Observational Causal Inference Using Network Information

Yan Leng, Martin Saveski, Alex 'Sandy' Pentland, Dean Eckles
Massachusetts Institute of Technology
{yleng,msaveski,pentland,eckles}@mit.edu

Abstract

In the absence of an experiment or quasi-experiment, causal inference typically rests on the assumption that conditioning on (i.e., controlling or adjusting for) observed covariates is sufficient to eliminate confounding. This assumption is typically controversial, if not implausible. Even when employing rich, high-dimensional covariates, which may make this assumption more plausible, we may suspect that there is some unobserved confounding. Even worse, controlling for covariates that are instrumental variables may instead increase bias. Here we use a common form of high-dimensional data about individuals — a social network — for counterfactual prediction, which may (i) encode information about the nodes' latent characteristics, and (ii) nodes may selecting into treatment according to their network positions. We critically examine how to use network information to improve causal inference, especially when concerned about the presence of remaining unobserved confounding. We propose a framework to address the bias amplification due to controlling for instruments or community-level fixed effects. In particular, we propose a representation learning approach with group lasso regularization, which results in learning representations that are highly associated with both the treatment and the outcome. Simulations with empirical social network data demonstrate the effectiveness of our approach and illustrate the potential of using network information for observational inference in general.

1 Introduction

Randomized experiments play a central role in science and practice, especially in the Internet industry where rapid experimentation (e.g., A/B testing) is possible. But we often lack the ability to randomly assign some treatments of the greatest interest, requiring reliance on observational (i.e., non-experimental) data to make causal inferences and subsequent decisions. Here we examine opportunities for observational causal inference to make use of rich, high-dimensional data about individuals, particularly that encoded in a network.

Credible causal inference from observational data has long been desired, if illusive, in many fields including public health [1], economics [2, 3], marketing [4], and political science. Observational causal inference also plays an explicit role in many recent analyses of traces of online behavior, sometimes leading to development or evaluation of methods [5, 6, 7, 8, 9, 10]. In the absence of a known source of random variation in treatments, analysis is performing by matching on or modeling using socio-demographic or other covariates relevant to the behavior of interest. There has been a rich literature focusing on estimating treatment effects with increasingly rich data and sophisticated techniques, e.g., propensity score matching [11], nearest neighbor matching [12], and tree-based methods [13]. Recently, with the gaining popularity of deep neural networks, several representation learning approaches to causal inference have been proposed. Johansson et al. [14] borrow ideas from domain adaptation to propose a representation learning approach to observational causal inference. Their framework learns representations that are (i) predictive of the outcomes,

NeurIPS 2019, Workshop on Graph Representation Learning.

but also (ii) balanced across the treatment and control units. In a following-up paper, Shalit et al. [15] provide a generalization bound for estimating individual-level causal effect under the strong ignorability assumption. Yao et al. [16] utilize local similarity information together with the global balance measure into the treatment effect estimation to decrease the generalization error.

These studies assume that all relevant confounding variables are observed [14, 12, 15, 17], or at least that we have suitable, if noisy, measures of these confounders. However, it is unrealistic to measure all possible confounders in real-world observational studies. When this assumption is violated, controlling for observed covariates may instead amplify bias [18, 19, 20]. There are two mechanisms by which bias amplification happens. The first mechanism is "Z-bias", referring to the controlling of instrumental variables. Many studies have demonstrated that controlling for instrumental variables (i.e., variables related to treatment but not directly influencing the outcome) can amplify bias [18, 19, 20], as after adjusting for an instrument, remaining variation in the treatment is driven more by the unobserved confounder [21]. Second, [18] show that the common practice of adjusting for many categorical variables with fixed effects may also induce bias [18]. This can be problematic, especially in network studies with community-level fixed effects, i.e., when the effect of the community does not co-vary in the treatment and the outcome.

Fortunately, in many cases, besides covariates about the units (e.g., demographic information), we also know how the units are related to each other. For instance, in the analysis of social media data, we often observe users' characteristics and behaviors, but we also observe their social network, i.e., the links between them. This additional high-dimensional information can be useful in answering causal questions in two ways. First, due to homophily [22], the social network can encode information about latent attributes that confound the treatment and outcome's causal relationship. Second, individuals may self-select into treatment according to their position in the network. For instance, students living in the same dorm are more likely to be exposed to the same education program, and consequently, are more likely to participate in the program. With these motivations, the objective of this paper is to study how network information can be used for counterfactual inference, and propose a method to utilize such information effectively.

