Modeling Interference for Individual Treatment Effect Estimation from Networked Observational Data

QIANG HUANG, School of Artificial Intelligence, International Center of Future Science, Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, China
JING MA, Department of Computer and Data Sciences, Case Western Reserve University, USA
JUNDONG LI, University of Virginia, USA
RUOCHENG GUO, ByteDance Research, United Kingdom
HUIYAN SUN, School of Artificial Intelligence, International Center of Future Science, Jilin University, China
YI CHANG, School of Artificial Intelligence, International Center of Future Science, Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, China

Estimating *individual treatment effect* (ITE) from observational data has attracted great interest in recent years, which plays a crucial role in decision-making across many high-impact domains such as economics, medicine, and e-commerce. Most existing studies of ITE estimation assume that different units at play are independent and do not influence each other. However, many social science experiments have shown that there often exist different levels of interactions between units in observational data, especially in a networked environment. As a result, the treatment assignment of one unit can affect the outcome of other units connected to it in the network, which is referred to as the *interference* or *spillover effect*. In this article, we study an important problem of ITE estimation from networked observational data by modeling the interference between different units and provide a principled framework to support such study. Methodologically, we propose a novel framework, *SPNet*, that first captures the influence of hidden confounders with the aid of graph convolutional network and then models the interference by introducing an environment summary variable and developing a masked attention mechanism. Experimental evaluations on several semi-synthetic datasets based on real-world networks corroborate the superiority of our proposed framework over state-of-the-art individual treatment effect estimation methods.

CCS Concepts: • **Computing methodologies** → *Causal reasoning and diagnostics;*

Additional Key Words and Phrases: Causal inference, ITE estimation, network interference

This work was supported in part by the National Natural Science Foundation of China under grants Nos. U19A2065, 61976102 and 62372210.

Authors' addresses: Q. Huang and Y. Chang (Corresponding author), School of Artiicial Intelligence, International Center of Future Science, Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, 2699 Qianjin Street, Changchun, Jilin, 130012, China; e-mails: huangqiang18@mails.jlu.edu.cn, yichang@jlu.edu.cn; J. Ma, Department of Computer and Data Sciences, Case Western Reserve University, 2101 Martin Luther King Jr Dr, Cleveland, OH, 44106, USA; e-mail: jing.ma5@case.edu; J. Li, University of Virginia, 351 McCormick Road, Charlottesville, VA, 22904, USA; e-mail: jundong@virginia.edu; R. Guo, ByteDance Research, 4 Lindsey St, Barbican, London, Great London, EC1A 9HP, United Kingdom; e-mail: rguo.asu@gmail.com; H. Sun (Corresponding author), School of Artiicial Intelligence, International Center of Future Science, Jilin University, 2699 Qianjin Street, Changchun, Jilin, 130012, China; e-mail: huiyansun@jlu.edu.cn.

1556-4681/2023/12-ART48 $15.00
https://doi.org/10.1145/3628449

ACM Reference format:
Qiang Huang, Jing Ma, Jundong Li, Ruocheng Guo, Huiyan Sun, and Yi Chang. 2023. Modeling Interference for Individual Treatment Effect Estimation from Networked Observational Data. *ACM Trans. Knowl. Discov. Data.* 18, 3, Article 48 (December 2023), 21 pages.
https://doi.org/10.1145/3628449

1 INTRODUCTION

Many scientific problems aim to investigate the causal effects between different variables at play. For example, scientists try to assess to what extent smoking can causally affect human health conditions. The ideal way to study the causal effect would be to conduct a randomized controlled trials [5, 14, 55], in which units (experimental subjects) are randomly divided into two groups, the treatment group and control group, and then the causal effect can be estimated by comparing the difference between the potential outcomes in these two groups. However, such randomized controlled trials are often unethical, time-consuming, and expensive to conduct [16, 53]. Fortunately, with the rocketing growth of big data in many high-impact domains (e.g., economics, health care, and social media), it is often effortless to collect a large amount of observational data, which provides great opportunities for causal inference research [16, 52, 53]. A natural question here is as follows: How can the outcome be precisely inferred if the unit has taken another treatment? Such a problem is well known as the *counterfactual outcome prediction problem*. After the counterfactual outcomes are predicted, we can easily estimate the treatment effect for each individual, i.e., **individual treatment effect (ITE)** [11, 40, 42].

