2018 IEEE 5th International Conference on Data Science and Advanced Analytics

Estimating Causal Effects On Social Networks

Laura Forastiere *, Fabrizia Mealli †, Albert Wu ‡ and Edoardo M. Airoldi‡
* *Yale Institute for Network Science*
Yale University,
New Haven, USA
laura.forastiere@yale.edu
† *Department of Statistics, Computer Science, Applications*
University of Florence
Florence, Italy
mealli@disia.unifi.it
‡ *Department of Statistics*
Harvard University
Cambridge, USA
albertwu@g.harvard.edu ; airoldi@stat.harvard.edu

***Abstract*—In most real-world systems units are interconnected and can be represented as networks consisting of nodes and edges. For instance, in social systems individuals can have social ties, family or financial relationships. In settings where some units are exposed to a treatment and its effects spills over connected units, estimating both the direct effect of the treatment and spillover effects presents several challenges. First, assumptions on the way and the extent to which spillover effects occur along the observed network are required. Second, in observational studies, where the treatment assignment is not under the control of the investigator, confounding and homophily are potential threats to the identification and estimation of causal effects on networks. Here, we make two structural assumptions: i) neighborhood interference, which assumes interference to operate only through a function of the the immediate neighbors' treatments, ii) unconfoundedness of the individual and neighborhood treatment, which rules out the presence of unmeasured confounding variables, including those driving homophily. Under these assumptions we develop a new covariate-adjustment estimator for treatment and spillover effects in observational studies on networks. Estimation is based on a generalized propensity score that balances individual and neighborhood covariates across units under different levels of individual treatment and of exposure to neighbors' treatment. Adjustment for propensity score is performed using a penalized spline regression. Inference capitalizes on a three-step Bayesian procedure which allows taking into account the uncertainty in the propensity score estimation and avoiding model feedback. Finally, correlation of interacting units is taken into account using a community detection algorithm and incorporating random effects in the outcome model. All these sources of variability, including variability of treatment assignment, are accounted for in the posterior distribution of finite-sample causal estimands.**

***Index Terms*—Causal Inference, Interference, Spillovers, Bayesian Inference, Social Impact**

I. Introduction

A. Motivation and Background

In many areas of social, economic and medical sciences, researchers are interested in assessing not just the association but the causal relationship between two variables, i.e., exposure to a condition and an outcome variable that is expected to be affected by the exposure. Many studies conducted for assessing the effect of the exposure to a certain observed condition, as well as many non-experimental or experimental studies designed for evaluating the impact of public policies and programs, are actually aiming at inferring causal effects of the exposure to the observed condition or the implementation of the program. In both cases, causal conclusions are typically used for predicting causal consequences of an hypothetical intervention that manipulates the exposure or implements the public policy or program eventually by selecting or improving specific components.

Causal inference can be drawn using experimental or non-experimental studies. In the former, the main challenges lies in the design, which involves both the sampling design and the assignment of subjects to different experimental conditions. In the latter, the main challenge is covariate imbalance across individual in different conditions. Covariate adjustment methods are needed to compare similar units in different treatment arms (Hernán & Robins, 2018; Imbens & Rubin, 2015).

Most estimators of causal effects rely on the assumption of no interference between units, that is, a unit's outcome is assumed to depend only on the treatment it received. When, instead, interference is present, common estimators fail to estimate the causal effect of the treatment. Interference mechanism are common in many fields, from economics to epidemiology. For example, de-worming a group of children may also benefit untreated children by reducing disease transmission (Miguel & Kremer, 2004).

In the past two decades the literature on causal inference in the presence of interference has rapidly started to grow, with increasingly rapid advances in both areas of experimental design (e.g., Eckles et al. 2014; Hudgens & Halloran 2008; Toulis & Kao 2013) and statistical inference. The recently proposed approaches for the estimation of spillover effects can be categorized as dealing with one of the following cases: randomized experiments on clusters (e.g., Hudgens & Halloran 2008), randomized experiments on networks (e.g., Aronow 2012; Aronow & Samii 2017; Athey et al. 2015; Bowers

978-1-5386-5090-5/18/$31.00 ©2018 IEEE
DOI 10.1109/DSAA.2018.00016

et al. 2013; Rosenbaum 2007), observational studies on clusters (e.g.,Hong & Raudenbush 2006; Tchetgen Tchetgen & VanderWeele 2012, and observational studies on networks (Forastiere et al., 2018; Van der Laan, 2014).

B. Contributions of the paper

Building on recent work by Forastiere et al. (2018), we extend the proposed generalized propensity score estimator to more flexible functional forms of the outcome model and to incorporate neighborhood correlation. We develop a Bayesian estimation method for finite sample causal effects, which relies on a modular technique and the imputation approach to causal inference. As an alternative to the commonly used frequentist and Fisherian perspectives, this paper pioneers the use of Bayesian inference for the estimation of treatment and spillover effects in the presence of interference. The proposed Bayesian methodology allows flexible estimation of a large range of causal effects, incorporates different sources of uncertainty and allows taking into account correlation among neighbors using community random effects. In addition, the modular technique enables preserving robustness to model misspecification.

II. Potential Outcome Framework

A. Introduction

The most widely used statistical framework for causal inference is the potential outcome framework (Rubin, 1974, 1978), also known as the Rubin Causal Model (RCM) (Holland , 1986). The first component of the RCM are potential outcomes defined as potential values of the outcome variable that each unit would experience under each level of the treatment condition. Causal effects are then defined as comparisons of potential outcomes under different treatment conditions for the same set of units. The fundamental problem of causal inference under the RCM is that, in one study, for each unit at most one of the potential outcomes is observed – the one corresponding to the treatment to the unit is actually exposed to –, and the other potential outcomes are missing. Therefore, unit-level causal effects are not identifiable without further assumptions. The second component of the RCM is the assignment mechanism: the process that determines which units receive which treatments, hence which potential outcomes are realized and thus can be observed, and which are missing. In randomized experiment the assignment mechanism is under the control of the experimenter. In contrast, in observational studies the assignment mechanism is the unknown process underlying the observed distribution of treatment and in general depends on units' characteristics. Identification of causal effects relies on specific assumptions on the assignment mechanism. The last optional component of the RCM is a model for the potential outcomes and covariates. In what follows, we will consider potential outcomes as random variables.

B. Set up and Notation

Consider a study where we observe a set of N units. Let i be the indicator of a unit in the sample, with $i = 1, \ldots, N$.

The variable the causal effect of which is under investigation can be the exposure to a certain condition (e.g., environmental exposure, socio-economic status, behavior) or a certain treatment or intervention. Throughout we will refer to this variable as treatment. Let $Z_i \in \mathcal{Z}$ be the treatment variable indicator for unit i and $Y_i \in \mathcal{Y}$ the observed outcome that we wish to estimate the effect on. For each unit we also collect a vector of baseline covariate $\mathbf{X}_i \in \mathcal{X}$. Let $\mathbf{O} = (\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ denote the observed data in the sample, where $\mathbf{X} = \{\mathbf{X}_i\}_i^N$ is the collection of baseline covariate across all units, $\mathbf{Z} = \{Z_i\}_i^N$ is the treatment vector in the sample and $\mathbf{Y} = \{Y_i\}_i^N$ is the the corresponding outcome vector.

In principle, we should define the potential outcome of each unit as the potential value of the outcome variable that the unit would experience under a specific assignment of the whole treatment vector $\mathbf{Z} = \mathbf{z}$. Under this general definition, a potential outcome is denoted by $Y_i(\mathbf{Z} = \mathbf{z})$ or simply $Y_i(\mathbf{z})$. We then have for each unit $|\mathcal{Z}|^N$ potential outcomes but we can only observe the one corresponding to the treatment vector that was actually observed, i.e., $Y_i(\mathbf{Z})\ \forall i$. A dimensionality reduction is needed both for the definition and for the identification of causal estimands as comparisons between potential outcomes under different treatment conditions.

C. Causal Estimands and Identifying Assumptions

The first basic assumption that is typically invoked is the stable unit treatment value assumption (SUTVA; Rubin, 1980). There are two components to this assumption. The first is that there is only one version of each treatment level possible for each unit (consistency). The second is the no interference assumption, that is, the treatment of one unit does not affect the potential outcomes of other units, formally:

Assumption 1: Stable Unit Treatment Value (SUTVA)

$$\text{If} \quad Z_i = Z_i' \quad \text{then} \quad Y_i(\mathbf{Z}) = Y_i(\mathbf{Z}')\ \forall \mathbf{Z}, \mathbf{Z}', \forall i$$

Under this assumption we can index potential outcomes of unit i only by the treatment received by unit i, i.e., $Y_i(Z_i = z)$ or simply $Y_i(z)$. Therefore, under SUTVA, for each unit there exist only one potential outcome for each treatment level, with the observed outcome $Y_i = \sum_{z \in \mathcal{Z}} I(Z_i = z) Y_i(z)$.