Though promising, there is a lack of effective method in utilizing network information for causal identification. The high-dimensional network—in the form of adjacency matrix or graph Laplacian—makes it difficult to apply traditional matching methods directly. More specifically, the bias of the matching estimator grows with the dimensionality of the covariates [23]. Auerbach [24] and Shalit et al. [15] utilize network information by assuming that network contains sufficient information for causal identification. [15] propose a two-step approach by inferring the latent network positions and then run a regression using these covariates as controls. [24] match individuals who have similar network positions.

Here we study how network information can be used to improve counterfactual inference. We propose a deep learning model designed to reduce bias amplification due to controlling for instrumental variables. In particular, we introduce a representation learning approach with group lasso regularization, which enforces the learned representation to be highly associated with both the treatment and the outcome. In other words, if the representations associate strongly with the treatment but weakly with the outcome, they will be omitted. To achieve this, we use group lasso regularization to select the representations learned from the network and jointly regularize the coefficients of the representations on the treatment and the outcome [25].

To demonstrate the effectiveness of this approach, we run naturalistic simulations using the Facebook 100 dataset and illustrate the potential of using network information for observational inference in general. We also explore how to tune hyperparameters when the goal is to infer the causal effect rather than maximize prediction accuracy, as in most machine learning scenarios.

To summarize, the contributions of our study are the following:

- We explore and demonstrate the effectiveness of using network information for counterfactual inferences, including proxies for the latent positions and treatment assignments with network characteristics.
- We conceptualize conditions in which the network information can be used in counterfactual predictions.
- We propose a novel framework—representation learning with group lasso regularization—to minimize bias amplification and bias unmasking driven by instrumental variables and fixed effects.

2 Problem formulation

The goal of our study is to estimate the average treatment effect (ATE) by conditioning on the observed covariates and the network information. We may also wish to estimate conditional average treatment effects (CATEs) for particular subgroups. We use potential outcomes notation [26] to define these quantities. For an individual i, we observe i's covariates $\mathbf{X}_i$, i's row of the adjacency matrix $\mathbf{G}_i$, and whether they were assigned to the treatment or control $T_i \in \{0, 1\}$. There are two corresponding potential outcomes $Y_0^{(i)}$ and $Y_1^{(i)}$. For notation convenience, we use bold symbols to represent vectors or matrices, and non-bold symbols to represent scalars. The treatment effect (CATE) for i is the difference between their potential outcomes under treatment and control, $Y_1^{(i)} - Y_0^{(i)}$. This quantity is unobservable because it involves i being observed in two different states. However, we may be able to summarize treatment effects of many units, some of which are observed in treatment and some in control. The average treatment effect (ATE) is

$$\begin{aligned} \text{ATE} &= \mathbb{E}\big(Y_1^{(i)} - Y_0^{(i)}\big) \\ &= \mathbb{E}\big(Y_1^{(i)}\big) - \mathbb{E}\big(Y_0^{(i)}\big) \end{aligned} \tag{1}$$

In order to estimate the ATE, which involves means potential outcomes and not yet any observed outcomes, researchers typically impose the following assumptions:

Assumption 1 *Stable Unit Treatment Value Assumption (SUTVA): The potential outcomes for any units do not vary with the treatments assigned to other units, and for each unit, there are no differences in the forms or versions of each treatment level, which lead to different potential outcomes.*

Assumption 2 ***(Consistency)**. The potential outcome of treatment t equals to the observed outcome if the actual treatment received is t, i.e., if $T_i = 0$, we observe $Y_0^{(i)}$ and if $T_i = 1$, we observe $Y_1^{(i)}$.*

Assumption 3 ***(Positivity)**. For any set of covariates $\mathbf{x}$, the probability to receive treatment 0 or 1 is positive, i.e., $0 < p(\mathbf{T} = t|\mathbf{X} = \mathbf{x}) < 1, \forall t$ and x.*

Assumption 4 ***(Ignorability)**. Conditional on observed covariates, the potential outcomes are independent of which treatment was received, i.e., $\big(Y_1^{(i)}, Y_1^{(i)}\big) \perp\!\!\!\perp T_i|X_i$*

Ignorability (i.e., conditional unconfoundedness, selection on observables, no-hidden confounders) is typically a controversial assumption in observational studies. It requires that the outcome variables are independent of treatment assignment given observed covariate $\mathbf{X}$ [14, 12]. However, this assumption is unrealistic in practice since it is typically impossible to directly measure confounders in the real world. There are some settings where this might be plausible, if we know something about how units select into treatment; for example, we may know that users were treated based on predictions of a machine learning model plus noise, but perhaps have not retained the details of the model, but only which variables it used. But in other cases, we have limited knowledge of how units became treated, as often units are self-selecting into treatment based on information known only to them. We consider two families of cases that relax this assumption somewhat. One in which we have noisy measures of all unobserved confounders and one in which there are further unobserved confounders.