A vast majority of existing works on ITE estimation assumes that units of observational data are independent of each other [2, 20, 26, 32, 50]. Nevertheless, units (or instances) of observational data in many settings are inherently connected (e.g., social network among users in social media and contact network among individuals during pandemic), and such data are often referred to as *networked observational data* [19, 41]. Over the past few years, many ITE estimation frameworks have been proposed for networked observational data [9, 17, 19]. Despite their empirical success in ITE estimation, these works mainly focus on utilizing network structure information to control for confounders (i.e., factors that influence both treatment assignment and outcome and thus often bring confounding bias to ITE estimation), while ignoring the interference of instances in a network [35, 38]. Here interference (a.k.a spillover effect) refers to a situation where experimental units usually interact with each other, and as a result, one unit's outcome is not only affected by its own features (covariates) and treatment but also influenced by the treatment assignments of other units that are connected to it. For example, a person who receives the COVID-19 vaccine will lower the infection risk of those people who do not get the vaccine in his or her social circle (e.g., families, friends, colleagues, etc.). Here receiving a vaccine can be viewed as a treatment, and the risk of COVID-19 infection can be viewed as an outcome. In fact, the significance of considering such interference has been demonstrated in many causal inference literature [1, 4, 12, 13, 21, 31]. Ignoring network interference can lead to biased or incorrect causal effect estimates and policy recommendations. For instance, if a treatment has positive effects on treated units but negative effects on their neighbors, then the causal effect of the treatment on the whole population may be small or negative, which would not be apparent if network interference is not taken into account.

In this work, we study an important problem of modeling the interference (i.e., spillover effect) for learning ITE from networked observational data. However, it remains a daunting task mainly because of the following challenges. First, as mentioned previously, a unit's outcome is not only affected by its features (i.e., covariates) and treatment; it is also influenced by the treatments of its neighbors. Hence, how to model these joint effects in a principled framework to characterize the

outcome of each unit is challenging. Second, networked data are often characterized by complex, high-dimensional covariates, and the interaction behavior of nodes is often non-linear. Therefore, how to effectively model network interference in networked data with complex interactions is very critical to estimate precise causal effects. Third, the magnitude of spillover effects between different pairs of neighboring units is naturally different (e.g., a vaccinated person will have a stronger effect in lowering the infection risk of her immediate family than a random acquaintance). Thus, how to measure the varied magnitude of spillover effects between neighboring units is an important problem. Fourth, the existence of hidden confounders (i.e., unobserved confounders) may yield a biased ITE estimation [39]. Although recent studies [17, 19] have shown that the auxiliary network information of observational data can be leveraged to control the influence of hidden confounders, how to better infer the hidden confounders from the networked observational data in the presence of interference remains an open problem.

To address the aforementioned challenges, we propose a novel individual treatment effect estimation framework for networked observational data under interference, named *SPNet*, which can capture the hidden confounders and modeling the varied magnitude of spillover effects by building a two-channel graph convolutional network and a masked-attention mechanism. The main contributions of this work are summarized as follows:

- We study an important research problem of modeling interference for learning individual treatment effect from networked observational data.
- We propose a novel individual treatment effect estimation framework, *SPNet*, which models interference between units in the network and controls the confounding bias by utilizing network information.
- We provide and prove a formal theorem about the identifiability of individual treatment effect under network interference.
- We conduct extensive experiments to show that the proposed framework *SPNet* is superior to existing state-of-the-art methods for estimating individual treatment effects on networked observational data.

2 PRELIMINARIES

In this section, we begin with notations used in this article and then formally define the problem of *modeling interference for learning individual treatment effects from networked observational data* by introducing the technical preliminaries.

