the bias as in Corollary 3.

- Theorem 2B says that if SUTVA is wrongly assumed and adjusting for $\mathbf{X}_i^*$ is insufficient for unconfoundedness, then the bias is due to both interference and due to unmeasured confounders. If the individual and neighbourhood treatments are conditionally independent given the covariates, then the bias is only due to the unmeasured confounders.

## 1.4 Generalized propensity scores under neighbourhood interference

Joint propensity score:

- The joint propensity score is defined as the probability for unit $i$ being exposed to individual treatment $z$ and neighbourhood treatment $g$ given observed covariates $\mathbf{x}$ and denoted

$$\psi(z;g;x) = \mathbb{P}(Z_i = z, G_i = g|\mathbf{X}_i = \mathbf{x})$$

  Note that this differs from the unit-level treatment assignment probability $\mathbb{P}(Z_i = z, G_i = g|\mathbf{X}, \{\mathbf{Y}(z,g), z \in \{0,1\}; g \in \mathcal{G}\})$ unless the assignment mechanism only depends on the unit-level variables $\mathbf{X}_i$ (instead of $\mathbf{X}$) and the unconfoundedness assumption (Assumption 3) holds given $\mathbf{X}_i$.

- Proposition 1 (Balancing property): The joint propensity score is a balancing score, i.e.,

$$\mathbb{P}(Z_i = z, G_i = g|\mathbf{X}_i, \psi(z; g; \mathbf{X}_i)) = \mathbb{P}(Z_i = z, G_i = g|\psi(z; g; \mathbf{X}_i))$$

  (Distribution of covariates $\mathbf{X}$ is the same for units with the same $\psi(z; g; \mathbf{x})$.)

- Proposition 2 (Conditional unconfoundedness): If Assumption 3 holds given $\mathbf{X}_i$, then for all $z \in \{0, 1\}$, $g \in \mathcal{G}_i$,

$$Y_i(z, g) \perp\!\!\!\perp Z_i, G_i|\psi(z; g; \mathbf{X}_i)$$

  (If unconfoundedness holds, the potential outcome is independent of individual and neighbourhood treatment of units with the same joint propensity score.)

Individual and neighbourhood propensity score:

- The joint propensity score can be factorized as

$$\begin{aligned}\psi(z; g; \mathbf{x}) &= \mathbb{P}(G_i = g|Z_i = z, \mathbf{X}_i^g = \mathbf{x}^g)\mathbb{P}(Z_i = z|\mathbf{X}_i^z = \mathbf{x}^x) \\ &= \lambda(g; z; \mathbf{x}^g)\phi(z; \mathbf{x}^z)\end{aligned}$$

  where $\mathbf{X}_i^g \in \mathcal{X}^g \subset \mathcal{X}$ is the subset of covariates affecting the neighbourhood treatment and $\mathbf{X}_i^z \in \mathcal{X}^z \subset \mathcal{X}$ is the subset of covariates affecting the individual treatment. These subsets may not be mutually exclusive. $\lambda(g; z; \mathbf{x}^g)$ is the neighbourhood (generalized) propensity score and $\phi(z; \mathbf{x}^z)$ is the individual (binary) propensity score.

- Proposition 3 (Conditional unconfoundedness given individual and neighbourhood propensities): if Assumption 3 holds, then for all $z \in \{0, 1\}$, $g \in \mathcal{G}_i$,

$$Y_i(z, g) \perp\!\!\!\perp Z_i, G_i|\lambda(g; z; \mathbf{X}_i^g), \phi(1; \mathbf{X}_i^z)$$

  (TODO: is there a typo?)

## 1.5 Propensity score-based estimator for main and spillover effects

Individual and neighbourhood propensity score estimator:

- Theorem 1 says that an unbiased estimator of the conditional mean $\bar{Y}_{z,g}^{\text{obs}}$ is unbiased for $\mu(z, g; V)$. If there are many covariates or if the covariates are continuous, then estimation of $\bar{Y}_{z,g}^{\text{obs}}$ is challenging.

- If Assumption 3 holds, by Proposition 2, adjusting for the joint propensity score can also lead to an unbiased estimator for $\mu(z, g; V)$, i.e.,

$$\mathbb{E}\left[\mathbb{E}\left[Y_i|Z_i = z, G_i = g, \psi(z; g; \mathbf{X}_i)\right]|Z_i = z, G_i = g\right]$$

  where the outer expectation is over the empirical distribution of the joint propensity score in the population. By Proposition 3, the propensity score adjustment can be done separately for individual and neighbourhood, leading to the unbiased estimator

$$\mathbb{E}\left[\mathbb{E}\left[Y_i|Z_i = z, G_i = g, \phi(1; \mathbf{X}_i^z), \lambda(g; z; \mathbf{X}_i^g)\right]|Z_i = z, G_i = g\right]$$

Estimation procedure:

1. Subclassification based on individual propensity score:

   (a) Predict $\phi(1; \mathbf{X}_i^z)$ using a logistic regression for $Z_i$ fitted on covariates $\mathbf{X}_i^z$.

(b) Identify $J$ subclasses $B_j$ based on similar values of $\phi(1; \mathbf{X}_i^z)$ and where there is sufficient balanced between individual treatment groups, i.e., $\mathbf{X}_i^z \perp\!\!\!\perp Z_i | i \in B_j$ (given a unit in a subclass, covariates are independent of treatment).

2. For each subclass $B_j$, estimate $\mu_j(z, g; V_g) = \mathbb{E}\left[Y_i(z, g) | i \in B_j^g\right]$ where $B_j^g = V_g \cap B_j$:

   (a) Estimate parameters of a model for $\lambda(g, z; \mathbf{x}^g)$ where $\lambda(g; z; \mathbf{X}_i^g) = \mathbb{P}(G_i = g | Z_i = z, \mathbf{X}_i^g) = f^G(g, z, \mathbf{X}_i^g)$.

   (b) Use observed data $(Y_i, Z_i, G_i, \mathbf{X}_i^g)$ and $\hat{\Lambda} = \lambda(G_i; Z_i; \mathbf{X}_i^g)$ to estimate parameters of a model $Y_i(z, g) | \lambda(g; z; \mathbf{X}_i^g) \sim f^Y(z, g, \lambda(g; z; \mathbf{X}_i^g))$.

   (c) For a particular level of the joint treatment $(Z_i = z, G_i = g)$, for each unit $i \in B_j^g$, predict $\lambda(g; z; \mathbf{X}_i^g)$ and use it to predict $Y_i(z, g)$.

   (d) Estimate the dose-response function $\mu_j(z, g; V_g)$ by

$$\hat{\mu}_j(z, g; V_g) = \frac{\sum_{i \in B_j^g} \hat{Y}_i(z, g)}{|B_j^g|}$$

3. Estimate the ADRF by

$$\hat{\mu}(z, g, V_g) = \sum_{j=1}^{J} \hat{\mu}_j(z, g; V_g) \pi_j^g$$

where $\pi_j^g = \frac{|B_j^g|}{v_g}$ are weights proportional to the subclass size.

# References

Forastiere, L., Airoldi, E. M., & Mealli, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, *116*(534), 901–918. https://doi.org/10.1080/01621459.2020.1768100