Causal estimands are defined as comparisons of potential outcomes under different treatment levels. Unit-level estimands are comparisons at the unit level, while average estimands are average comparisons on the same sets units. A common comparison is the average difference. For the (finite) population of units, under SUTVA, the average treatment effect (ATE) is defined as:

$$ATE(z, z') = \frac{1}{N} \sum_{i=1}^{N} (Y_i(z) - Y_i(z'))$$

· The problem of identifying and estimating causal estimands relies on the fundamental problem of causal inference (Holland , 1986), that is, the inability to simultaneously observe all the potential outcomes of the same unit. In fact, for each unit, we can observe at most the potential outcome corresponding to the the treatment to which the unit is exposed, i.e., $Y_i = Y_i(Z_i)$,

whereas all the other potential outcomes $Y_i(z)$, with $Z_i \neq z$, are missing. We could estimate a marginal contrast in potential outcomes, e.g., $ATE(z, z')$, if we could recover the distribution of $Y_i(z)$ for units $Z_i \neq z$. This could be done by assuming that the distribution of potential outcomes is independent of the actual treatment received, i.e., $Y_i(z) \perp\!\!\!\perp Z_i \; \forall z, \forall i$. This assumption is known as *unconconfoudedness*.

The unconfoudedness assumption holds by design in randomized experiments, where the assignment mechanism is known and the treatment is randomly assigned. Conversely, in observational studies the investigator does not control the assignment of treatments and cannot ensure that similar subjects receive different levels of treatment. In an observational setting we typically relax the unconditional unconfoudedness assumption by assuming exchangeability conditional on a set of observed covariates. The conditional unconfoudedness assumption can be formally stated as follows.

Assumption 2: Conditional Unconconfoudedness

$$Y_i(z) \perp\!\!\!\perp Z_i | \mathbf{X}_i \; \forall z \in \mathcal{Z}, \forall i$$

Assumption 2 implies that the treatment is as randomized among units with the same value of the observed covariates.

D. Propensity Score for Binary Treatment

Most of causal inference literature is concerned with settings with a binary treatment $Z_i \in \{0, 1\}$ In this case, for each unit there are only two potential outcomes $Y_i(0)$ and $Y_i(1)$, one observed and one missing.

In such settings, propensity score-based methods are particularly useful. For each unit, the propensity score, denoted by $\phi(\mathbf{X}_i)$, is defined as the probability of receiving the active treatment given the unit's covariates: $\phi(\mathbf{X}_i) = Pr(Z_i = 1|\mathbf{X}_i)$. The propensity score $\phi(\mathbf{X}_i)$ has two important properties: (i) it is a balancing score, that is, it satisfies $\mathbf{X}_i \perp\!\!\!\perp Z_i|\phi(\mathbf{X}_i)$; (ii) if the treatment assignment is unconfounded given $\mathbf{X}_i$, then it is also unconfounded given $\phi(\mathbf{X}_i)$, that is, $Y_i(z) \perp\!\!\!\perp Z_i|\phi(\mathbf{X}_i)$.

Therefore, under unconfoundedness, adjusting for the difference in the propensity scores between treated and control units would remove all biases due to covariate imbalance, i.e., we can compare the observed outcomes between treated and control units within groups defined by the value of the propensity score rather than by the value of covariates.

Propensity score methods usually involve two stages: in the first stage ('PS stage'), the propensity scores $\phi(\mathbf{X}_i)$ are determined by estimating the parameters of a model, usually through a logistic regression ($\text{logit}(Pr(Z_i = 1|\mathbf{X}_i)) = \alpha + \beta_Z^T \mathbf{X}_i$), and then by computing the individual probabilities given covariates; in the second stage ('outcome stage'), estimate the causal effects based on the estimated propensity scores, through three main alternative strategies: matching, stratification, weighting or combination of these methods (for a review, see Imbens & Rubin 2015).

E. Propensity Score for Continuous Treatment

In many studies the treatment is not binary but units may receive different treatment levels. In such settings, adjusting for the difference in the propensity score corresponding to one specific treatment level $z^\star$ does not yield unbiased estimate of causal estimands comparing potential outcomes under different treatment levels, i.e., $ATE(z, z')$ with $z, z \neq z^\star$. In addition, when the treatment is discrete or continuous, we are usually interested in estimating an average dose-response function.

Over the last years, propensity score methods have been generalized to the case of discrete and continuous treatments. Here we review the work by Hirano & Imbens (2004), who introduced the concept of the Generalized Propensity Score (GPS) and use it to estimate an average dose-response function (ADRF) $\mu(z) = E[Y_i(z)]$. The GPS, denoted by $\lambda(z; \mathbf{x})$, is the conditional density of the treatment given the covariates: $\lambda(z; \mathbf{X}_i) = p(z|\mathbf{X}_i)$. Each unit is then characterized by a different density of the treatment. For each unit the GPS corresponding to the actual treatment to which the unit is exposed, $\Lambda_i = \lambda(Z_i; \mathbf{x})$, is the probability for that unit of receiving the treatment it actually received given his characteristics $\mathbf{X}_i$. As the classic propensity score, the GPS is a balancing score. In the continuous case, this means that, within strata with the same value of $\lambda(z; \mathbf{X}_i)$, the probability that $Z_i = z$ does not depend on the value of $\mathbf{X}_i$. Furthermore, the unconfoundedness assumption, combined with the balancing score property, implies that the treatment is unconfounded given the GPS, i.e., $Y_i(z) \perp\!\!\!\perp Z_i|\lambda(z; \mathbf{X}_i) \; \forall z \in \mathcal{Z}, \forall i$. Thus, any bias caused by covariate unbalance across groups with different treatment levels can be removed adjusting for the difference in the GPS.

Hirano & Imbens (2004) estimate the DRF using the estimated GPS by employing a parametric partial mean approach. Specifically, we posit a model for the treatment, $Z_i \sim f(\mathbf{X}_i; \boldsymbol{\theta}^{(Z)})$, and a model for the potential outcomes given the GPS: $Y_i(z) \sim f(z, \lambda(z; \mathbf{X}_i); \boldsymbol{\theta}^{(Y)})$. In Hirano & Imbens (2004) the linear predictor is a cubic polynomial function of the treatment z and the generalized propensity score $\lambda(z; \mathbf{X}_i)$, including the interaction term. Bia et al. (2014) replace the parametric approach with a semiparametric estimator based on penalized spline techniques. In particular, they use penalized bivariate splines, with radial basis functions. Relying on the two models for the treatment and the outcome, the average DRF is derived using a two-step estimator is used. In Algorithm 1 we describe the two-step estimator proposed by Hirano & Imbens (2004) for the estimation of the average DRF based on the generalized propensity score.

F. Bayesian Propensity Score Adjustment

After Rubin (1985) reflected on the usefulness of propensity scores for Bayesian inference, only recently has Bayesian estimation of causal effects been combined with propensity score methods (e.g., Kaplan & Chen 2012; McCandless et al. 2010, 2009; Zigler et al. 2013). The first advantage of the Bayesian propensity score approach is that it allows embedding propensity score adjustment within broader Bayesian modeling strategies, incorporating prior information as well as complex models for hierarchical data, measurement error or missing data (Rubin, 1985). In addition, Bayesian meth-

Algorithm 1: Generalized Propensity Score for Multivalued Treatment

Input: Dataset $\mathcal{D}^\star$, PS model, outcome model
Output: Average DRF $\mu(z)^\star, z \in \mathcal{Z}$

GPS Stage:

Estimate the parameters $\boldsymbol{\theta}^{(Z)}$ of the GPS model
for $i = 1$ *to* N **do**
  Predict $\widehat{\Lambda}_i = \lambda(Z_i; \mathbf{X}_i)$
end

Outcome Stage:

Estimate the parameters $\boldsymbol{\theta}^{(Y)}$ of the outcome model, given the data $(\mathbf{Z}^\star, \mathbf{Y}^\star)$ and $\widehat{\boldsymbol{\Lambda}}$
Impute potential outcomes and Compute average DRF:
for $z \in \mathcal{Z}$ **do**
  for $i = 1$ *to* N **do**
    Impute the potential outcome $\widehat{Y}_i(z)$:
      a. Predict the GPS $\phi(z; \mathbf{X}_i)$
      b. Predict $\widehat{Y}_i(z)$, given z, $\lambda(z; \mathbf{X}_i)$, and the estimated parameters $\widehat{\boldsymbol{\theta}^{(Y)}}$
  end
  Average the potential outcomes over all units:
  $\widehat{\mu}(z) = \frac{1}{N} \sum_{i=1}^{N} \widehat{Y}_i(z)$
end

ods offer a natural strategy for modeling uncertainty in the propensity scores. McCandless et al. (2009) proposed to model the joint distribution of the data and parameters with the propensity score as a latent variable. Let $\boldsymbol{\theta}^{(Z)}$ and $\boldsymbol{\theta}^{(Y)}$ be the vectors of parameters of the propensity score and the outcome model, respectively. Markov chain Monte Carlo (MCMC) methods allow to draw from the posterior distribution for model parameters,

$$p(\boldsymbol{\theta}^{(Z)}, \boldsymbol{\theta}^{(Y)} | \mathbf{Y}, \mathbf{Z}, \mathbf{X}) \propto p(\mathbf{Y}, \mathbf{Z}, \mathbf{X} | \boldsymbol{\theta}^{(Z)}, \boldsymbol{\theta}^{(Y)}) p(\boldsymbol{\theta}^{(Z)}) p(\boldsymbol{\theta}^{(Y)})$$

by successively drawing from the full conditional distributions

$$p(\boldsymbol{\theta}^{(Z)} | \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}^{(Y)}) \propto \prod_{i}^{N} p(Y_i, Z_i, \mathbf{X}_i | \boldsymbol{\theta}^{(Z)}, \boldsymbol{\theta}^{(Y)}) p(\boldsymbol{\theta}^{(Z)})$$

$$p(\boldsymbol{\theta}^{(Y)} | \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}^{(Z)}) \propto \prod_{i}^{N} p(Y_i, Z_i, \mathbf{X}_i | \boldsymbol{\theta}^{(Z)}, \boldsymbol{\theta}^{(Y)}) p(\boldsymbol{\theta}^{(Y)})$$

In McCandless et al. (2009), as well as in the whole literature on Bayesian propensity score, the focus is on binary treatment.

Within the Bayesian framework, we denote the propensity score of a binary treatment by $\phi(\mathbf{X}_i; \boldsymbol{\theta}^{(Z)})$ to highlight the dependence on the parameters. With propensity score adjustment, we do not directly model the dependence of the outcome of covariates, but rather we adjust for covariates by modeling parametrically or semi-parametrically the propensity score $\phi(\mathbf{X}_i; \boldsymbol{\theta}^{(Z)})$. As a consequence, the outcome model indirectly depends on all parameters, including the set of parameters of the propensity score models, i.e., $\boldsymbol{\theta}^{(Z)}$, and, thus, the likelihood cannot be factorized into two parts, $p(Y_i | Z_i, \mathbf{X}_i, \boldsymbol{\theta}^{(Y)}) p(Z_i | \mathbf{X}_i, \boldsymbol{\theta}^{(Z)})$, that separately depend on different sets of parameters. Therefore, the posterior distribution of the parameters $\boldsymbol{\theta}^{(Z)}$ of the PS stage are in part informed by the outcome stage. Because of this phenomenon, referred to as 'model feedback' (e.g., McCandless et al. 2010; Zigler et al. 2013), the joint Bayesian PS estimation has raised some concerns. First, since the propensity score adjustment is meant to approximate the design stage in a randomized experiment it should be done without access to the outcome data (Rubin, 2007). Furthermore, a practical consequence is the propagation of error due to model misspecification. In fact, when the model for the relationship between the outcome and the propensity score is misspecified, then the joint Bayesian approach was shown to provide invalid inferences for $\boldsymbol{\theta}^{(Z)}$, which distorts the balancing property of the propensity score and yields incorrect estimates of the treatment effect (Zigler et al., 2013).

Various methods described as 'two-step Bayesian' have been recently proposed to 'cut the feedback' between the propensity score and outcome stages (e.g., Kaplan & Chen 2012; McCandless et al. 2010). These methods represent a special case of the so-called 'modularization' in Bayesian inference (Liu et al., 2009). To limit feedback, the general idea is based on an approximate Bayesian technique that uses the posterior distribution of the PS model as an input when fitting the outcome model. Specifically, the posterior distribution of the parameters $\boldsymbol{\theta}^{(Z)}$ of the PS model is not updated from the full conditional, but rather from the approximate conditional distribution

$$p(\boldsymbol{\theta}^{(Z)} | \mathbf{Y}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}^{(Y)}) \propto \prod_{i}^{N} p(Z_i | \mathbf{X}_i, \boldsymbol{\theta}^{(Z)}) p(\boldsymbol{\theta}^{(Z)})$$

which ignores the likelihood contribution from the outcome. This restricts the flow of information between models during MCMC computation, and is similar in spirit to two-stage estimation (Lunn et al., 2009).

Algorithm 2: Bayesian Two-Stage Propensity Score for Binary Treatment

Input: Dataset $\mathcal{D}$, PS model, outcome model, priors
Output: Posterior distribution of causal estimand

for $m = 1$ *to* M **do**

PS Stage:

Draw the parameters $\boldsymbol{\theta}^{(\boldsymbol{Z})} \sim p(\boldsymbol{\beta}^{(\boldsymbol{Z})} | \mathbf{X}, \mathbf{Z})$
for $i = 1$ *to* N **do**
  Predict $\widehat{\Phi}_i = \phi(\mathbf{X}_i; \boldsymbol{\theta}^{(\boldsymbol{Z})})$
end

Outcome Stage:

Draw the parameters $\boldsymbol{\theta}^{(\boldsymbol{Y})} \sim p(\boldsymbol{\beta}^{(\boldsymbol{Y})} | \mathbf{X}, \mathbf{Z}, \mathbf{Y}, \widehat{\boldsymbol{\Phi}})$
for $i = 1$ *to* N **do**
  Impute potential outcomes $Y_i^{mis} = Y_i(1)(1 - Z_i) + Y_i(0) Z_i$ from the posterior predictive distribution $p(Y_i^{mis} | Z_i, \mathbf{X}_i, \widehat{\Phi}_i, \boldsymbol{\theta}^{(\boldsymbol{Y})})$
end
Compute the causal estimand
$ATE = \frac{1}{N} \sum_{i=1}^{N} (Y_i - Y_i^{mis})(2Z_i - 1)$

end

III. Causal Inference under Interference on Network Data

In this section we describe the problem of interference on network data in observational studies. The presence of interference on networks poses two major challenges: i) spillover effects of a unit's treatment on other units' outcomes, including

through a contagion mechanism, ii) homophily, that is, the tendency of units with similar characteristics of forming ties, which creates a dependence structure among interacting units in both pre-treatment characteristics and in the outcome.

In addition, in observational studies, where the treatment is not randomized, homophily can generate correlation among neighbors' treatment due to similar propensity of taking the treatment given similar covariates, as well as peer influence in the treatment uptake.

A. Network Data

Consider a network $\mathcal{N}$ of N units, indexed by i, with adjacency matrix $\mathbf{A}$, where element $A_{ij} > 0$ represents the presence of a tie between unit i and unit j. Ties are assumed to be fixed and known. Recall that for each unit we measure a vector of covariates $\mathbf{X}_i$, a treatment variable Z_i, and an outcome variable Y_i. Here we focus on binary treatments $Z_i \in \{0, 1\}$. The adjacency matrix A defines for each unit i the set of units that have a direct tie with i. We refer to this set as neighborhood of unit i, denoted by $\mathcal{N}_i = \{j : A_{ij} > 0\}$, and to the units belonging to this set as neighbors of unit i. In real-world applications, neighbors can be geographical neighbors, or friends, partners or collaborators. Let $N_i = \sum_{j \neq i} I(A_{ij} > 0)$ denote the number of neighbors of unit it, referred to as *degree* of unit i. The complement of $\mathcal{N}_i$ in $\mathcal{N}$, excluding i, is denoted by $\mathcal{N}_{-i} \backslash \mathcal{N}_i$.

For each unit i, the partition $(i, \mathcal{N}_i, \mathcal{N}_{-i} \backslash \mathcal{N}_i)$ defines the following partitions of the treatment and outcome vectors: $(Z_i, \mathbf{Z}_{\mathcal{N}_i}, \mathbf{Z}_{\mathcal{N}_{-i} \backslash \mathcal{N}_i})$ and $(Y_i, \mathbf{Y}_{\mathcal{N}_i}, \mathbf{Y}_{\mathcal{N}_{-i} \backslash \mathcal{N}_i})$. With non-network data, $\mathbf{X}_i$ typically includes individual characteristics or cluster-level characteristics representing contextual factors (e.g., demographic or socio-economic factors) or contextual covariates (e.g., geographical factors or presence of infrastructures). On the contrary, in network data $\mathbf{X}_i$ might also include variables describing the network. In particular, it might contain variables representing of the neighborhood $\mathcal{N}_i$, including the topology but also the distribution of individual-level characteristics, and it can contain network properties at node-level representing the position of unit's neighborhood in the graph (e.g., centrality, betweenness, ...).