Networks are common data sources and may reveal relevant latent information due to homophily and past social influence [22]. Therefore, we use the information encoded in the network to account for some of the otherwise unobserved confounders. Our empirical evaluation uses a social network, but the ideas are readily applicable to other similar data structures, including bipartite graphs.

To illustrate how a network can be useful in causal inference, we illustrate with four causal directed acyclic graphs (DAGs; Figure 1). We can use the backdoor criterion [27] to determine in which DAGs the ATE of T on Y is identified.

In Figure 1(a) and (c), network reveals the latent covariates. In Figure 1 (b) and (d), network positions influence one's exposures to treatments. Conditional on $\mathbf{X}$ and $\mathbf{\Phi}$, ignorability holds for the models in Figures 1 (a) and (b), but not for the models in Figures 1 (c) and (d). In Figure 1(a), the observed covariates $\mathbf{X}$ and the latent position $\mathbf{\Phi}$ inferred from the network G both influence the treatment and the outcome. In addition to $\mathbf{\Phi}$ and $\mathbf{X}$, one's network position also influences the treatment and outcome in Figure 1(b). In these two figures, there exists no confounder other than $\mathbf{X}$ and $\mathbf{\Phi}$,

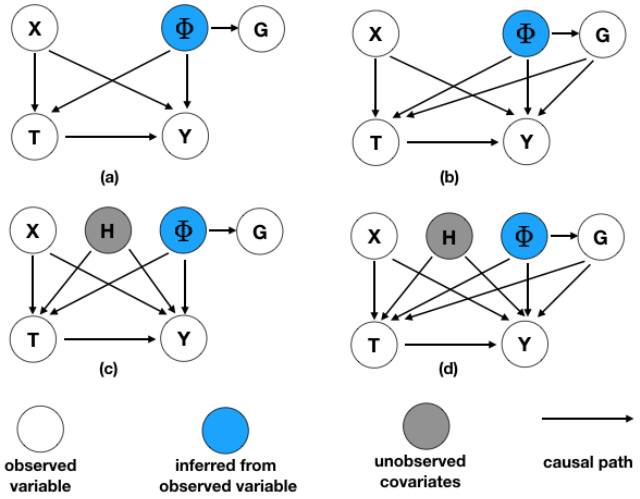

Figure 1: Illustration of the causal structures. $\mathbf{X}$: observed non-network covariates; $\mathbf{G}$: network; $\mathbf{\Phi}$: latent representation of the network; $\mathbf{H}$: unobserved covariates; $\mathbf{T}$ treatment; $\mathbf{Y}$: outcomes. White, blue and grey nodes correspond to observed variables, variables that are inferred from the observed and unobserved variables.

ignorability holds, and observed variables and proxies are causally sufficient to estimate the effect of $\mathbf{T}$ on $\mathbf{Y}$. Effective usage of network information helps to identify the treatment effect.

In Figure 1 (c)-(d), the latent confounders $\mathbf{H}$ contribute to both $\mathbf{T}$ and $\mathbf{Y}$. In this case, ignorability is typically violated and the ATE is not identified because $\{\mathbf{X}, \mathbf{\Phi}, \mathbf{G}\}$ are not sufficient to block the backdoor path between $\mathbf{T}$ and $\mathbf{X}$. Even worse, if the weights on the dashed arrow are zero, $\mathbf{X}$, $\mathbf{\Phi}$ and $\mathbf{G}$ can be instrumental variables, which may amplify bias if being controlled [20]; similar problems apply if they are "nearly" instruments, having only small effects on the outcome.

Adjusting for covariates will amplify the bias in the following conditions. Readers are encouraged to refer to Middleton et al. [18] and Pearl [20] for more details.

- *Instrumental variable*. If instrumental variables are controlled, bias will be amplified. The bias amplification will be more problematic if the controlled covariates account for a great deal of variation in the instrument. However, with the existing representation learning framework, such instrumental variables are more likely to be selected due to its strong predictive power in the treatment.
- *Fixed effects*. Including coefficients for each level of a categorical variable ("fixed effects") is often thought to be useful for absorbing unmeasured group-level confounding. They can function as bias amplifiers when the treatment and outcome coefficients for the same level of the categorical variable have little correlation.

