Biometrika (2025), **112**, 1, asae041 https://doi.org/10.1093/biomet/asae041
Advance Access publication 20 August 2024

Estimating causal effects under non-individualistic treatments due to network entanglement

By P. TOULIS

Booth School of Business, University of Chicago,
5807 South Woodlawn Avenue, Chicago, Illinois 60637, U.S.A.
panos.toulis@chicagobooth.edu

A. VOLFOVSKY

Department of Statistical Science, Duke University,
Box 90251, Durham, North Carolina 27708, U.S.A.
alexander.volfovsky@duke.edu

and E. M. AIROLDI

Department of Statistics, Operations and Data Science, Fox School of Business,
Temple University, 1801 Liacouras Walk, Philadelphia, Pennsylvania 19122, U.S.A.
airoldi@temple.edu

Summary

In many observational studies, the treatment assignment mechanism is not individualistic, as it allows the probability of treatment of a unit to depend on quantities beyond the unit's covariates. In such settings, unit treatments may be entangled in complex ways. In this article, we consider a particular instance of this problem where the treatments are entangled by a social network among units. For instance, when studying the effects of peer interaction on a social media platform, the treatment on a unit depends on the change of the interactions network over time. A similar situation is encountered in many economic studies, such as those examining the effects of bilateral trade partnerships on countries' economic growth. The challenge in these settings is that individual treatments depend on a global network that may change in a way that is endogenous and cannot be manipulated experimentally. In this paper, we show that classical propensity score methods that ignore entanglement may lead to large bias and wrong inference of causal effects. We then propose a solution that involves calculating propensity scores by marginalizing over the network change. Under an appropriate ignorability assumption, this leads to unbiased estimates of the treatment effect of interest. We also develop a randomization-based inference procedure that takes entanglement into account. Under general conditions on network change, this procedure can deliver valid inference without explicitly modelling the network. We establish theoretical results for the proposed methods and illustrate their behaviour via simulation studies based on real-world network data. We also revisit a large-scale observational dataset on contagion of online user behaviour, showing that ignoring entanglement may inflate estimates of peer influence.

Some key words: Causal inference; Misspecification; Network; Non-individualistic assignment; Observational study; Peer influence; Propensity score; Randomization inference.

1. Introduction

In causal inference, the goal is usually to evaluate the effects of treatments applied individually to units (Imbens & Rubin, 2015, Ch. 3). However, when units form networks, the treatment is typically applied to pairs or groups of connected units and is therefore not individualistic.

Settings with such non-individualistic entangled treatments are common in many fields. For example, professional connections affect labour market outcomes (Montgomery, 1991, 1992; Podolny & Baron, 1997; Calvo-Armengol & Jackson, 2004) or knowledge diffusion and innovation (Topa, 2001; Granovetter, 2005; Kim & Marschke, 2005; Agrawal et al., 2006 and the 2003 Phd thesis from Aalborg University by M. S. Dahl); centrality in political networks affects coalition development (Keller, 2014); and online friendships have value in marketing (Ellison et al., 2007; Manchanda et al., 2015; Hobbs et al., 2016; Hanna et al., 2017) and affect how peer influence propagates (Aral et al., 2009).

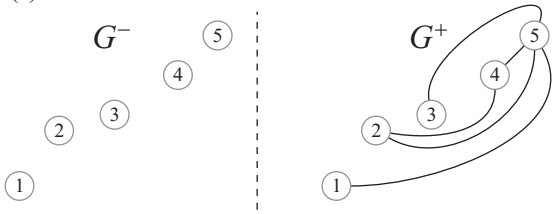
In these settings, treatment entanglement poses new methodological challenges that have not been addressed in the literature despite increased interest in evaluating treatment effects on networks. Traditionally, the concern about treatments on networks is interference (Cox, 1958; Rubin, 1974), where a unit's outcome can depend on other units' treatments. In recent years, a rich literature has emerged to deal with interference in statistics (Rosenbaum, 2007; Hudgens & Halloran, 2008; Bowers et al., 2013; Toulis & Kao, 2013; Ogburn et al., 2014; Aronow & Samii, 2017; Choi, 2017; Eckles et al., 2017; Sussman & Airoldi, 2017; Basse & Airoldi, 2018; Karwa & Airoldi, 2018; Basse et al., 2019; Jagadeesan et al., 2020; Puelz et al., 2022; Mathews & Volfovsky, 2023) and econometrics (Manski, 1993; Graham, 2008; Bramoullé et al., 2009; Manski, 2013; Angrist, 2014; Belloni et al., 2022; Vazquez-Bare, 2023), with a split focus on design and identification, respectively. However, treatment entanglement may still exist under no interference, and so the two problems are separate.

To illustrate the problem of entanglement, Fig. 1 depicts six users in a hypothetical professional network. The units form an empty network G^- at time t^-, and the network evolves endogenously to G^+ at time t^+. Suppose that the individual treatment on unit i, denoted by Z_i, is the number of new professional connections i makes from t^- to t^+ and is thus a function of G^- and G^+. Short-term outcomes $Y_i \in \mathbb{R}$ are measured for each i at t^+ and may represent, say, whether i moved to a higher-income job. Suppose that we want to estimate the causal effect of Z on Y, i.e., the effect of professional networking on job mobility. Owing to endogeneity, estimation of this causal effect may be confounded with units' covariates. For example, in Fig. 1 it would be tempting to associate improved wages with making new professional connections, but being more sociable confounds making more new connections and having better job outcomes.

One classical approach to mitigating such endogeneity bias relies on the propensity score methodology (Rosenbaum & Rubin, 1983, 1984, 2023; Heckman, 1990). When treatment is binary, the idea is to model the propensity score function, $\mathrm{pr}(Z_i = 1 \mid X_i)$, and then compare outcomes of units with similar propensities. However, the classical methodology tacitly assumes that the treatment is applied individually to each unit i. This approach may be biased under entanglement because in that setting the individual treatments depend on the change in the network from G^- to G^+, which is a population quantity.

In this article, we extend the classical propensity score methodology to settings with treatment entanglement. The idea is to model the propensity score of unit i by taking into account information from every other unit j that could connect to i during the

(a)



(b)

User	X_i	Z_i	Y_i
1	−5	1	0
2	−1	2	0
3	0	1	1
4	3	2	1
5	10	4	0

Fig. 1. (a) The networks before (G^-) and after (G^+) the presumed intervention. (b) Observed data: X_i is the covariate value for worker i; Z_i is the treatment on i, i.e., the number of new connections that unit i made in G^+; Y_i denotes the outcome of unit i, where $Y_i = 1$ if unit i's income increased after the treatment period and $Y_i = 0$ otherwise.

evolution from G^- to G^+. We also develop a nonparametric approach based on randomization methodology that, within a class of graphon-like network models (Borgs & Chayes, 2017), delivers valid inference without explicitly modelling the network. We show that this approach rectifies the classical propensity score method under a certain condition of network ignorability, and we illustrate its application through examples.

2. Preliminaries

2.1. *Definitions and assumptions*

Before specializing to networks, we present the general definition of treatment entanglement.

Definition 1 (Entangled treatments). *Treatments are said to be entangled if the assignment mechanism is not individualistic so that the probability of treatment assignment of a unit may depend on the treatment assignments of other units.*

Intuitively, Definition 1 describes entanglement as a form of interference between treatments. This definition is broad and encompasses large classes of study designs. For example, it includes completely randomized designs where the proportion of treated individuals is fixed a priori. The analysis of completely randomized designs generally accounts for this dependence (Imbens & Rubin, 2015).

We now specialize the definition to network settings. There are N units indexed by i. The units form a pre-treatment network G^- that evolves to the post-treatment network G^+. Let $N_i(G)$ denote the neighbourhood of unit i in $G \in \{G^-, G^+\}$ and $d_i(G) = |N_i(G)|$ the unit's degree, i.e., the number of immediate connections unit i has in G.