B. Neighborhood interference

In general, the potential outcome for unit i depends on the entire treatment assignment vector $\mathbf{Z}$, i.e., $Y_i(\mathbf{Z} = \mathbf{z})$. The no interference assumption, or SUTVA, restricts the dependency to only the treatment received by unit i, i.e., $Y_i(Z_i = z)$. On the contrary, under interference the potential outcome for unit i depends on the treatment received by other units. However, if each outcome depends on the whole treatment vector, then for each treatment vector $\mathbf{Z}$ each unit would be observed under the same treatment $\mathbf{Z}$. Therefore, the data would not provide any information on missing potential outcomes under different treatment conditions. Oftentimes, we have reasons to assume that the outcome of a unit only depends on the treatment received by the neighbors, that is, by the units that individual is in direct contact with. In such case, the potential outcome could be written as $Y_i(\mathbf{Z}_{\mathcal{N}_i} = \mathbf{z}_{\mathcal{N}_i})$. This assumption excludes spillover effects of the treatment received by higher order connections. Nevertheless, when the number of neighbors is substantial then under each treatment vector $\mathbf{Z}$ the variability on the treatment vector $\mathbf{Z}_{\mathcal{N}_i}$ across units is still very low, and information on the $2^{N_i} - 1$ missing potential outcomes would be hard to extrapolate. For this reason, we introduce the concept of *exposure mapping*. In general terms, we define an exposure mapping as a function that maps a treatment vector $\mathbf{z}$, the adjacency matrix $\mathbf{A}$ and unit-level characteristics $\mathbf{X}$ to an exposure value denoted by G_i: $G_i = g(\mathbf{z}, \mathbf{A}, \mathbf{X}_i, \mathbf{X}_{-i})$, with $g : \mathcal{Z}^N \times \mathcal{A}^{N^2} \times \mathcal{X}^N \rightarrow \mathcal{G}_i$. Hudgens & Halloran (2008) consider the 'partial interference' assumption, that allows units to be affected only by the treatment received by units belonging to the same clusters. This can be expressed by a specific exposure mapping function that only depends on group indicators. In network data, a special case is the function $g_{\mathcal{N}}(\mathbf{z}_{\mathcal{N}_i}, \mathbf{A}_i, \mathbf{X}_i, \mathbf{X}_{\mathcal{N}_i})$, which receives as input only the treatment vector in the neighborhood, the unit's row of the adjacency matrix, the unit's covariates, and the covariates of units in the neighborhood. Given this definition, we can formalize the neighborhood interference assumption.

Assumption 3 (Neighborhood Interference): Given a function $g_{\mathcal{N}} : \mathcal{Z}^{N_i} \times \mathcal{A}^{N_i} \times \mathcal{X}^{N_i+1} \rightarrow \mathcal{G}_i$, $\forall i \in \mathcal{N}$, $\forall \mathbf{Z}_{\mathcal{N}_{-i}}, \mathbf{Z}'_{\mathcal{N}_{-i}}$ and $\forall \mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}^N : g_{\mathcal{N}}(\mathbf{Z}_{\mathcal{N}_i}, \mathbf{A}_i, \mathbf{X}_i, \mathbf{X}_{\mathcal{N}_i}) = g_{\mathcal{N}}(\mathbf{Z}'_{\mathcal{N}_i}, \mathbf{A}_i, \mathbf{X}_i, \mathbf{X}_{\mathcal{N}_i})$, then $Y_i(\mathbf{Z}) = Y_i(\mathbf{Z}')$.

Assumption 3 rules out the dependence of the potential outcomes of unit i from the treatment received by units outside its neighborhood, i.e., $\mathbf{Z}_{\mathcal{N}_{-i} \backslash \mathcal{N}_i}$, but allows Y_i to depend on the treatment received by his neighbors, i.e., $\mathbf{Z}_{\mathcal{N}_i}$. Moreover, this dependence is assumed to be through a specific exposure mapping function $g_{\mathcal{N}}(\cdot)$. This formulation is similar to the 'exposure mapping' introduced by Aronow & Samii (2017) and the one in Van der Laan (2014). In Assumption 3 the function $g_{\mathcal{N}}(\cdot)$ is assumed to be known and well-specified. We refer to $G_i = g_{\mathcal{N}}(\mathbf{Z}_{\mathcal{N}_i}, \mathbf{A}_i, \mathbf{X}_i, \mathbf{X}_{\mathcal{N}_i})$ as the *neighborhood treatment*, and, by contrast, we refer to Z_i as the *individual treatment*. In general, we can write $G_i = \sum_{j \in \mathcal{N}_i} A_{ij} w(\mathbf{X}_i, \mathbf{X}_j) Z_j$, where A_{ij} is the element of the adjacency matrix, which could be binary or weighted, and $w(\mathbf{X}_i, \mathbf{X}_j)$ is a function of unit-level characteristics of the interacting units. In the simplest case, G_i can be the number or the proportion of treated neighbors, i.e., $G_i = \sum_{j \in \mathcal{N}_i} Z_j$ or $G_i = \frac{\sum_{j \in \mathcal{N}_i} Z_j}{N_i}$, respectively. The domain of G_i depends on how the function $g_{\mathcal{N}}(\cdot)$ is defined. For example, if we consider the simple number of treated neighbors, then $\mathcal{G}_i = \{0, 1, \ldots, N_i\}$. We denote the overall domain by $\mathcal{G} = \bigcup_i \mathcal{G}_i$.

Under Assumption 3, potential outcomes can be indexed just by the the individual treatment and the neighborhood treatment, i.e., $Y_i(Z_i = z, G_i = g)$, which can be simplified to $Y_i(z, g)$. The potential outcome $Y_i(z, g)$ represents the outcome that unit i would exhibit under individual treatment $Z_i = z$ and if exposed to the value g of a function $g_{\mathcal{N}}(\cdot)$ of the treatment vector of his neighbors, $\mathbf{Z}_{\mathcal{N}_i}$.

A potential outcome $Y_i(z,g)$ is defined only for a subset of nodes where G_i can take on value g. We denote this subset by $V_g = \{i : g \in \mathcal{G}_i\}$. For instance, in the case where G_i is the number of treated neighbors, V_g is the set of nodes with degree $N_i \geq g$, that is, with at least g neighbors. It is worth noting that each unit can belong to different subsets V_g, depending on the cardinality of $\mathcal{G}_i$.

C. Causal estimands: Treatment and Spillover effects

We define here the causal estimands of interest under the neighborhood interference. We first define the average potential outcome $Y_i(z,g)$ in a set of units V as:

$$\mu(z,g;V) = \frac{1}{|V|}\sum_{i\in V} Y_i(z,g) \qquad z \in 0,1, g \in \mathcal{G} \tag{1}$$

V is a set of units, possibly defined by covariates, including individual or network characteristics.

We can now define causal estimands as comparisons between average potential outcomes. We define the average (individual) *treatment effect* at neighborhood level $g \in \mathcal{G}$ by

$$\tau(g;V) = \mu(1,g;V) - \mu(0,g;V) \tag{2}$$

which denotes the average causal effect of the individual treatment when the neighborhood treatment is set to level g. Again here $V \subseteq V_g$. Instead of fixing the neighborhood treatment, we can consider an hypothetical intervention that assigns the neighborhood treatment to unit i based on a probability distribution $\pi^\star(g;\mathbf{X}_{\mathcal{N}_i})$. Thus, we define the average *treatment effect* $\tau(\pi^\star,V)$ given the neighborhood treatment assignment $\pi^\star(g;\mathbf{X}_{\mathcal{N}_i})$ by the average effect of the individual treatment marginalized over the probability distribution of the neighborhood treatment, that is

$$\tau(\pi^\star,V) = \int \Big(\mu(1,g;V) - \mu(0,g;V)\Big)\pi^\star(g;\mathbf{X}_{\mathcal{N}_i})dg. \tag{3}$$

We now define the causal effects of the neighborhood treatment, often referred to as spillover effects or peer effects. We define the *average spillover effect* of having the neighborhood treatment set to level g versus g', when the unit is under the individual treatment z, by

$$\delta(g,g',z;V) = \mu(z,g;V) - \mu(z,g';V) \tag{4}$$

Notice that V must be a subset of units belonging to both V_g and $V_{g'}$, i.e., $V \subseteq V_g \cap V_{g'}$. Finally, define the *average spillover effect* of intervention $\pi^\star$ vs π' by

$$\begin{aligned}\Delta(\pi^\star,\pi',z;V) = &\int \mu(z,g;V)\pi^\star(g;\mathbf{X}_{\mathcal{N}_i})dg \\ &- \int \mu(z,g;V)\pi'(g;\mathbf{X}_{\mathcal{N}_i})dg\end{aligned} \tag{5}$$

Hypothetical Intervention $\pi^\star(g;\mathbf{X}_{\mathcal{N}_i})$ vs $\pi'(g;\mathbf{X}_{\mathcal{N}_i})$ can be given by real experiments with assignment mechanism $p(\mathbf{Z}=\mathbf{z})$, which reflects into the probability distribution of the neighborhood treatment, or it can be directly defined as a probability distribution of G_i given covariates.