Notations. In this work, we use unbold capital letters to denote random variables (e.g., $Y_i^{T_i=1}$) and use unbold lowercase letters (e.g., $y_i^{t_i}$), bold lowercase letters (e.g., $\boldsymbol{x}_i$), and bold capital letters (e.g., $\boldsymbol{A}$) to denote specific values of scalars, vectors, and matrices, respectively. We use $\boldsymbol{A}_{ij}$ to represent the (i, j)-th entry of matrix $\boldsymbol{A}$.

Networked observational data. We use $\boldsymbol{x}_i$, t_i, and y_i to denote the covariates (i.e., features), the assigned treatment, and the observed outcome of the ith instance, respectively. Without loss of generality, we use $t_i = 1$ ($t_i = 0$) to denote that the ith instance is in the treatment (control) group. The potential outcome $y_i^{t_i}$ of the treatment assignment is defined as the value that the outcome would have taken if the treatment of unit i had been set to t_i. Additionally, the network structure among n instances is encoded in an adjacency matrix $\boldsymbol{A} \in \{0, 1\}^{n \times n}$ such that $\boldsymbol{A}_{ij} = 1$ ($\boldsymbol{A}_{ij} = 0$) denotes that there exits an (no) edge between instance i and j. As a summary, we refer $\{\{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n, \boldsymbol{A}\}$ as the networked observational data, where $\boldsymbol{x}_i \in \mathbb{R}^m$ and m is the dimension of the observed instance features.

Definition 1 (Network Interference). Interference (i.e., spillover effect), exists when different units are related with each other (e.g., social network among users), such that one unit's outcome

is causally affected by the treatment assignments of other units that are connected to it in the network.

It should be noted that many existing causal inference studies [9, 17, 19] on networked environment fail to consider the existence of spillover effects among different units. In this work, we study this important research problem of modeling interference for ITE estimation from networked observational data.

Traditional ITE without network interference aims to measure the difference between potential outcomes under treatment and control, i.e., $\tau_i = \mathbb{E}[Y_i^{T_i=1}|\boldsymbol{x}_i] - \mathbb{E}[Y_i^{T_i=0}|\boldsymbol{x}_i]$. However, such estimation is not applicable when interference (spillover effect) exists among different units. To account for the influence of neighboring units, in this work, we introduce an environment summary $\boldsymbol{e}_i$ for each unit i, which can summarize (e.g., mean) over the treatment assignment information of the unit i's neighbors. With this, the individual treatment effect of unit i under network interference is defined as $\tau_i = \mathbb{E}[Y_i^{T_i=1}|\boldsymbol{x}_i, \boldsymbol{e}_i] - \mathbb{E}[Y_i^{T_i=0}|\boldsymbol{x}_i, \boldsymbol{e}_i]$. One of the most important tasks in this work is to model the interference over the current environment. It should be noted that we assume the interference from neighboring units for each unit i remains unchanged when we intervene its treatment assignment T_i. Using the above, we provide a formal definition of the studied problem as follows:

Definition 2 (Modeling Interference for Learning Individual Treatment Effect from Networked Observational Data). Given the networked observational data $\{\{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n, \boldsymbol{A}\}$, our goal is to learn the ITE τ_i $(i = 1, 2, \cdots, n)$ for each unit i while accounting for the interference among different units in $\tau_i = \mathbb{E}[Y_i^{T_i=1}|\boldsymbol{x}_i, \boldsymbol{e}_i] - \mathbb{E}[Y_i^{T_i=0}|\boldsymbol{x}_i, \boldsymbol{e}_i]$.

It is noteworthy that many existing works [25, 40, 53] on individual treatment effect estimation are developed based upon the following two fundamental assumptions: The **Stable Unit Treatment Value Assumption (SUTVA)** and the Strong Ignorability [19, 24]. SUTVA assumption emphasizes the independence of each unit, such that the treatment of one unit does not affect the potential outcomes of other units. Most of the early works on causal inference relied on this SUTVA assumption. However, there are many scenarios whereby the interference between units matters [23, 44, 48]. In this work, we aim to model the spillover effect and control the confounding bias for more accurate individual treatment effects, which does not rely on the SUTVA assumption. In other words, we assume different units interact with each other. Strong Ignorability assumes that there are no hidden confounders and all the confounding variables can be measured by the observed features. However, the assumption is often untenable in many real-world scenarios, especially when the observed feature space is very limited. In this work, we consider a more practical scenario where the hidden confounders exist, and we will exploit the network structure among instances to better control these hidden confounders toward unbiased ITE estimation.