Definition 2 (Network-entangled treatments). *Treatments are said to be network-entangled if for each unit i its treatment Z_i is a function of G^- and G^+, i.e., $Z_i = f_i(G^-, G^+) \in \mathcal{Z}$ for a known function f_i.*

This definition aims to cover settings where the treatment is a function of the change in a network of units. The following are particular examples of f_i:

(i) $f_i(G^-, G^+) = d_i(G^+) - d_i(G^-)$, the change in the neighbourhood size of unit i;
(ii) $f_i(G^-, G^+) = \mathbb{I}\{d_i(G^+) > d_i(G^-)\}$, whether the neighbourhood of i grew;

(iii) $f_i(G^-, G^+) = \{\sum_{j \in V(G^+)} \text{dist}(i,j)\}^{-1} - \{\sum_{j \in V(G^-)} \text{dist}(i,j)\}^{-1}$ where $V(G)$ denotes the node set of graph G and $\text{dist}(i,j)$ is a measure of distance between nodes i and j in the network, a measure that captures, say, a change in individual closeness centrality.

To highlight the issue of entanglement, specification (i) defines treatment as

$$Z_i = f_i(G^-, G^+) = d_i(G^+) - d_i(G^-). \tag{1}$$

This is representative of several real-world applications; see § 2.2. In this specification, there is treatment entanglement because individual degrees are codependent; for instance, the sum of degrees of all units in a network needs to be even, and this constraint entangles the units' treatments. In § 5.2 and § 6 we consider a more general form of entanglement that is common in network seeding experiments. In such settings, Z_i is, not only a function of a change in the degree of individual units, but also a function of additional characteristics of individuals, such as whether they were given early access to a particular application. We provide details of this type of entanglement in those sections.

Throughout, $Z = (Z_i : i = 1, \ldots, N) \in \mathcal{Z}^N$ denotes the population treatment assignment vector. The potential outcome for unit i under treatment Z is denoted by $Y_i(Z) \in \mathbb{R}$. Each unit i has covariate $X_i \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^p$ and p is fixed.

The above definitions of treatment, such as the change in degree described in (1), generally lead to a multilevel treatment. We therefore need to extend the stable unit treatment value assumption, SUTVA (Rubin, 1980), also known as effective treatment in econometrics (Manski, 2013). Specifically, we assume that the value of Z_i affects only the outcome of i and that there are no hidden versions of the treatment.

Assumption 1 (*Multilevel treatment SUTVA*). For any population assignment vectors Z and Z', we have $Y_i(Z) = Y_i(Z')$ if $Z_i = Z'_i$ for every unit i.

With a slight abuse of notation, Assumption 1 allows us to write $Y_i(Z)$ as $Y_i(Z_i)$. Since Z_i takes only integer values by the definition in (1), let $\mathbb{Y}_i = \{Y_i(-N), \ldots, Y_i(N)\}$ denote all possible potential outcomes for unit i. The observed outcome for i is denoted by $Y_i \in \mathbb{Y}_i$.

Regarding causal estimands, we consider the following type of average treatment effect:

$$\tau_m = E\big\{Y_i(m) - Y_i(m-1)\big\}.$$

Here, the expectation may be either over the sample or over an infinite population. Under entanglement, this estimand captures the incremental causal effect from adding, or losing, m new connections in the treatment period over adding, or removing, $m-1$ new connections. With only two levels of treatment, the estimand is the classical average treatment effect.

2.2. *Examples*

We discuss some examples from the literature that exhibit treatment entanglement. These examples are used to illustrate our set-up and also to motivate the technical challenges.

Example 1 (Aral et al., 2009). Prior to the treatment of interest, a fraction of individuals in the population are provided with early access to a product, labelled $D_i = 1$. The treatment for unit i measures whether that unit is exposed to zero, one or more early adopters of the product. Thus, G^- and G^+ are graphs that represent exposures to product adoptions. Assuming for simplicity that G^- is empty, the treatment can be defined as

$$Z_i = \mathbb{I}\{d_i(G^+ \circ \vec{1}D^{\mathrm{T}}) > 0\} + \mathbb{I}\{d_i(G^+ \circ \vec{1}D^{\mathrm{T}}) > 1\}, \tag{2}$$

where $\vec{1}$ is the length-N vector of ones and $\circ$ denotes elementwise multiplication. Individual treatments are therefore entangled. For instance, two units that share a common neighbour that adopts the product are both exposed to the treatment together. We revisit this example in § 6.

Example 2 (Banerjee et al., 2013). A microfinance programme is introduced in some parts of a networked population, and then the information about this opportunity is diffused through the network. The goal is to understand peer effects on information diffusion. In this context, unit i is treated if i is informed about the programme by a friend. In this case, G^- and G^+ represent the directed interactions between units. This interaction network overlaps with the social network, but the two networks need not be identical. The definition of the treatment Z_i is the same as in (2), given the new definitions of G^- and G^+ as interaction networks.

Example 3 (Keller, 2014, 2015). These two papers study how different notions of network centrality, such as betweenness and closeness centrality, affect coalition formation. Considering individual centrality as a treatment leads to a form of probabilistic entanglement because the centralities of nearby units are correlated. While these studies do not include a formal causal analysis, their goal remains the same as in the other two examples, that is, to understand the effect of an entangled network treatment on individual-level outcomes.

3. Challenges under treatment entanglement

In our multilevel treatment setting with integer-valued treatments, the propensity score definition can be generalized to

$$e(l, X_i) = \mathrm{pr}(Z_i = l \mid X_i, G^-) \quad (l = -N, \ldots, N). \tag{3}$$

In standard methodology, conditional on similar propensity scores the treatment is as if randomly assigned under an appropriate ignorability assumption. This allows valid causal inference conditional on the propensity score (Rosenbaum & Rubin, 1983).

However, this standard approach ignores treatment entanglement. The subtle issue is that the probability in (3) implicitly conditions on G^+, since Z_i is a function of both G^- and G^+ by (1). Hence, the classical propensity score methodology is actually modelling $\mathrm{pr}(Z_i = l \mid X_i, G^+, G^-)$, not $\mathrm{pr}(Z_i = l \mid X_i, G^-)$ as claimed above. Conditioning on the post-treatment network and then estimating the propensity scores is incorrect because in the presence of entanglement the treatment is a function of the network.

One appropriate way to compute the propensity scores in (3) is to marginalize over the post-treatment network, accounting for uncertainty in G^+. This relies on a statement about the ignorability of treatment, which we formalize as follows.

Assumption 2 (*Ignorability under entanglement*). Let $\mathbb{Y} = (\mathbb{Y}_1, \ldots, \mathbb{Y}_N)$, where $\mathbb{Y}_i$ is the set of all possible potential outcomes of unit i. Then G^+ is conditionally independent of $\mathbb{Y}$ given pre-treatment information $X = (X_1, \ldots, X_N) \in \mathcal{X}^N$ and G^-; that is,

$$G^+ \perp\!\!\!\perp \mathbb{Y} \mid X, G^-.$$

This assumption is an extension of the standard ignorability assumption of treatment in settings with no entanglement (Rosenbaum & Rubin, 1983, §1.3). With this assumption in place, we can correctly calculate the propensity scores and get unbiased estimates of τ_m, according to the following theorem.