D. Identifying Assumption: Unconfoundedness

Because the causal effects of interest depend on the comparison between two quantities $\mu(z,g;V)$ with different values of the individual and neighborhood treatments, identification results can focus on the identification of the ADRF $\mu(z,g;V)$.

Assumption 4 (Unconfoundedness of Individual and Neighborhood Treatment):

$$Y_i(z,g) \perp\!\!\!\perp Z_i, G_i \mid \mathbf{X}_i \qquad \forall z \in \{0,1\}, g \in \mathcal{G}_i, \forall i \in V.$$

This assumption states that the individual and neighborhood treatments are independent of the potential outcomes of unit i, conditional on the vector of covariates $\mathbf{X}_i$.

Assumption 4 states that the vector $\mathbf{X}_i$ contains all the potential confounders of the relationship between the individual and the neighborhood treatment and the potential outcomes for each unit i. The plausibility of this assumption depends on how the vector $\mathbf{X}_i$ is defined in relation to the probability distribution of the treatment and to the network structure. Assumption 4 rules out the presence of latent variables (not included in $\mathbf{X}_i$) that affect both the probability of taking the treatment and/or the value of neighborhood treatment and the outcome. Neighborhood covariates, that is, the topology of the neighborhood or individual-level covariates among neighbors, are potential confounders only if they affect the outcome of the unit.

Forastiere et al. (2018) show that under Assumption 4 the ADRF $\mu(z,g;V)$ is identified from the observed data, and estimation can be conducted by taking the average of the observed outcomes of units with $Z_i = z$ and $G_i = g$ within groups of units defined by covariates $\mathbf{X}_i$.

IV. Bayesian Generalized Propensity Score Estimator for Causal Effects under Neighborhood Interference

Here we develop a new Bayesian semi-parametric estimator for the ADRF $\mu(z,g;V)$, and in turn for the causal estimands in Section III-C. The idea is to combine results on the generalized propensity score for multivalued treatment proposed by Hirano & Imbens (2004) (Section II-E) and extended by Forastiere et al. (2018) to interference settings, with the two-step Bayesian propensity score estimator for a binary treatment without interference (see Section II-F). The proposed estimator lies within the Bayesian imputation approach to causal inference. In addition we will replace the parametric partial mean approach of Hirano & Imbens (2004) with a semiparametric technique based on penalized Bayesian multivariate splines. To take into account the dependence in the outcome, we also include random effects in the outcome model, with groups defined by a community detection algorithm. This Bayesian approach allows us to easily quantify the unceratinty due to the assignment of Z and G and to the inherent variability of the (missing) potential outcomes.

A. Individual and Neighborhood Propensity Scores

Under the unconfoudedness assumption (Assumption 4) the ADRF $\mu(z,g;V)$ could be estimated by taking the average

of the observed outcomes within cells defined by covariates. Nevertheless, the presence of continuous covariates or a large number of covariates poses some challenges in the estimation. Under SUTVA, propensity score-based estimators are common solutions (see Section II-D). Conversely, under the neighborhood interference assumption, Forastiere et al. (2018) propose a new propensity score-based estimator, based on the adjustment for the so-called individual and neighborhood propensity scores.

The *individual propensity score*, denoted by $\phi(z;\mathbf{X}_i^z)$, is the probability of having the individual treatment at level z conditional on covariates $\mathbf{X}_i^z$, i.e., $P(Z_i = z|\mathbf{X}_i^z = \mathbf{x}^z)$. Similarly, the *neighborhood propensity score*, denoted by $\lambda(g;z;\mathbf{X}_i^g)$, is the probability of having the neighborhood treatment at level g conditional on a specific value z of the individual treatment and on the vector of covariates $\mathbf{X}_i^g$, i.e., $P(G_i = g|Z_i = z, \mathbf{X}_i^g = \mathbf{x}^g)$. $\mathbf{X}_i^z \in \mathcal{X}^z \subset \mathcal{X}$ is the subset of covariates affecting the individual treatment, and $\mathbf{X}_i^g \in \mathcal{X}^g \subset \mathcal{X}$ is the subset of covariates affecting the neighborhood treatment. Typically, $\mathbf{X}_i^z$ should include individual characteristics and $\mathbf{X}_i^g$ is likely to include neighborhood characteristic. Nevertheless, $\mathbf{X}_i^z$ and $\mathbf{X}_i^g$ could also coincide and both include all kind of covariates.

Forastiere et al. (2018) show that the individual and neighborhood propensity scores satisfy the balancing and unconfoundedness properties. In particular, if Assumption 4 holds given $\mathbf{X}_i$, then the unconfoundedness assumption holds conditional on the two propensity scores separately, i.e., $Y_i(z,g) \perp\!\!\!\perp Z_i, G_i|\lambda(g;z;\mathbf{X}_i^g), \phi(z;\mathbf{X}_i^z), \forall z \in \{0,1\}, g \in \mathcal{G}_i$. This property allows deriving a covariate-adjustment method that separately adjusts for the individual propensity score $\phi(z;X_i^z)$ and for the neighborhood propensity score $\lambda(g;z;X_i^g)$. Because $\phi(z;\mathbf{X}_i^z)$ is the propensity score of a binary treatment, we can always adjusts for the propensity score $\phi(\mathbf{X}_i^z) = \phi(\ \mathbf{X}_i^z)$. Forastiere et al. (2018) propose the use of a subclassification approach on the individual propensity score $\phi(1;x^z)$ and, within subclasses that are approximately homogenous in $\phi(1;x^z)$, a model-based approach for the neighborhood propensity score, similar to the one in Hirano & Imbens (2004). Here we replace the frequentist subclassification and generalized propensity score-based estimator with a semiparametric Bayesian approach.

B. Propensity Scores and Outcome Models

Individual and Neighborhood Propensity Scores Models. We first posit a model for the binary individual treatment Z_i

$$\begin{aligned} Z_i &\sim Ber(\phi(1;\mathbf{X}_i^z)) \\ \mathrm{h}^{(Z)}\big(\phi(1;\mathbf{X}_i^z)\big) &= \boldsymbol{\beta}^{(Z)T}\mathbf{X}_i^z \end{aligned} \tag{6}$$

where $h^{(G)}(\cdot)$ is the logit or probit link function, and a model for the neighborhood treatment G_i

$$\mathrm{h}^{(G)}\big(G_i\big) \sim f^{(G)}\left(q^{(G)}\big(\mathbf{X}_i;\boldsymbol{\beta}^{(G)}\big), \nu^{(G)}\right) \tag{7}$$

where again $h^{(G)}(\cdot)$ is a link function, $f^{(G)}$ is a probability density function (pdf), $q^{(G)}(\cdot)$ is a flexible function of the covariates depending on a vector of parameters $\boldsymbol{\beta}^{(G)}$, and $\nu^{(G)}$ is a scale parameter.