3 Method

In this section, we describe our method to utilize information in the network effectively and to reduce potential bias amplification. In particular, our objective function predicts the treatment and the outcome simultaneously. A desirable property is that if certain representation associates weakly with the outcome and strongly with the treatment, it should not be included in the model. With these properties, we aim to learn representations whose coefficients are large on *both* the treatment and the outcome.

To achieve this, we use group lasso to select the representations learned from the network and, in particular, regularize the coefficients of such representations on the treatment and the outcome [25]. In this case, the groups in the group lasso are the coefficients that correspond to the same embedding dimension in the treatment and outcome prediction layers. See illustration in Figure 2.

Formally, we define the model as follows:

$$\begin{aligned}
\Phi^{(1)} &= \rho(AW^{(1)}), \\
\Phi^{(i)} &= \rho(\Phi^{(i-1)}W^{(i)}), \quad (\text{for } i = 2 \dots k), \\
\hat{y} &= [\Phi^{(k)}; X; t]w_y + \lambda_2 ||w_y||_2, \\
\hat{t} &= \sigma([\Phi^{(k)}; X]w_t) + \lambda_2 ||w_t||_2, \\
y_{loss} &= ||\hat{y} - y||_2, \\
t_{loss} &= \mathrm{H}(\hat{t}, t), \\
\mathcal{L} &= \gamma y_{loss} + (1-\gamma) t_{loss} + \lambda_g ||w_y, w_t||_g,
\end{aligned} \tag{2}$$

where $W^{(1)} \dots W^{(k)}$ are fully-connected layers; w_y, w_t are the weights in the final layers; $||\bullet||_2$ correspond to the l_2 norm; $[\bullet;\bullet]$ is matrix concatenation; ρ is the ReLU and σ is the sigmoid function; H is cross-entropy; and finally the group lasso term is:

$$||w_y, w_t||_g = \sum_{l=1}^{L} \sqrt{(w_y^{(l)})^2 + (w_t^{(l)})^2}. \tag{3}$$

The hyper-parameters λ_2, and λ_g control the importance of the l_2 and group lasso regularization terms, and γ trades off minimizing the outcome versus treatment prediction loss.

At a high-level, the model consists of two parts: an embedding network and a prediction network. In the embedding network, we feed the adjacency matrix $\mathbf{A}$ and learn an embedding $\mathbf{\Phi}$. The layers in the embedding network are fully connected. In the prediction network, we use the network embeddings $\mathbf{\Phi}$, and the observed covariates $\mathbf{X}$ to simultaneously predict the treatment and the outcome.

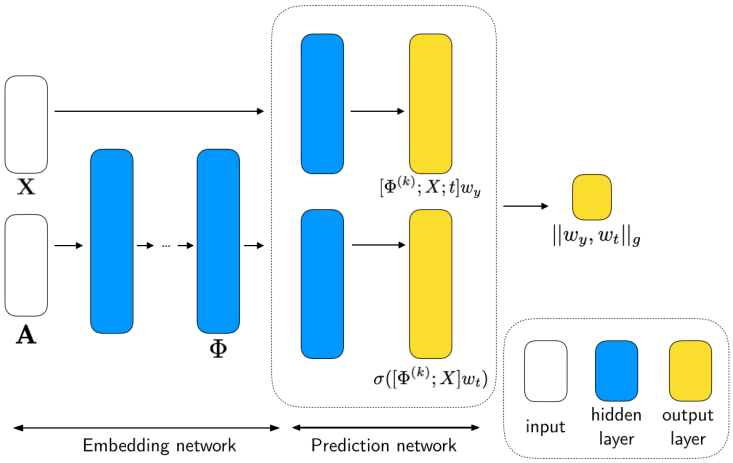

Figure 2: Illustration of the model architecture. The model consists of two main parts: an embedding network and a prediction network. The goal of the embedding network is to learn the representations from the network input $\mathbf{A}$. The output $\mathbf{\Phi}$ is then feed into the prediction network where we jointly predict the outcome $\mathbf{Y}$ and treatment $\mathbf{T}$. $\boldsymbol{w_t}$ and $\boldsymbol{w_y}$ are regularized jointly via group lasso.