Theorem 1. *Suppose that Assumption 2 holds. Define the generalized propensity score as*

$$e(l, i; X) = \text{pr}(Z_i = l \mid X, G^-) = \int_{f_i(G^-, G^+)=l} \text{pr}(G^+ \mid G^-, X)\, \text{d}\mu(G^+), \tag{4}$$

where μ is a Lebesgue measure on G^+ and $\text{pr}(G^+ \mid G^-, X)$ is the network evolution model. Let $S_m(i; X) = \{e(m-1, i; X), e(m, i; X)\}$. If the network model is correctly specified and

$$0 < e(m-1, i; X), e(m, i; X) < 1$$

for all $m \in \{-N+1, \ldots, N\}$, units i and covariates X, then

$$\begin{aligned} &E\{Y_i \mid Z_i = m, S_m(i; X)\} - E\{Y_i \mid Z_i = m-1, S_m(i; X)\} \\ &\quad = E\{Y_i(m) - Y_i(m-1) \mid S_m(i; X)\}. \end{aligned}$$

The proof of Theorem 1 is provided in the Supplementary Material. Intuitively, the key result is that we can use the standard propensity score methodology as usual provided that we have computed the correct propensity scores in (4). In the literature, such propensity scores can be used in a variety of methods, including matching, subclassification and inverse weighting (Rosenbaum, 2002; Imbens & Rubin, 2015). Although here we suggest particular methods, the choice of the appropriate method is separate from the problem of entanglement.

Remark 1. Standard propensity score-based methods need to be adjusted to accommodate the fact that treatment is generally multilevel under entanglement. These problems, including inverse propensity score weighting approaches, have been partially addressed by Imbens (2000), Hirano & Imbens (2004), Cattaneo (2010), Lopez et al. (2017) and Lee (2018). Theorem 1 contributes to this literature, showing that to estimate τ_m one can use the classical methodology through the two-dimensional propensity score, $S_m(i; X)$.

Remark 2. The choice of network model, namely $\text{pr}(G^+ \mid X, G^-)$, is crucial but, ultimately, application-specific. Possible choices include simple rewiring models (Dietz & Hadeler, 1988), temporal exponential random graph models, which are a generalization of the ERGM framework of Hanneke & Xing (2007) and Hanneke et al. (2010), and dynamic latent space models (Sarkar & Moore, 2006; Durante & Dunson, 2014; Sewell & Chen, 2015). When it is reasonable to assume that G^- and $G^+ \setminus G^-$ are conditionally independent, one can appeal to the generalizability of latent space models (Hoff et al., 2002); that is, one can model G^- conditional on unit and dyadic covariates and use the fitted model to compute the probability of edges in $G^+ \setminus G^-$. In the next section, we develop a randomization-based procedure that under certain conditions can deliver valid inference without explicitly modelling the network.

4. Concrete methodology

4.1. *Estimation*

Theorem 1 implies that modelling the change in the network and then marginalizing over the treatment definition should allow proper causal estimation when treatments are entangled. Given a statistical model $\text{pr}(\cdot)$ of network evolution, this estimation can be accomplished via a simple Monte Carlo procedure as follows:

Step 1. Calculate the treatment assignments, $Z_i = f_i(G^-, G^+)$.

Step 2. Let $G^+ \mid G^-, X$ be modelled via $\text{pr}(\cdot \mid \theta, G^-, X)$. Obtain an estimate $\hat{\theta}$ of the model parameters.

Step 3. Use $\hat{\theta}$ to sample $G^+_{(b)}$ for $b = 1, \ldots, B$, conditional on the observed G^-.

Step 4. Use the samples from Step 3 to compute $\hat{e}(l, i; X)$ using the empirical frequencies:

$$\hat{e}_{i,l} = \hat{e}(l, i; X) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\{f_i(G^-, G^+_{(b)}) = l\}. \tag{5}$$

Although this procedure relies on a parametric network model, it is also possible to use nonparametric network models (Airoldi et al., 2013; Wolfe & Olhede, 2013; Borgs & Chayes, 2017) or models based on econometric or game-theoretic considerations (Galeotti et al., 2006; Jackson, 2010; Chandrasekhar & Lewis, 2011; Graham, 2015).

Given the estimates of propensity scores in (5), point estimation is possible via standard techniques. One approach would be to use the inverse-propensity weighted estimator,

$$\hat{\tau}^{\text{ipw}}_m(Z, Y) = \sum_{i=1}^{N} \frac{\mathbb{I}(Z_i = m)\, Y_i}{\hat{e}_{i,m}} - \sum_{i=1}^{N} \frac{\mathbb{I}(Z_i = m-1)\, Y_i}{\hat{e}_{i,m-1}}.$$

This estimator can be unbiased for τ_m as long as the propensity scores are consistently estimated. Another option is to use the standard subclassification estimator, which proceeds as follows:

Step 1. Subclassify units into K classes according to pairs $(\hat{e}_{i,m-1}, \hat{e}_{i,m})$; that is, units should be grouped together if the pair of propensity score values are similar. Let N_k denote the set of units in class k.

Step 2. Obtain estimates of τ_m within classes,

$$\hat{\tau}_{m,k} = \frac{1}{\sum_{i \in N_k} \mathbb{I}(Z_i = m)} \sum_{i \in N_k} \mathbb{I}(Z_i = m)\, Y_i - \frac{1}{\sum_{i \in N_k} \mathbb{I}(Z_i = m-1)} \sum_{i \in N_k} \mathbb{I}(Z_i = m-1)\, Y_i,$$

and combine estimates across classes into the estimator

$$\hat{\tau}^{\text{class}}_m(Z, Y) = \sum_{k=1}^{K} \frac{|N_k|}{N}\, \hat{\tau}_{m,k}. \tag{6}$$

A subclassification estimator such as (6) is generally more robust than the inverse-propensity weighted estimator (Imbens & Rubin, 2015, Ch. 17), but it also requires the specification of a clustering technique in Step 1. In § 5 we employ k-means clustering, but other options are available (Friedman et al., 2001). A comparative study of different clustering techniques in terms of bias/variance would be interesting, but we leave this for future work.

While point estimation is straightforward with these standard techniques, statistical inference is challenging under treatment entanglement. The key technical issue is that treatment assignment is not individualistic under entanglement. Variance estimation based on matching (Abadie & Imbens, 2006, 2016) or design-based sampling variation (Imbens & Rubin, 2015) cannot be applied because treatments are not independent across units. To overcome this challenge, we propose to use randomization inference, which we describe next.

4.2. *Randomization inference*

The key idea in randomization inference is to randomize treatment conditional on covariates and a particular null hypothesis. This approach is justified on the basis that treatment is as if randomized conditional on covariates in line with Assumption 2. The added benefit is that inference can be finite-sample valid for certain null hypotheses whenever exact matching on covariates is possible. This corresponds to settings where the propensity score model is well specified and relatively low-dimensional.

In particular, suppose that we want to test the global null hypothesis of no treatment effect,

$$H_0 : Y_i(Z) = Y_i(Z') \quad \text{for all } i, Z, Z'. \tag{7}$$

This hypothesis is, of course, restrictive as it immediately implies that $\tau_m = 0$ for all $m \geqslant 0$. However, it is useful as a building block for testing local null hypotheses related to τ_m for fixed m, which can be used to construct randomization-based confidence intervals for τ_m. To test the global null hypothesis, we can apply the classical Fisherian randomization test:

Step 1. Calculate the observed test statistic, $T^{\text{obs}} = T(Z, Y)$.

Step 2. Sample $G^+_{(r)} \sim \text{pr}(\cdot \mid \theta, G^-, X)$ according to the network model for $r = 1, \ldots, R$.

Step 3. Recalculate the treatment vector $Z^{(r)}$ where $Z^{(r)}_i = f_i(G^-, G^+_{(r)})$.

Step 4. Calculate the randomization p-value

$$\text{pval} = \frac{1}{1+R}\left[1 + \sum_{r=1}^{R} \mathbb{I}\left\{T(Z^{(r)}, Y) > T^{\text{obs}}\right\}\right].$$

This baseline procedure is valid in finite samples for the global null hypothesis H_0 in (7) and is feasible if the true model parameters θ are known. While this setting is limiting, the procedure can be extended to more general settings where knowing, or estimating, θ is not necessary, and also to more specialized null hypotheses. We consider both of these extensions next.