Outcome Model.We now postulate a model for the potential outcomes given $\phi(1;\mathbf{X}_i^z)$ and $\lambda(g;z;\mathbf{X}_i^g)$:

$$\begin{aligned} &\mathrm{h}^{(Y)}\big(Y_i(z,g)\big) \sim \\ &\quad f^{(Y)}\left(q^{(Y)}\big(z,g,\phi(1;\mathbf{X}_i^z),\lambda(g;z;\mathbf{X}_i^g);\boldsymbol{\beta}^{(Y)}\big),\nu^{(Y)}\right) \end{aligned} \tag{8}$$

where as usual $h^{(Y)}(\cdot)$ is a link function, $f^{(Y)}$ is a probability density function (pdf), and $\nu^{(Y)}$ is a scale parameter. The key feature here is $q^{(Y)}(\cdot)$, which we model semiparametrically using a set of penalized spline basis functions. Splines yield several advantages that include flexibility as well as interpretability via representations that use a compact set of basis functions and coefficients. In particular, the predictor $q^{(Y)}(z,\mathbf{V}(z,g)_i)$, where $\mathbf{V}(z,g)_i = [g,\phi(1;\mathbf{X}_i^z),\lambda(g;z;\mathbf{X}_i^g)]^T$, can be written in the mixed model representation (Ruppert et al., 2003):

$$\begin{aligned} q^{(Y)}\big(z,\mathbf{V}_i(z,g)\big) =& \boldsymbol{\beta}_V^{(Y)T}\mathbf{V}_i'(z,g) + \boldsymbol{\beta}_{Vz}^{(Y)T}\mathbf{V}_i'(z,g)z \\ &+ \boldsymbol{b}_U^T\mathbf{U}_i(z,g) + \boldsymbol{b}_{Uz}^T\mathbf{U}_i(z,g)z + u_j \\ u_j \sim \mathcal{N}(0,\ \Sigma_u);\ \boldsymbol{b}_U &\sim \mathcal{N}(0,\sigma_{b_u}^2 I_K);\ \boldsymbol{b}_{Uz} \sim \mathcal{N}(0,\sigma_{b_{uz}}^2 I_K) \end{aligned} \tag{9}$$

where $\mathbf{V}_i'(z,g) = [1,g,\phi(1;\mathbf{X}_i^z),\lambda(g;z;\mathbf{X}_i^g),\lambda(g;z;\mathbf{X}_i^g)g]^T$, such that the first two terms of Equation (9) represent the linear predictor with interactions, and $\mathbf{U}_i(z,g)$ contains spline basis functions. In particular, we use multivariate smoothing splines with radial basis functions of the form

$$\begin{aligned} \mathbf{U}_{ik}(z,g) =& C(||\mathbf{V}_i(z,g) - \mathbf{k}_k||)\Omega^{-1/2}; \\ \Omega =& \big[C(||\mathbf{k}_k - \mathbf{k}_{k'}||)\big]_{1\le k,k',\le K} \end{aligned} \tag{10}$$

where $||\cdot||$ is the euclidean norm and $C(\cdot)$ is a basis function. Here our choice goes to thin plate splines of the form

$$C(||\mathbf{r}||) = \begin{cases} ||\mathbf{r}||^{2m-|\mathbf{r}|} \\ ||\mathbf{r}||^{2m-|\mathbf{r}|} log(\mathbf{r}) \end{cases} \tag{11}$$

where m is an integer satisfying $2m-|\mathbf{r}| > 0$, that controls the order of the spline and its smoothness (Ruppert et al., 2003; Wood, 2003). The default is to use the smallest integer satisfying that condition. The advantage of radial basis functions in multivariate smoothing is that they are rotational invariant. The distribution of the coefficients $\boldsymbol{b}_U$ and $\boldsymbol{b}_{Uz}$ is a mixed model representation of penalties. The variances $\sigma_{b_u}^2$ and $\sigma_{b_{uz}}^2$ are indeed the parameters controlling the degree of smoothness.

In Equation 8, the outcome model depends on the individual treatment, the neighborhood treatment and both the individual and the neighborhood propensity scores. An alternative approach is to replace the model-based adjustment for the individual propensity score with a matching approach. The idea is to match units on the individual propensity score $\phi(1;\mathbf{X}_i^z)$ to create a matched sample where covariates $\mathbf{X}_i^z$ are balanced across the treated group ($Z_i = 1$) and the control group ($Z_i = 0$). Adjustment for the neighborhood propensity score is then handled, as previously, by a model-based generalized propensity score method applied to the matched samples. For matching with replacement or variable ratio matching weights

need to be incorporated into the analysis. When matching with replacement, individuals receive a frequency weight that reflects the number of matched sets they belong to. When using variable ratio matching, units receive a weight that is proportional to the actual number of units matched to them. In a Bayesian framework, weights can be incorporated by weighting the scale parameter. Therefore, we would assume a model for the outcome as in (8), with $V_i = [g, \lambda(g; z; \mathbf{X}_i^g)]$, $\mathbf{V}_i' = [1, g, \lambda(g; z; \mathbf{X}_i^g), \lambda(g; z; \mathbf{X}_i^g)g]^T$ and $\nu^{(Y)}$ scaled by the matching weights.

Finally, the last term in Equation (9), u_j, $j = 1, \ldots, J$, is the random effect for community j, with $j = 1, \ldots, J$. We include this term to take into account any dependence in the outcome data between a unit and his neighbors. Clusters are defined using a community detection algorithm that identifies groups of nodes that are heavily connected among themselves, but sparsely connected to the rest of the network.

Community Detection. Unfolding the communities in real networks is widely used to determine the structural properties of these networks. Community detection or clustering algorithms aim at finding groups of related nodes that are densely interconnected and have fewer connections with the rest of the network. As communities are often associated with important structural characteristics of a complex system, community detection is a common first step in the understanding and analysis of networks. The problem of how to find communities in networks has been extensively studied and a substantial amount of work has been done on developing clustering algorithms (an overview can be found in Fortunato (2010)).

Priors. Within the Bayesian framework we should posit a prior distribution for the parameter vector

$$\boldsymbol{\theta} = \left[\boldsymbol{\theta^{(Z)}}, \boldsymbol{\theta^{(G)}}, \boldsymbol{\theta^{(Y)}}\right]$$

where $\boldsymbol{\theta^{(Z)}} = \boldsymbol{\beta^{(Z)}}$, $\boldsymbol{\theta^{(G)}} = [\boldsymbol{\beta^{(G)}}, \nu^{(G)}]$, $\boldsymbol{\theta^{(Y)}} = [\boldsymbol{\beta^{(Y)}}, \nu^{(Y)}, \Sigma_u, \sigma^2_{b_u}, \sigma^2_{b_{uz}}]$. In particular, we recommend the use of weakly informative priors to provide moderate regularization and help stabilize computation. We assume a multivariate normal prior for all regression coefficients:

$$\boldsymbol{\beta^{(Z)}} \sim \mathcal{N}(\boldsymbol{\eta}^Z, K^Z); \boldsymbol{\beta^{(G)}} \sim \mathcal{N}(\boldsymbol{\eta}^G, K^G); \boldsymbol{\beta^{(Y)}} \sim \mathcal{N}(\boldsymbol{\eta}^Y, K^Y)$$

The priors on the scale parameters $\nu^{(G)}$ and $\nu^{(Y)}$ should depend on the neighborhood treatment distribution $f^{(G)}$ and on the outcome distribution $f^{(Y)}$, respectively. However, a general prior distribution could be

$$\nu^{(G)} \sim \text{Exp}(\gamma_{\nu^G}) \qquad \nu^{(Y)} \sim \text{Exp}(\gamma_{\nu^Y})$$

The random effect covariance matrix Ω_u is decomposed into a diagonal matrix of standard deviations and the correlation matrix (McElreath , 2016): $\Sigma_u = \text{diag}(\boldsymbol{\sigma}_u)\Omega_u\text{diag}(\boldsymbol{\sigma}_u)$. For the correlation matrix we use a prior distribution called LKJ, whose density is proportional to the determinant of the correlation matrix raised to the power of a positive regularization parameter minus one:$\Omega_u \sim LKJ(\zeta); p(\Omega_u) \propto \det(\Omega_u)^{\zeta-1}, \zeta > 0$. The standard deviations $\boldsymbol{\sigma}_u$ are in turn decomposed into the product of a simplex vector $\boldsymbol{\pi}_u$) and the trace of the covariance matrix, i.e., $\text{diag}(\boldsymbol{\sigma}_u) = \text{tr}(\Sigma_u)\boldsymbol{\pi}_u = Je^2\boldsymbol{\pi}_u$, where the trace (total variance) is the product of the order of the matrix and the square of a scale parameter and the element π_j of the simplex vector is the proportion of the total variance attributable to the corresponding random effect u_j. For the scale parameter e we posit a Gamma prior, with shape and scale parameters both set to 1. For the simplex vector $\boldsymbol{\pi}_u$ we use a symmetric Dirichlet prior, which has a single concentration parameter $\chi > 0$.

For the smoothing parameters we posit the following prior distribution: $\sigma^2_{b_u} \sim \text{Exp}(\gamma_{b_u}); \sigma^2_{b_{uz}} \sim \text{Exp}(\gamma_{b_{uz}})$.

C. Three-Step Estimating Procedure

Here we propose a three-step Bayesian estimator that extends the 'two-step Bayesian' estimator proposed by McCandless et al. (2010) and Kaplan & Chen (2012) to the neighborhood interference setting, with the individual and the neighborhood propensity score.