4 Evaluation

Datasets. To test our approach, we perform a naturalistic simulation: we take the network and covariate information form a real-world dataset, but we use a model to simulate the treatment assignment and the outcomes. We use data from the Facebook 100 [28, 29] dataset, which contains data from the first 100 U.S. colleges that were on the platform in 2005. It includes (i) the online friendship networks among the users and (ii) the users' attributes, including student/faculty status, gender, major, dorm, year, and their high school. We also run the Louvain community detection

Table 1: Error in ATE when network positions reveal relevant characteristics of individuals

	CFR (covariates)	CFR (covariates + network)	Proposed
Reed	0.61	0.61	0.08
Haverford	0.53	0.08	0.02
Caltech	1.20	1.20	0.65

Table 2: Error in ATE when the community influences the treatment

	CFR (covariates)	CFR (covariates + network)	Proposed
Reed	0.50	0.50	0.34
Haverford	0.33	1.05	0.32
Caltech	0.70	0.70	0.66

algorithm [30] and assign each user to a community based on their network position. In order to run repeated evaluations, we focus on the three smallest colleges: Reed College (Reed), California Institute of Technology (Caltech) and Haverford College (Haverford).

We treat each covariate as an observed or unobserved confounder, or as an instrument. We simulate the treatment assignments and the outcomes as follows. First, we generate the treatment (w_t) and outcomes (w_y) weights: (a) for confounders we generate the treatment and outcome weights jointly by drawing from a 2D Gaussian distribution with mean $[0, 0]$ and covariance $[1.0, 0.5; 0.5, 1.0]$, (b) for instruments we generate the treatment weights by drawing them from 1D Gaussian with mean 0 and variance 1 and set the outcome weights to 0. Second, we compute the propensity scores,

$$\mu = \sigma(x \cdot w_t), \tag{4}$$

which is a linear model where σ is the logistic function, and we generate the treatment assignment t by drawing from a Bernoulli distribution with probability of success μ. Finally, to generate the outcomes we multiply the outcome weights by the observed covariates and add the ground truth ATE if the unit is assigned to treatment:

$$y = x \cdot w_y + \Delta t. \tag{5}$$

Settings. We explore two experimental settings where the network plays a different role:

- Network positions reveal latent and relevant characteristics of individuals (Figure 1c). However, one's network positions do not directly appear in the treatment and outcome function. More specifically, we set the person's status (student or faculty) and gender as observed confounders, the person's dorm and year as unobserved confounders, and their major as an instrument. Note that controlling for the major may amplify bias.
- The community in which each individual belongs to influences the treatment they receive (Figure 1d), but it does not influence the outcome. That is, we explore the robustness of the method with a categorical variable describing network position that is an instrument. We keep the same observed and unobserved confounders.

We compare our method to two variations of the Counterfactual Regression (CFR) model [15], which uses Wasserstein distance to balance the distribution between the treatment and control group: one only takes in covariates, and the other takes in both the covariates and the network.

The methods are evaluated in the following way. We first select the best hyper-parameters using 10-fold cross-validation. The model is trained on the nine splits of the data, and evaluated, according to the average prediction error on the outcome, on the remaining one split. The best hyperparameter is chosen to be the set that has the smallest average prediction error on the validation sets. After that, we train the full dataset with the chosen hyperparameters and compute the average treatment effect using the full data set.

The performance for setting 1 is show in Table 1. Using the network with CFR reduces bias for Haverford, but not for Reed and Caltech. The proposed method reduces error in all three cases.

The performance for setting 2, where community membership is a categorical instrument, are shown in Table 2. Here using the network with CFR increases bias in the same subset where it previously reduced bias (Haverford), reflecting that some aspects of latent position are functioning as bias amplifying instruments. On the other hand, the proposed method constributes to reduce bias in all cases, albeit less so than in setting 1.

5 Related work

Here we highlight the relationship with other work on representation learning for causal inference and network data.

There has been a recent interest in using representation learning in causal inference. This literature is started by Johansson et al. [14], and followed by [15] with theoretical guarantees. In particular, they tackled the problem by using representation learning in estimating individual treatment effects. [16, 12] further this literature by adding balanced representation to reduce the high-dimensional covariates to a lower subspace for causal inference. Alaa and van der Schaar [31] proposes an approach to infer the individual causal effect of a treatment using a Bayesian approach to learn the treatment effects through a multitask Gaussian process prior to the population's potential outcomes. This literature builds upon the unconfoundedness assumption, which is unrealistic in most real-world situations.