As the first extension, let G^- be empty without loss of generality, and suppose that the network model for G^+ follows a graphon specification (Borgs & Chayes, 2017); that is, edges are sampled independently and for every pair of nodes (i, j),

$$\text{pr}(g_{ij}^{+} = 1 \mid X) = w(X_i, X_j), \tag{8}$$

where $w(\cdot, \cdot)$ is a fixed, but unknown function. If, in addition, the covariate dimension p is relatively small, we may apply a simpler version of the above Fisherian randomization test to test the global null hypothesis via permutations of treatment levels within units having identical X values. Notably, this procedure is valid without requiring knowledge of $w(\cdot, \cdot)$, as long as all relevant covariates are observed.

More formally, let S_N denote the symmetric group. For any $\pi \in \mathsf{S}_N$, let πx denote the permutation of vector $x \in \mathbb{R}^p$ according to π; πX denotes the covariate matrix after permuting the rows of X according to π; and πG denotes the graph obtained from G by shuffling the nodes according to π. Define $\Pi(X)$ as the stabilizer group of permutations of units that leave X unchanged; that is, $\Pi(X) = \{\pi \in \mathsf{S}_N : \pi X = X\}$. Then the proposed permutation test replaces Steps 2 and 3 of the above Fisherian randomization test procedure with the following:

Step 2′. Resample the treatment as $Z^{(r)} = \pi^r Z$, where π^r is a random sample from $\Pi(X)$.

This results in a conditional Fisherian randomization test, since the randomization test described by Step 2′ resamples in the restricted space of treatment assignments defined by $\Pi(X)$. Similar conditional Fisherian randomization tests have been derived to study a different problem in network causal inference relating to outcome interference (Athey et al., 2018; Basse et al., 2019).

THEOREM 2. *Under the network model described by* (8)*, suppose that the population treatment vector,* $Z = f(G^{+}) = (f_i(G^{+}) : i = 1, \ldots, N) \in \mathcal{Z}^N$*, is equivariant such that*

$$f(\pi G^{+}) = \pi f(G^{+}) \quad \textit{for all } \pi \in \Pi(X), G^{+}.$$

Then, the permutation test described by Step 2′ *is finite-sample valid for* H_0 *in* (7)*.*

The key result in Theorem 2 is that permutation of the treatment levels of units that have identical X values is valid for testing the global null hypothesis. Notably, validity holds in finite samples and does not require estimation of the unknown graphon function $w(\cdot, \cdot)$ of the network model. Moreover, the permutation procedure is simple and can easily be scaled up to massive datasets, which we put to use in the application of § 6. The key condition identified by Theorem 2 is that the entanglement function needs to be equivariant. The importance of equivariance for permutation tests was highlighted recently by Basse et al. (2024), and our result can be considered an extension of theirs to the entanglement setting. Importantly, equivariance is a simple condition to check; in fact, all definitions of f_i in § 2 satisfy equivariance. To see this, observe that $(d_i(\pi G) : i = 1, \ldots, N) = \pi(G\vec{1}) = \pi\{d_i(G) : i = 1, \ldots N\}$. Consequently, any entanglement function that depends on degrees is equivariant. We give another example of treatment equivariance by revisiting Example 1.

Example 1 (*continued*). The treatment in this setting is defined in terms of the degree in a modification of the observed graph. First, notice that

$$d(G^{+} \circ \vec{1}D^{\mathrm{T}}) = (d_i(G^{+} \circ \vec{1}D^{\mathrm{T}}) : i = 1, \ldots, N) = (G^{+} \circ \vec{1}D^{\mathrm{T}})\vec{1} = \text{diag}(G^{+}D\vec{1}^{\mathrm{T}}) = G^{+}D,$$

where $\text{diag} : \mathbb{R}^{n \times n} \to \mathbb{R}^n$ retrieves the diagonal of its square matrix argument. Then, for any permutation $\pi \in \mathsf{S}_N$ we have that $d(\pi G^{+} \circ 1D^{\mathrm{T}}) = (\pi G^{+})D = \pi(G^{+}D) = \pi d(G^{+} \circ 1D^{\mathrm{T}})$, and so equivariance holds. From Theorem 2 we can test treatment effects between various levels through permutation tests conditional on covariates as described in Step 2′ above.

As a second extension, suppose that we are interested in testing a local sharp null hypothesis between two specific treatment levels,

$$H_0 : Y_i(m-1) = Y_i(m). \tag{9}$$

We encounter such local null hypotheses in the mobile app data example of § 6. In this setting, we wish to compare, for instance, the effects of communications between peers. The local null, as expressed in (9), immediately implies that $\tau_m = 0$, but is not as stringent as the global null hypothesis in (7). To test the local null hypothesis, we can simply replace Step 1 of the above Fisherian randomization test with the following:

Step 1′. Filter the data to include only units observed at treatment levels $m-1$ or m, i.e., the set $\{i : Z_i = m-1, m\}$. Calculate the test statistic using data from only those units.

Then we employ the permutation test described in Step 2′ above. The validity of this testing procedure is proved in the Supplementary Material.

Remark 3 (*Inference*). To perform inference on the local treatment effect τ_m, we could apply the method of test inversion (Rosenbaum, 2002, Ch. 10) to the randomization procedures developed in this section. Specifically, if we let $\text{pval}(u)$ denote the p-value from the randomization test on τ_m described by Step 1′, we can derive $\{u : \text{pval}(u) \geqslant \alpha\}$ as the $(1-\alpha)$-level randomization-based confidence interval for a constant local treatment effect.

Remark 4 (*Weak null hypotheses*). One additional important extension of the main Fisherian randomization test procedure is to test the 'weak local null' hypothesis $H_0 : \tau_m = 0$. This tests the equality in means between potential outcomes $Y(m-1)$ and $Y(m)$ rather than their equality in distribution. To test the weak null we could follow standard techniques (Chung & Romano, 2013; DiCiccio & Romano, 2017; Wu & Ding, 2021) and apply the permutation test described in Step 1′ with a test statistic that is normalized by a conservative estimate of its standard error. Such estimates are available for the subclassification estimator (Imbens & Rubin, 2015, § 17.6), though an adjustment would be needed to accommodate non-individualistic treatments.

Remark 5 (*Approximate Fisherian randomization test*). A second important extension of the main Fisherian randomization test deals with settings where a parametric form of the network model, $\text{pr}(\cdot)$, is available, but does not necessarily adhere to the graphon structure defined in (8). A general approach would then be to use an estimator of θ in Step 2, resulting in a procedure known as the approximate Fisherian randomization test. If this estimator of θ is consistent, then the resulting approximate Fisherian randomization test can have the correct level asymptotically, which has been proven for various individualistic treatment settings (Toulis, 2019; Berrett et al., 2020; Shaikh & Toulis, 2021; Pimentel, 2023). As before, proving this result under entanglement would need to account for the non-individualistic nature of treatment assignment. We provide empirical validation of this approach in § 5.2, reserving theoretical validation for future work.