The three steps refer to the posterior distributions of the parameters of the individual propensity score, the neighborhood propensity score and the outcome models. Since the outcome model involves the individual and neighborhood propensity scores, in principle the outcome model indirectly depends on all parameters, including the set of parameters of the two propensity score models, i.e., $\boldsymbol{\theta}^{(Z)}$, and $\boldsymbol{\theta}^{(G)}$. Therefore, the posterior distribution of these parameters should in part be informed by the outcome stage. To avoid 'model feedback', we take a three-step approach which approximate the joint posterior distribution by drawing the parameters of the propensity score models from the approximate conditional distributions

$$p(\boldsymbol{\theta}^{(Z)}|\mathbf{Y}, \mathbf{G}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}^{(Y)}, \boldsymbol{\theta}^{(G)}) \propto \prod_i^N p(Z_i|\mathbf{X}_i, \boldsymbol{\theta}^{(Z)})p(\boldsymbol{\theta}^{(Z)})$$

$$p(\boldsymbol{\theta}^{(G)}|\mathbf{Y}, \mathbf{G}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}^{(Y)}, \boldsymbol{\theta}^{(Z)}) \propto \prod_i^N p(G_i|Z_i, \mathbf{X}_i, \boldsymbol{\theta}^{(G)})p(\boldsymbol{\theta}^{(G)})$$

which ignore the likelihood contribution from the outcome and, hence, do not depend neither on $\boldsymbol{\theta}^{(Y)}$) nor on the parameters of the neighborhood or individual propensity score model, respectively. The posterior distributions of the individual and the neighborhood propensity score models are then used as an input when deriving the posterior distribution of the parameters of the outcome model:

$$p(\boldsymbol{\theta}^{(Y)}|\mathbf{Y}, \mathbf{G}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}^{(Z)}, \boldsymbol{\theta}^{(G)}) \propto \prod_i^N p(Y_i, G_i, Z_i, \mathbf{X}_i|\boldsymbol{\theta}^{(Z)}, \boldsymbol{\theta}^{(G)}, \boldsymbol{\theta}^{(Y)})p(\boldsymbol{\theta}^{(Y)})$$

After the posterior distribution of all the parameters $\boldsymbol{\theta}$ is drawn, the posterior distribution of our finite-sample average dose-response function ADRF $\mu(z, g; V)$ is obtained by drawing from the posterior predictive distribution the potential outcomes $Y_i(z, g)$, for each value of z and g and for each unit $i \in V$. Then for each draw the ADRF $\mu(z, g; V)$ is computed by taking the average of the imputed potential outcomes $Y_i(z, g)$ over all units of the set V. Causal estimands are simply computed from the ADRF as comparisons of average potential

Algorithm 3: Bayesian Three-Step Generalized Propensity Score

Input: Dataset $\mathcal{D}$, Adjacency Matrix $\mathbf{A}$, Z model, G model, Y model, priors, Matching=FALSE
Output: Posterior distribution of ADRF $\mu(z,g)$

Community Detection Stage:
Run a *Community Detection algorithm* on $\mathbf{A} \Longrightarrow$ community indicators $C_i \in \{1,\dots,J\}, \forall i \in \mathcal{N}$
Initialize parameters $\boldsymbol{\theta}^{(0)} = [\boldsymbol{\theta}^{(Z)(0)}, \boldsymbol{\theta}^{(G)(0)}, \boldsymbol{\theta}^{(Z)(0)}]$
for $m = 1$ *to* M **do**
  PS Stage:
    Define $\mathbf{X}_i^z \in \mathbf{X}_i$ Draw the parameters $\boldsymbol{\beta}^{(\boldsymbol{Z})(\boldsymbol{m})} \sim p(\boldsymbol{\beta}^{(\boldsymbol{Z})}|\mathbf{X}^z, \mathbf{Z})$
    for $i = 1$ *to* N **do**
      Predict $\widehat{\Phi}_i^{(m)} = \phi(1; \mathbf{X}_i^z; \boldsymbol{\beta}^{(\boldsymbol{Z})(\boldsymbol{m})})$
    end
    if *Matching=TRUE* **then**
      Run a *Matching algorithm* with distance metric=$\widehat{\lambda}_i \Longrightarrow$ matched sets $\mathcal{S}_k$, $k = 1,\dots,K$
      Given $S_{ik} = I(i \in \mathcal{S}_k)$, define weights as $w_i = \frac{1}{\sum_{k=1}^K S_{ik}}$, set $\mathcal{M} = \{i : \sum_{k=1}^K S_{ik} > 0\}$, and define the matched sample as $\mathcal{D}^\star = \{(\mathbf{X}_i)_{i\in\mathcal{M}}, (Z_i)_{i\in\mathcal{M}}, (Y_i)_{i\in\mathcal{M}}\}$
    end
    else
      Set $w_i = 1, \forall i \in \mathcal{N}$; $\mathcal{M} = \mathcal{N}$; $\mathcal{D}^\star = \mathcal{D}$
    end
  GPS Stage:
    Compute $G_i = g_{\mathcal{N}}(\mathbf{Z}_{\mathcal{N}_i}, \mathbf{A}_i, \mathbf{X}_i, \mathbf{X}_{\mathcal{N}_i}), \forall i \in \mathcal{N}$
    Define $\mathbf{X}_i^g \in \mathbf{X}_i$ Draw the parameters $\boldsymbol{\beta}^{(\boldsymbol{G})(\boldsymbol{m})} \sim p(\boldsymbol{\beta}^{(\boldsymbol{G})}|\mathbf{X}^g, \mathbf{Z}, \mathbf{G})$ and $\nu^{(G)(m)} \sim p(\nu^{(G)}|\mathbf{X}^g, \mathbf{Z}, \mathbf{G})$
    for $i = 1$ *to* N **do**
      Predict $\widehat{\Lambda}_i^{(m)} = \lambda(G_i; Z_i; \mathbf{X}_i^g; \boldsymbol{\beta}^{(\boldsymbol{G})(\boldsymbol{m})}, \nu^{(G)(m)})$
    end
  Outcome Stage:
    if *Matching=TRUE* **then**
      Define $\mathbf{V}^{(m)}(Z_i, G_i)_i = [G_i, \widehat{\Lambda}_i^{(m)}]^T$
    end
    else
      Define $\mathbf{V}_i^{(m)}(Z_i, G_i) = [G_i, \widehat{\Phi}_i^{(m)}, \widehat{\Lambda}_i^{(m)}]^T$
    end
    Compute spline basis functions $\mathbf{U}_i^{(m)}(Z_i, G_i)$ as in (10) and (11)
    Draw the parameters $\boldsymbol{\theta}^{(Y)(m)}$ of the outcome model in (8) and (9) using the Gibbs sampler algorithm 4
    for $z = 0, 1$ **do**
      for $g \in \mathcal{G}$ **do**
        for $i = 1$ *to* N **do**
          Impute potential outcomes $\widehat{Y}_i(z,g)$:
            a. Predict the neighborhood GPS $\lambda(g; z; \mathbf{X}_i^g; \boldsymbol{\beta}^{(\boldsymbol{G})(\boldsymbol{m})}, \nu^{(G)(m)})$
            b. Define $\mathbf{V}_i^{(m)}(z,g)$ and compute $\mathbf{U}_i^{(m)}(z,g)$
            c. Predict $\widehat{Y}_i(z,g)$, given z,g, $\mathbf{V}_i^{(m)}(z,g)$, $\mathbf{U}_i^{(m)}(z,g)$, the random effects $\mathbf{u}^{(m)}$, and the parameters $\boldsymbol{\beta}^{(\boldsymbol{Y})(\boldsymbol{m})}$, $\boldsymbol{b}_U^{(m)}$, $\boldsymbol{b}_{Uz}^{(m)}$ and $\nu^{(Y)(m)}$
        end
        Average the potential outcomes over all units: $\widehat{\mu}(z,g) = \frac{1}{N}\sum_{i=1}^N \widehat{Y}_i(z,g)$
      end
    end
end

outcomes at different levels. In Algorithm 3 we describe all the steps of the algorithm.