There is also work using network data to account for otherwise unobserved confounding. [32] shows that when the network grows according to either a latent community model or continuous latent space model, latent attributes can be consistently estimated from the global pattern of social ties, and hence the causal estimates can be identified. This is based on the assumption that after controlling for the network connections, there is no additional information on the treatment and the control function one does not observe. A similar assumption is used in [24]. Auerbach [24] introduces a method based on matching pairs of agents with similar columns of $\mathbb{G}^2$. For a class of network formation models, the columns of this matrix characterize all of the identifiable information about individual linking behavior. Unlike the present work, both of these papers assumme that there is not additional unobserved confounding.

Louizos et al. [17] is closer to our study, which also deals with hidden confounders and does not constrain the relationship between the observed covariates and the latent attributes. They propose to learn the causal effect with noisy measurement of the hidden confounders using a standard variational autoencoder framework. This framework helps to relax the parametric assumptions on the relationship between latent and observed covariates. This paper can be regarded as a non-parametric version of Shalizi and McFowland III [32]. While Shalizi and McFowland III [32] rely on prior knowledge about the network formation process. However, Louizos et al. [17] rely on the neural network to learn such complex relationships. They focuses only on the causal graph, as depicted in Figure 1(a), in which the observed covariates contain all the variations in the unobserved confounders. This is a special case of our framework and will perform poorly in the case of unobserved confounding and bias amplification or bias unmasking.

6 Conclusion

Networks, if utilized effectively, can provide information about otherwise unobserved variables of the nodes they connect. Yet other unobserved variables may still be present, such that naively using this data may even increase bias. This type of bias amplification is neglected by most of the literature, including recent developments in computer science on utilizing high-dimensional covariates in causal inference.

In this paper, we provide the conditions in which network information can amplify or unmask the bias. We propose a framework to utilize the information encoded network for counterfactual inferences effectively. Our framework integrates representation learning with group lasso regularization to minimize bias amplification driven by instrumental variables and near instruments, such as community fixed effects. We apply our method to the naturalistic simulations on empirical data, showing that our method outperforms the state-of-the-art.

An important future direction is to provide theoretical bound for the bias in the average treatment effect estimation when networks encode relevant, but not all, information for treatment and outcome predictions.

References

[1] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

[2] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[3] Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448): 1053–1062, 1999.

[4] Randall A Lewis, Justin M Rao, and David H Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th international conference on World wide web*, pages 157–166. ACM, 2011.

[5] Dean Eckles and Eytan Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*, 2017.

[6] Tomasz Kusmierczyk and Manuel Gomez-Rodriguez. On the causal effect of badges. In *Proceedings of the 2018 World Wide Web Conference*, pages 659–668. International World Wide Web Conferences Steering Committee, 2018.

[7] Yan Leng, Xiaowen Dong, Esteban Moro, et al. The rippling effect of social influence via phone communication network. In *Complex Spreading Phenomena in Social Systems*, pages 323–333. Springer, 2018.

[8] David Stück, Haraldur Tómas Hallgrímsson, Greg Ver Steeg, Alessandro Epasto, and Luca Foschini. The spread of physical activity through social networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 519–528. International World Wide Web Conferences Steering Committee, 2017.

[9] Jessica Su, Aneesh Sharma, and Sharad Goel. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*, pages 1157–1167. International World Wide Web Conferences Steering Committee, 2016.

[10] Nir Rosenfeld, Yishay Mansour, and Elad Yom-Tov. Predicting counterfactuals from large historical data and small randomized trials. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 602–609. International World Wide Web Conferences Steering Committee, 2017.

[11] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[12] Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pages 929–939, 2017.

[13] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[14] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

[15] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[16] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2638–2648, 2018.

[17] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[18] Joel A Middleton, Marc A Scott, Ronli Diakow, and Jennifer L Hill. Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323, 2016.

[19] Judea Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*, 2012.

[20] Judea Pearl. Invited commentary: Understanding bias amplification. *American journal of epidemiology*, 174(11):1223–1227, 2011.

[21] Jessica A Myers, Jeremy A Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–1222, 2011.

[22] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[23] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[24] Eric Auerbach. Identification and estimation of a partially linear regression model using network data. *arXiv preprint arXiv:1903.09679*, 2019.

[25] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[26] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[27] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[28] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[29] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[30] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[31] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[32] Cosma Rohilla Shalizi and Edward McFowland III. Estimating causal peer influence in homophilous social networks by inferring latent locations. *arXiv preprint arXiv:1607.06565*, 2016.