5. NUMERICAL EXAMPLES

5.1. *Small multiplicative covariates simulation*

Consider again the example in Fig. 1. There are five units, each having a one-dimensional covariate $X_i \in \mathbb{R}$. The pre-treatment network G^- has no edges, as might be expected in

Table 1. *Propensity scores from two different models: the left panel is based the methodology described in § 4 using the true model (10), while the right panel is based on the misspecified Poisson regression in (11). Units enclosed by dashed lines are subclassified together as having similar propensities to receive $Z_i = 1$ and $Z_i = 2$; the misspecified model leads to incorrect subclassification and, consequently, bias in causal inference.*

	Propensity score for $Z_i = \cdots$							Propensity score for $Z_i = \cdots$						
Unit (i)	0	1	2	3	4	…	Unit (i)	0	1	2	3	4	…
1	0.00	0.27	0.73	0.00	0.00	…	1	0.37	0.37	0.18	0.06	0.02	…
2	0.00	0.24	0.67	0.09	0.00	…	2	0.24	0.34	0.25	0.12	0.04	…
3	0.01	0.06	0.23	0.42	0.28	…	3	0.21	0.33	0.26	0.13	0.05	…
4	0.00	0.24	0.68	0.09	0.00	…	4	0.13	0.26	0.27	0.19	0.10	…
5	0.00	0.27	0.73	0.00	0.00	…	5	0.02	0.08	0.15	0.20	0.20	…

a product adoption study, and the post-treatment network $G^+ = (g_{ij}^+)$ has a probability distribution such that the connection g_{ij}^+ between two units i and j is independent Bernoulli:

$$\text{pr}(g_{ij}^+ = 1 \mid G^-, X) \propto \exp(X_i X_j + 1.0). \tag{10}$$

Our goal is to use the data shown in Fig. 1 to estimate $\tau_2 = E\{Y_i(2) - Y_i(1)\}$, i.e., the causal effect of making two new connections relative to making just one.

The proposed method in § 4 requires conditioning on the propensity scores for making one or two connections. We compare two models for the propensity scores. The first relies on the true model in (10). The second method follows the classical propensity score approach, which ignores the network structure and instead fits a Poisson regression model. Based on the data in Fig. 1, the fitted model is

$$\text{pr}(Z_i = l \mid X_i) = \lambda_i^l \exp(-\lambda_i)/l! \quad (l = 0, 1, \ldots), \tag{11}$$

where $\log \lambda_i = 0.45 + 0.09 X_i$. The parameter estimates are rounded to two decimal places. Table 1 displays the estimated propensity scores from the two aforementioned models and outlines the resulting subclassification based on these estimates. The scores based on (10) are obtained by explicitly marginalizing over G^+, assuming that the underlying network model is known. Unsurprisingly, the subclassifications lead to different estimates of the causal effect: $\hat{\tau}_2 = 0.5$ using the true model (10), and $\hat{\tau}_2 = 0$ using the classical model (11). In absolute value the bias is 0.5, which is substantial because the range of estimands is $[-1, 1]$ as outcomes are binary.

The explanation for this bias is straightforward. The graph G^+ in the data of Fig. 1 is an unlikely sample from its true distribution implied by (10). Specifically, unit 3 has only one connection in G^+. However, from the left panel of Table 1 we get $E(Z_3) \approx 2.9$. As mentioned earlier, the classical methodology conditions on G^+ and thus underestimates the propensity scores for unit 3. Additionally, the unlikely large number of new connections for unit 5, $Z_5 = 4$, influences the Poisson model substantially, leading to the association of higher covariate X values with a higher number of connections. This contributes to underestimating the propensity scores of unit 3 and leads to wrong subclassification and biased estimates of the causal effect.

The example of this section highlights an important issue with the nonentangled approach. The Poisson-based propensity model does not incorporate the constraints imposed by the network topology, e.g., that the maximum degree of any unit is four. Without this, we may mischaracterize the space of possible treatments and underestimate the true propensities of individuals to select into treatment. Such constraints can potentially be included in the naive approach, but they clearly complicate computation and do not resolve the bigger-picture failure to model the entangled dependence between individual treatments.

5.2. *Large simulation study*

In this simulation study, we demonstrate that standard propensity score models can lead to severely biased causal inference whenever treatment assignment is not individualistic due to network entanglement. In contrast, methods that take entanglement into account perform better, even when they employ a misspecified treatment model.

Our simulation is designed to resemble the real-world application of the following section and is motivated by network seeding experiments (Kim et al., 2015; Chin et al., 2022). There are $N = 300$ units each with covariate X_i independent and identically distributed as Un[0, 1], indicating a latent social attribute. The units communicate with each other, forming a network G^+ where each edge is sampled in an independent and identically distributed manner as

$$\mathrm{pr}(g^+_{ij} = 1 \mid X) = \mathrm{expit}\big\{-\mu - a(X_i X_j)^{1/2} - \beta_{\mathrm{mis}}|X_i - X_j|\big\},$$

where $\mu, a, \beta_{\mathrm{mis}} > 0$. As before, G^- has no edges and G^+ is made to be symmetric. We set $\mu = 2$ and $a = 4$. With this definition, edges form mainly between units with smaller X values. The parameter β_{mis} controls the level of misspecification when fitting the treatment model, which we make precise later. Approximately 30% of the units are designated as seeds, via a label $D_i = 1$, and treatment describes whether someone was a seed and whether communication between seeds and non-seeds happened along G^+. The treatment can take three different levels and is defined as

$$Z_i = 2D_i + (1 - D_i)\,\mathbb{I}\{U_i < (G^+ D)_i - 3\}, \tag{12}$$

where the U_i are independent and identically distributed logistic random variables representing latent individual characteristics. This is a similar entanglement structure to that of Example 1 and the application of §6. The entangled treatment has the following intuitive interpretation:

$$Z_i = \begin{cases} 2 & \text{if unit } i \text{ was a seed;} \\ 1 & \text{if unit } i \text{ was not a seed, but communicated with many seeds;} \\ 0 & \text{if unit } i \text{ was not a seed and did not communicate with many seeds.} \end{cases}$$

We simulate outcome data based on pre-treatment covariate information:

$$Y_i(0) = -\beta_{\mathrm{con}} \exp\{-\beta_{\mathrm{con}}(X_i X_j)^{1/2}\} + \epsilon_i, \quad Y_i(1) = Y_i(0) + 10, \; Y_i(2) = Y_i(1) + \xi_i,$$

where $Y_i = \sum_{z=0}^{2} \mathbb{I}(Z_i = z)\, Y_i(z)$ is the outcome and $\epsilon_i, \xi_i \overset{\mathrm{iid}}{\sim} N(0, 1)$ are unobserved noise.

Table 2. *Simulation study of § 5.2: $N = 300$ units are treated according to the entangled treatment model in (12); each row is calculated as the average over 15 000 samples with fixed covariates X and G^+, but with varying D. The three panels show (a) percentage coverage of the true parameter ($\tau_1 = 10$) achieved by the subclassification estimator based on the propensity scores estimated by each method;* (b) *bias of each method;* (c) *root mean square error of each method.*

		(a) % Coverage			(b) Bias			(c) Root mean square error		
β_{con}	β_{mis}	Naive	Ent	Oracle	Naive	Ent	Oracle	Naive	Ent	Oracle
0.00	0.00	94.09	94.73	95.01	0.00	0.00	0.00	0.08	0.08	0.00
0.00	0.10	93.85	94.59	94.92	−0.00	−0.00	0.00	0.08	0.08	0.00
1.00	0.00	76.89	92.83	94.95	−0.54	−0.25	−0.06	0.73	0.38	0.07
1.00	0.10	75.18	92.47	94.88	−0.55	−0.24	−0.05	0.75	0.36	0.08
3.00	0.00	71.73	91.77	95.10	−3.72	−1.32	−0.18	4.89	2.10	0.35
3.00	0.10	70.50	92.05	95.22	−3.87	−1.32	−0.18	5.04	2.09	0.35

Ent, the entanglement-aware method.

The parameter β_{con} controls the level of confounding between potential outcomes and treatment. Our goal is to perform inference on $\tau_1 = E\{Y_i(1) - Y_i(0)\} = 10$.

We employ three different methods of estimating propensity scores, and use those propensity scores as plug-ins for the 10-class subclassification estimator of (6) and for inference on τ_1. For inference, we use the classical Fisherian randomization test outlined in § 4.2, adjusted to resample treatments according to the estimated propensity scores. To calculate these propensity scores for the entanglement-aware method defined below requires marginalizing over G^+ as prescribed by our main procedure in § 4.2. The model structure in (12) implies that the Z_i are conditionally independent given the seed status D_i, and so the randomization test resampling treatment based on the propensity scores is valid.