Algorithm 4: Gibbs Sampler for Parameters of the Outcome Model

Input: Dataset $\mathcal{D}$, Adjacency Matrix $\mathbf{A}$, Y model, priors
Output: Posterior distribution of $\boldsymbol{\theta}^{(Y)(m)}$

a. Draw the random effects $u_j^{(m)} \sim p(u_j|\Sigma_u^{m-1})$
b. Draw $\boldsymbol{b}_U^{(m)} \sim p(\boldsymbol{b}_U|\sigma_{b_u}^{2(m-1)})$ and $\boldsymbol{b}_{Uz}^{(m)} \sim p(\boldsymbol{b}_{Uz}|\sigma_{b_{uz}}^{2(m-1)})$
c. Draw $\Sigma_u^{(m)} \sim p(\Sigma_u|\mathbf{Y}, \mathbf{Z}, \mathbf{V}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{U}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{u}^{(m)}, \boldsymbol{\beta}^{(\boldsymbol{Y})(\boldsymbol{m-1})}, \boldsymbol{\nu}^{(\boldsymbol{Y})(\boldsymbol{m-1})}, \boldsymbol{b}_U^{(m-1)}, \boldsymbol{b}_{Uz}^{(m-1)})$
d. Draw the smoothing parameters:
$\sigma_{b_u}^{2(m)} \sim p(\sigma_{b_u}^{2(m)}|\mathbf{Y}, \mathbf{Z}, \mathbf{V}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{U}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{u}^{(m)}, \boldsymbol{b}_U^{(m)}, \boldsymbol{b}_{Uz}^{(m)}, \boldsymbol{\beta}^{(\boldsymbol{Y})(\boldsymbol{m-1})}, \boldsymbol{\nu}^{(\boldsymbol{Y})(\boldsymbol{m-1})})$

$\sigma_{b_{uz}}^{2(m)} \sim p(\sigma_{b_{uz}}^{2(m)}|\mathbf{Y}, \mathbf{Z}, \mathbf{V}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{U}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{u}^{(m)}, \boldsymbol{b}_U^{(m)}, \boldsymbol{b}_{Uz}^{(m)}, \boldsymbol{\beta}^{(\boldsymbol{Y})(\boldsymbol{m-1})}, \boldsymbol{\nu}^{(\boldsymbol{Y})(\boldsymbol{m-1})})$

e. Draw $\boldsymbol{\beta}^{(\boldsymbol{Y})(\boldsymbol{m})} \sim p(\boldsymbol{\beta}^{(\boldsymbol{Y})}|\mathbf{Y}, \mathbf{Z}, \mathbf{V}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{U}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{u}^{(m)}, \boldsymbol{b}_U^{(m)}, \boldsymbol{b}_{Uz}^{(m)}, \boldsymbol{\nu}^{(\boldsymbol{Y})(\boldsymbol{m-1})})$
f. Draw $\nu^{(Y)(m)} \sim p(\nu^{(Y)}|\mathbf{Y}, \mathbf{Z}, \mathbf{V}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{U}^{(m)}(\mathbf{Z}, \mathbf{G}), \mathbf{u}^{(m)}, \boldsymbol{b}_U^{(m)}, \boldsymbol{b}_{Uz}^{(m)}, \boldsymbol{\beta}^{(\boldsymbol{Y})(\boldsymbol{m})})$

V. Conlcluding Remarks

We have discussed identification and estimation of causal effects of interventions in contexts with interconnected and interfering units. We have introduced the neighborhood interference assumption, on the way and the extent to which spillover effects occur along the observed network are required. Under unconfoundedness of the individual and neighborhood treatments, which rules out the presence of unmeasured confounding variables – including those driving homophily –, we have developed a new covariate-adjustment estimator for treatment and spillover effects in observational studies on networks. Estimation is based on a generalized propensity score and adjustment is performed using a penalized spline regression. We proposed a three-step Bayesian procedure which allows taking into account the uncertainty in the propensity score estimation and avoiding model feedback. Correlation of interacting units is taken into account using random effects on clusters defined by a community detection algorithm. We conducted preliminary simulations showing promising performance of the proposed methods in terms of bias and coverage.

Aronow, P. M. (2012). A General Method for Detecting Interference Between Units in Randomized Experiments. *Sociological Methods & Research*, 41(3), 3–16.

Aronow, P. M. & Samii, C.(2017). Estimating Average Causal Effects Under General Interference. *Forthcoming. Annals of Applied Statistics. Preprint:arxiv:1305.6156*.

Athey, S., Eckles, D., & Imbens, G.W. (2015). Exact p-values for network interference. *NBER Working Paper 21313*.

Bia, M., Flores, C.A., Flores-Lagunes, A., & Mattei, A. (2014). A Stata package for the application of semiparametric estimators of doseresponse functions. *Stata Journal, StataCorp LP*, 14(3), 580–604.

BOWERS, J., FREDRICKSON, M.M., & PANAGOPOULOS, C. (2013). Reasoning about interference between units: a general framework. *Political Analysis*, 21(1), 97–124.

ECKLES, D., KARRER, B., & UGANDER, J. (2014). Design and analysis of experiments in networks: Reducing bias from interference. *arxiv:1404.7530*.

FORASTIERE, L., AIROLDI, E.M., & MEALLI, F. (2018). Identification and estimation of treatment and interference effects in observational studies on networks. *arxiv:1609.06245*.

FORTUNATO, S. (2010). Community detection in graphs. *Physics Reports*, 486, 75–174.

GELMAN, A., MENG, X.,& STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–807.

HERNÁN, M. A., & ROBINS, J. M. (2018). Causal Inference. *Boca Raton: Chapman & Hall/CRC, forthcoming*.

HILL, J. & REITER, J.(2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, 25, 2230–2256.

HIRANO, K., & IMBENS, G. W. (2004). The propensity score with continuous treatments. *In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, ed. A. Gelman and X.-L. Meng, 73–84. West Sussex, England: Wiley InterScience*.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81, 945–970.

HONG, G. & RAUDENBUSH, S. W. (2006). Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data. *Journal of the American Statistical Association*, 101, 901–910.

HUDGENS, M. G. & HALLORAN, M. E. (2008). Towards causal inference with interference. *Journal of the American Statistical Association*, 103, 832–842.

IMBENS, G. W., & RUBIN, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. *Cambridge University Press New York, NY, USA*.

KAPLAN, D. & CHEN, J. (2012). A two-step bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, 77, 581–609.

LIU, F., BAYARRI, M.J., & BERGER, J.O. (2009). Modularization in bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4, 119–150.

LUNN, D., BEST, N., SPIEGELHALTERS, D., GRAHAM, G., & NEUENSCHWANDER, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36, 19–38.

MCCANDLESS, L.C., DOUGLAS, I.J., EVANS, S.J. & SMEETH, L. (2010). Cutting feedback in bayesian regression adjustment for the propensity score. *The international journal of biostatistics*, 6, 122.

MCCANDLESS, L.C., GUSTAFSON, P., & AUSTIN, P.C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28, 94–112.

MCCANDLESS, L.C., RICHARDSON, S., & BEST, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association*, 107, 40–51.

MCELREATH, R. (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan. *Chapman & Hall/CRC Press*.

MIGUEL, E. & KREMER, M. (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 72(1), 159–217.

ROSENBAUM, P. R. (2002). Observational studies. *New York, NY: Springer-Verlag*.

ROSENBAUM, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, 102(477), 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.

RUBIN, B. D. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology* 66, 688–701.

RUBIN, B. D. (1978). Bayesian inference for causal effects. *Annals of Statistics*, 6, 34–58.

RUBIN, D. B. (1980). Comment on "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by D. Basu. *Journal of the American Statistical Association*, 75, 591–593.

RUBIN, D. B. (1985). The use of propensity scores in applied Bayesian inference. *In Bayesian Statistics 2, Bernardo JM, De Groot MH, Lindley DV, Smith AFM (eds). Valencia University Press, North-Holland: Amsterdam*, 63–72.

RUBIN, D. B. (1986). Which Ifs have Causal Answers? Comment on "Statistics and Causal Inference" by P. Holland. *Journal of the American Statistical Association*, 81, 961–962.

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–36.

RUPPERT, D., WAND, M.P. & CARROL, R.J. (2003). Semiparametric Regression. *Cambridge University Press*.

TCHETGEN TCHETGEN, E. J. & VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21, 55–75.

TOULIS, P. & KAO, E.K. (2013). Estimation of Causal Peer Influence Effects. *ICML*, (3) 2013: 1489–1497.

VAN DER LAAN, M.J. (2014). Causal Inference for a Population of Causally Connected Units. *Journal of Causal Inference*, 0, 2193–3677.

ZIGLER, C.M., WATT, K., YEH, R. W., , WANG, Y., COULL, B. A., &, DOMINICI, F. (2013). Model feedback in {b}ayesian propensity score estimation. *Biometrics*, 69, 263–273.

WOOD, S.N. (2003). Thin plate regression splines.. *Journal of the Royal Statistical Society: Series B - Statistical Methodology*, 65, 95.