We report the bias, root mean square error and coverage results for the following propensity score methods.

(i) Naive method: this fits a multinomial model $Z \sim X + D$ using the `multinom` package in `R` (R Development Core Team, 2024). As an aside, the results in this section remain unchanged if the naive model omits D.

(ii) Entanglement-aware method: this method follows the Monte Carlo approach of § 4.2 by fitting the treatment model

$$\text{pr}(g_{ij}^+ = 1 \mid X) = \text{expit}\{-\mu - a(X_i X_j)^{1/2}\}.$$

Whenever $\beta_{\text{mis}} = 0$, this corresponds to a correctly specified model. However, the model is misspecified when $\beta_{\text{mis}} \neq 0$. This allows us to check the robustness of our entanglement-aware methodology.

(iii) Oracle method: this uses the true definition of the treatment vector in (12) and calculates propensity scores assuming that G^+ and D are known.

The naive and entanglement-aware methods are fit conditionally on all observable quantities, captured by X and D. Importantly, neither method has access to the full network G^+.

The results are shown in Table 2. First, we see that all methods, including the naive method that assumes individualistic treatment, achieve the nominal coverage level when there is

no confounding, i.e., $\beta_{\text{con}} = 0$. This is true even when the treatment model fitted by the entanglement method is misspecified, i.e., $\beta_{\text{mis}} \neq 0$. However, the naive method can undercover severely when there is confounding, $\beta_{\text{con}} \neq 0$. For example, when $\beta_{\text{con}} = 3$, the naive method has a coverage rate of the true parameter that can be as low as 71%. Moreover, model misspecification, i.e., $\beta_{\text{mis}} \neq 0$, appears to introduce more confounding, resulting in additional deterioration of the performance of the naive method. On the other hand, the entanglement-aware method performs better, with a coverage rate that generally exceeds 92% across all settings. Notably, this method does not deteriorate under model misspecification, $\beta_{\text{mis}} \neq 0$. Predictably, the oracle method that uses the true propensity scores covers at the nominal level throughout all settings.

The bias and root mean square error results shown in Table 2(b) and (c) are analogous to the coverage results. We point out that the bias for the oracle method, e.g., -0.18 for $\beta_{\text{con}} = 3$ and $\beta_{\text{mis}} = 0.1$, is due to finite samples. Despite this bias, the oracle method achieves nominal coverage because it uses the correct randomization distribution of the estimator based on the true propensity scores.

6. Application

6.1. *Introduction*

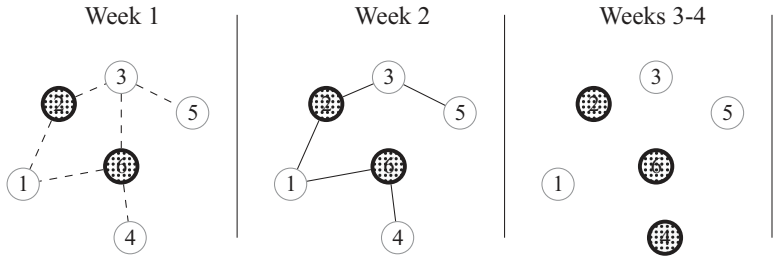
In this section we revisit Example 1 and analyse data from Aral et al. (2009). These data describe use of the Yahoo! Go mobile service in a social network of users over time. The key finding of Aral et al. (2009) is that adjusting for the individual propensity score to adopt the mobile service can help to distinguish between peer effect and homophily. Our goal here is different as we focus on estimation of causal effects, and in particular we aim to illustrate the differences between using a standard propensity score model that ignores the presence of entanglement and using a model that takes entanglement into account.

6.2. *Data and definitions*

The dataset consists of a universe of 13.02 million Yahoo! users in the U.S.A. tracked over the month of October 2007. Up to this point the vast majority of users, 98.3%, had not used the Go service that had been launched the year before. About 60% of the users were male with an age range of 18–70 years and an average age of 30. In week 1 of the month, 115.6 thousand users adopted the Go service, and we refer to these users as seeds, in the network seeding sense discussed in § 5.2. The units of our analysis are users who are not seeds and who communicated with someone else, seed or non-seed, during week 2 through Yahoo! Messenger. There are roughly 3.54 million units in the dataset. For each unit we have covariates X including the following: age; sex, recorded as binary; country location; full communication history; and Go activity, i.e., summary page view counts, if they were Go users.

These communications are recorded as messages between pairs of units. During week 2, a total of 178 million messages were exchanged over 10.5 million distinct conversations. The distribution of communications per user is severely right-skewed, with a median of 54 messages and a mean of 234 messages. There is also a user with more than 125 000 recorded messages. Additional descriptive statistics and visualization are included in the Supplementary Material.

The treatment of a unit has three levels, which indicate whether the unit communicated with 0, 1 or more seeds during week 2. With a slight adjustment to our notation, let G^+ be the



Summary statistics	
Users	13.02m
Average age	29.9
% male	60.2%
Seeds	115.6k
Units	3.54m
Messages	178m
Adopters	9.6k

Fig. 2. Yahoo! Go example. Week 1: users have potential communication channels marked with dashed edges; some users, i.e., 2 and 6, adopt the Go mobile app and form the seeds (marked with dots), while the rest of the users form the units. Week 2: users exchange messages with each other (solid edges); the treatment Z denotes the number of seeds a unit communicated with, e.g., $Z_1 = 2$ and $Z_4 = 1$. Weeks 3 and 4: outcomes Y are measured on the units and capture whether they adopt Go, e.g., $Y_4 = 1$ (unit 4 adopted Go) whereas $Y_1 = 0$.

communication network in week 2, i.e., $g_{ij}^{+} = 1$ means that user i exchanged messages with user j. As discussed above, this communication network is vast and contains 10.5 million edges across 3.55 million nodes. Then, for every unit i we define

$$Z_i = \mathbb{I}\Big(\sum_{j \in \text{seeds}} g_{ij}^{+} > 0 \Big) + \mathbb{I}\Big(\sum_{j \in \text{seeds}} g_{ij}^{+} > 1 \Big).$$

The outcome is binary and indicates whether a unit adopted the Go service in weeks 3 and 4 of October 2007. A total of 9581 units adopted the Go service in weeks 3 and 4, which we refer to as adopters. We visualize our applied example, along with key summary statistics, in Fig. 2.

Remark 6. A seed may communicate with units on the basis of similar age or gender, or other characteristics related to the user browsing interests. Our goal is to estimate the effect of communication with a peer who adopted Go in week 1, and so entanglement is likely owing to an underlying social network that can exhibit homophilous patterns of communication between peers and motifs such as triadic closures. This form of entanglement is similar to that of Example 1 and § 5.2, allowing us to employ our randomization approach for inference.

6.3. *Methods and results*

To analyse the data, we employ the permutation inference procedure of § 4.2 that does not require estimation of a particular network model, i.e, Step 1′. Both the standard approach and the entanglement-aware approach use exact matching, each using a different set of covariates.

For the standard individualistic approach, we perform exact matching on age and sex. For the entanglement-aware approach we construct additional covariates that capture the communication profile of each user in week 2. Specifically, for every communication between units i and j in week 2, we calculate whether the units are of similar age, $\texttt{same\_age}_{ij} = 1$, and whether the units are of the same sex, $\texttt{same\_sex}_{ij} = 1$. Then, for every unit i, the model includes four additional variables counting the number of communications between i and other units, for all four possible subgroups defined by $(\sum_j \texttt{same\_age}_{ij}, \sum_j \texttt{same\_sex}_{ij})$, where the sum is over all units that i communicated with. These covariates are normalized

Table 3. *Point estimates and randomization-based confidence intervals for constant treatment effects based on a standard propensity score model that ignores treatment entanglement and is based on an entanglement-aware model*

Model		$H_0^{\{0,1\}}$: $Y_i(0) = Y_i(1)$	$H_0^{\{1,2\}}$: $Y_i(1) = Y_i(2)$
Standard	Estimate	0.32%	0.32%
	95% confidence interval	[0.22%, 0.43%]	[0.18%, 0.48%]
Entanglement	Estimate	0.68%	−0.21%
	95% confidence interval	[0.54%, 0.80%]	[−0.45%, 0.01%]

as proportions, in percentages, over the total communications of unit i with all other units, thus forming the communication profile of unit i.

To enforce positivity, we eliminate matched groups in which fewer than 5% or more than 95% of the units are treated (Lee et al., 2011); this removes 11% of the data. As mentioned before, we conduct inference based on the randomization framework of §4.2 using the subclassification estimator as the test statistic. Specifically, we test two sharp null hypotheses, namely $H_0^{\{0,1\}} : Y_i(1) = Y_i(0)$ and $H_0^{\{1,2\}} : Y_i(1) = Y_i(2)$. For each test, we condition on the units that receive one of the two treatment levels, construct matches and permute the treatment vector within every match. We then invert these tests to construct randomization-based confidence intervals for a constant treatment effect; see Remark 3. This approach leverages Theorem 2, which guarantees that the restricted permutation test on Z is equivalent to a test that marginalizes over the post-treatment network, G^+.

As a result, the standard approach may permute treatments between units with potentially very different communication profiles, whereas the entanglement-aware approach permutes only between units with identical profiles. This may be important because the communication profile seems to be a significant factor for user-to-user communication in Yahoo! Messenger, which could bias the results from the standard model. Importantly, in settings where the standard model is valid, the entangled model would also be valid since every randomization within the entanglement-aware model is also valid under the standard model.

The results are shown in Table 3. For the standard model, we see that the point estimates are positive and very similar for both effects at 0.32%. The randomization-based 95% confidence intervals under this model are also similar and range from 0.18% to 0.48%. This implies a strong effect in adopting Go from communicating with one seed versus communicating with no one, the so-called τ_0 effect, and the effect is also positive when comparing communication with multiple seeds versus communication with one seed, known as the τ_1 effect. However, the results are substantively and statistically different under the entanglement-aware model and also suggest strong effect heterogeneity. Specifically, the τ_0 effect is stronger under the entanglement model and ranges between 0.54% and 0.80%. On the other hand, the point estimate for the τ_1 effect is negative, −0.21%, but not statistically significant. This suggests that ignoring entanglement may result in grouping together highly dissimilar units, leading to possible bias.

One potential criticism of the nonsignificant result in Table 3 is that the randomization-based confidence intervals may be conservative. To address this, we investigate the Type II error of the randomization procedure through a power simulation study calibrated on the real data. The results of this study are reported in the Supplementary Material and show that the randomization procedure is high-powered for effects of similar magnitude to those observed in the real data.

7. Discussion

This article studies the problem of treatment entanglement in causal inference in the context of network data. This leads to non-individualistic treatment assignment, which has been largely ignored by standard causal inference. Our work, however, leaves several open problems.

First, it would be interesting to know theoretically the extent of bias of classical propensity score methodology under non-individualistic, entangled treatments, as well as the bias reduction achieved by entanglement-aware methods. In § 5.2 we showed empirically that even misspecified network models can reduce bias relative to standard propensity scores. Determining the kinds of theoretical conditions that guarantee such reduction is a question for future research, including possible sensitivity analysis. Second, as discussed in § 4, it would be interesting to know how to select appropriate network models. Third, subclassification on a multilevel propensity score as in Theorem 1 is never exact, and we did not address the resulting bias from subclassification error. Finally, there are interesting open problems concerning randomization tests for weak null hypotheses under entanglement, as discussed in Remarks 4 and 5.

As a concluding remark, while our focus has been on treatments that occur between two individuals in a network, in practice we often observe treatments on larger subsets of units. This scenario is common in fields such as education and business, where entire subsets of a school or a sector receive a treatment while being connected to other subsets. Since the fundamental building block for such treatment assignment is the underlying network, we believe that our methodology is general enough to encompass it. In particular, after adjusting the definition of treatment assignment to apply to larger subgraphs, the procedure outlined in § 4 could be applied directly. This flexibility allows our methodology to adapt to various real-world scenarios where treatment occurs within interconnected subsets of a network.

Acknowledgement

This work was partially supported by grants from the U.S. National Science Foundation (CAREER IIS-1149662 and IIS-1409177) and the Office of Naval Research (YIP N00014-14-1-0485 and N00014-17-1-2131) to Harvard University, by a John E. Jeuck Faculty Fellowship awarded to Toulis, by grants from the U.S. National Institutes of Health (R01-1R01EB025021), Army Research Institute (W911NF1810233) and National Science Foundation (CAREER DMS-2046880 and DMS-2230074) to Duke University, and by a Shutzer Fellowship and an Alfred Sloan Research Fellowship awarded to Airoldi. Airoldi is also affiliated with the Data Science Institute at Temple University.

Supplementary material

The Supplementary Material includes the proofs of Theorems 1 and 2 as well as power simulation results.

References

Abadie, A. & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–67.

Abadie, A. & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica* **84**, 781–807.

Agrawal, A., Cockburn, I. & McHale, J. (2006). Gone but not forgotten: Knowledge flows, labor mobility, and enduring social relationships. *J. Econ. Geogr.* **6**, 571–91.

Airoldi, E. M., Costa, T. B. & Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Proc. 26th Int. Conf. Neural Information Processing Systems*. Red Hook, New York: Curran Associates, pp. 692–700.

Angrist, J. D. (2014). The perils of peer effects. *Labour Econ.* **30**, 98–108.

Aral, S., Muchnik, L. & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Nat. Acad. Sci.* **106**, 21544–9.

Aronow, P. M. & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Statist.* **11**, 1912–47.

Athey, S., Eckles, D. & Imbens, G. W. (2018). Exact *p*-values for network interference. *J. Am. Statist. Assoc.* **113**, 230–40.

Banerjee, A., Chandrasekhar, A. G., Duflo, E. & Jackson, M. O. (2013). The diffusion of microfinance. *Science* **341**, 1236498.

Basse, G., Ding, P., Feller, A. & Toulis, P. (2024). Randomization tests for peer effects in group formation experiments. *Econometrica* **92**, 567–90.

Basse, G. W. & Airoldi, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika* **105**, 849–58.

Basse, G. W., Feller, A. & Toulis, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106**, 487–94.

Belloni, A., Fang, F. & Volfovsky, A. (2022). Neighborhood adaptive estimators for causal inference under network interference. *arXiv:* 2212.03683.

Berrett, T. B., Wang, Y., Barber, R. F. & Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *J. R. Statist. Soc.* B **82**, 175–97.

Borgs, C. & Chayes, J. (2017). Graphons: A nonparametric method to model, estimate, and design algorithms for massive networks. In *Proc. 2017 ACM Conf. Economics and Computation*. New York: Association for Computing Machinery, pp. 665–72.

Bowers, J., Fredrickson, M. M. & Panagopoulos, C. (2013). Reasoning about interference between units: A general framework. *Polit. Anal.* **21**, 97–124.

Bramoullé, Y., Djebbari, H. & Fortin, B. (2009). Identification of peer effects through social networks. *J. Economet.* **150**, 41–55.

Calvo-Armengol, A. & Jackson, M. O. (2004). The effects of social networks on employment and inequality. *Am. Econ. Rev.* **94**, 426–54.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *J. Economet.* **155**, 138–54.

Chandrasekhar, A. & Lewis, R. (2011). Econometrics of sampled networks. Unpublished manuscript, MIT.[422].

Chin, A., Eckles, D. & Ugander, J. (2022). Evaluating stochastic seeding strategies in networks. *Manag. Sci.* **68**, 1714–36.

Choi, D. (2017). Estimation of monotone treatment effects in network experiments. *J. Am. Statist. Assoc.* **112**, 1147–55.

Chung, E. Y. & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41**, 484–507.

Cox, D. R. (1958). *Planning of Experiments*. New York: John Wiley & Sons.

DiCiccio, C. J. & Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *J. Am. Statist. Assoc.* **112**, 1211–20.

Dietz, K. & Hadeler, K. (1988). Epidemiological models for sexually transmitted diseases. *J. Math. Biol.* **26**, 1–25.

Durante, D. & Dunson, D. B. (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika* **101**, 883–98.

Eckles, D., Karrer, B. & Ugander, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Infer.* **5**, DOI: 10.1515/jci–2015–0021.

Ellison, N. B., Steinfield, C. & Lampe, C. (2007). The benefits of Facebook 'friends:' social capital and college students' use of online social network sites. *J. Computer-Mediated Commun.* **12**, 1143–68.

Friedman, J., Hastie, T. & Tibshirani, R. (2001). *The Elements of Statistical Learning*, vol. 1. New York: Springer.

Galeotti, A., Goyal, S. & Kamphorst, J. (2006). Network formation with heterogeneous players. *Games Econ. Behav.* **54**, 353–72.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica* **76**, 643–60.

Graham, B. S. (2015). Methods of identification in social networks. *Annu. Rev. Econ.* **7**, 465–85.

Granovetter, M. (2005). The impact of social structure on economic outcomes. *J. Econ. Perspect.* **19**, 33–50.

Hanna, B., Kee, K. F. & Robertson, B. W. (2017). Positive impacts of social media at work: Job satisfaction, job calling, and Facebook use among co-workers. SHS Web of Conferences, **33**. doi: 10.1051/shsconf/20173300012.
Hanneke, S., Fu, W. & Xing, E. P. (2010). Discrete temporal models of social networks. *Electron. J. Statist.* **4**, 585–605.
Hanneke, S. & Xing, E. P. (2007). Discrete temporal models of social networks. In *Statistical Network Analysis: Models, Issues, and New Directions (Proc. ICML 2006 Workshop, Pittsburgh, PA, USA)*. Berlin: Springer, pp. 115–25.
Heckman, J. (1990). Varieties of selection bias. *Am. Econ. Rev.* **80**, 313–18.
Hirano, K. & Imbens, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Chichester: Wiley & Sons, pp. 73–84.
Hobbs, W. R., Burke, M., Christakis, N. A. & Fowler, J. H. (2016). Online social integration is associated with reduced mortality risk. *Proc. Nat. Acad. Sci.* **113**, 12980–4.
Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.* **97**, 1090–8.
Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *J. Am. Statist. Assoc.* **103**, 832–42.
Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706–10.
Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.
Jackson, M. O. (2010). *Social and Economic Networks*. Princeton, New Jersey: Princeton University Press.
Jagadeesan, R., Pillai, N. S. & Volfovsky, A. (2020). Designs for estimating the treatment effect in networks with interference. *Ann. Statist.* **48**, 679–712.
Karwa, V. & Airoldi, E. (2018). A systematic investigation of classical causal inference strategies under mis-specification due to network interference. *ArXiv*: 1810.08259.
Keller, F. (2014). A theory for networks of power: Coalition formation on networks. https://api.semanticscholar.org/CorpusID:53504674.
Keller, F. (2015). Networks of power: Using social network analysis to understand who will rule and who is really in charge in the Chinese Communist Party. https://api.semanticscholar.org/CorpusID:202586017.
Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H. & Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: A cluster randomised controlled trial. *Lancet* **386**, 145–53.
Kim, J. & Marschke, G. (2005). Labor mobility of scientists, technological diffusion, and the firm's patenting decision. *RAND J. Econ.* **36**, 298–317.
Lee, B. K., Lessler, J. & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS One* **6**, e18174.
Lee, Y.-Y. (2018). Efficient propensity score regression estimators of multivalued treatment effects for the treated. *J. Economet.* **204**, 207–22.
Lopez, M. J. & Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statist. Sci.* **32**, 432–54.
Manchanda, P., Packard, G. & Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Market. Sci.* **34**, 367–87.
Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Studies* **60**, 531–42.
Manski, C. F. (2013). Identification of treatment response with social interactions. *Economet. J.* **16**, S1–23.
Mathews, H. & Volfovsky, A. (2023). Community informed experimental design. *Statist. Meth. Applic.* **32**, 1141–66.
Montgomery, J. D. (1991). Social networks and labor-market outcomes: Toward an economic analysis. *Am. Econ. Rev.* **81**, 1408–18.
Montgomery, J. D. (1992). Job search and network composition: Implications of the strength-of-weak-ties hypothesis. *Am. Sociol. Rev.* **57**, 586–96.
Ogburn, E. L. & VanderWeele, T. J. (2014). Causal diagrams for interference. *Statist. Sci.* **29**, 559–78.
Pimentel, S. D. (2023). Covariate-adaptive randomization inference in matched designs. *arXiv:* 2207.05019v2.
Podolny, J. M. & Baron, J. N. (1997). Resources and relationships: Social networks and mobility in the workplace. *Am. Sociol. Rev.* **62**, 673–93.
Puelz, D., Basse, G., Feller, A. & Toulis, P. (2022). A graph-theoretic approach to randomization tests of causal effects under general interference. *J. R. Statist. Soc.* B **84**, 174–204.
R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Rosenbaum, P. & Rubin, D. (2023). Propensity scores in the design of observational studies for causal effects. *Biometrika* **110**, 1–13.

Rosenbaum, P. R. (2002). Observational studies. In *Observational Studies*. New York: Springer, pp. 1–17.
Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *J. Am. Statist. Assoc.* **102**, 191–200.
Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Assoc.* **79**, 516–24.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701.
Rubin, D. B. (1980). Comment on 'Randomization analysis of experimental data: The Fisher randomization test'. *J. Am. Statist. Assoc.* **75**, 591–3.
Sarkar, P. & Moore, A. W. (2006). Dynamic social network analysis using latent space models. In *Proc. 18th Int. Conf. Neural Information Processing Systems*. Cambridge, Massachusetts: MIT Press, pp. 1145–52.
Sewell, D. K. & Chen, Y. (2015). Latent space models for dynamic networks. *J. Am. Statist. Assoc.* **110**, 1646–57.
Shaikh, A. M. & Toulis, P. (2021). Randomization tests in observational studies with staggered adoption of treatment. *J. Am. Statist. Assoc.* **116**, 1835–48.
Sussman, D. L. & Airoldi, E. M. (2017). Elements of estimation theory for causal effects in the presence of network interference. *arXiv:* 1702.03578.
Topa, G. (2001). Social interactions, local spillovers and unemployment. *Rev. Econ. Studies* **68**, 261–95.
Toulis, P. (2019). Life after bootstrap: Residual randomization inference in regression models. *arXiv:* 1908.04218.
Toulis, P. & Kao, E. (2013). Estimation of causal peer influence effects. *Proceedings of the 30th International Conference on Machine Learning, in Proceedings of Machine Learning Research.* PMLR, Vol. 28, pp. 1489–97. Available from https://proceedings.mlr.press/v28/toulis13.html.
Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *J. Economet.* **237**, 105237.
Wolfe, P. J. & Olhede, S. C. (2013). Nonparametric graphon estimation. *arXiv:* 1309.5936.
Wu, J. & Ding, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *J. Am. Statist. Assoc.* **116**, 1898–913.

[*Received on* 4 *March* 2021. *Editorial decision on* 13 *June* 2024]