3 THE PROPOSED FRAMEWORK: *SPNET*

An overview of the proposed framework, *SPNet*, which models the interference for learning individual treatment effect from networked observational data, is shown in Figure 1. The proposed framework has the following key components: (1) *Learning Partial Representations*, (2) *Modeling Interference*, and (3) *Predicting Potential Outcome and Treatment Assignment*. We will describe the three key components in detail in the following sections.

3.1 Learning Partial Representations

The first component of the proposed framework *SPNet* is to learn two partial representations of hidden confounders for each unit by utilizing both instance features and network structure

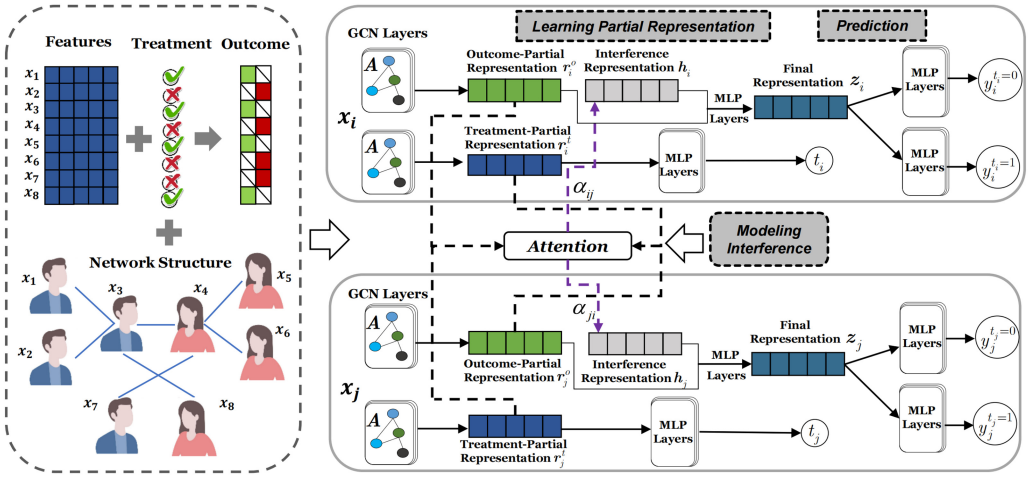

Fig. 1. The overview of the proposed *SPNet*, which models interference, taking neighboring pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ as an example.

information. Among them, one partial representation is mainly for the potential outcome prediction and the other one is for the treatment assignment prediction. There are two main reasons for learning these two partial representations. First, learning two partial representations to capture hidden confounders can help control the influence of confounding bias for ITE estimation and treatment assignment prediction. Second, learning two partial representations can help to model the interference between different connected units. For these two partial representations, *SPNet* aims to learn two partial representation learning functions g^o and $g^t : \boldsymbol{X} \times \boldsymbol{A} \rightarrow \mathbb{R}^d$, which map the observed features along with the network structure information to a d-dimensional latent space with the supervision of factual outcomes and observed treatment assignments, respectively. Specifically, we parameterize these two partial representation learning functions by a twin-channel **graph convolutional network (GCN)** [10, 30], which has shown to be effective in handling non-Euclidean data (e.g., networked structure data) in diverse settings.

To be more specific, we propose stacking multiple GCN layers to develop a twin-channel network model to obtain these two partial representations of hidden confounders. For the simplicity of notation, we represent the partial representation learning functions g^o and g^t with a single GCN layer. Each partial representation learning function can be formulated as follows:

$$\begin{aligned} \boldsymbol{r}_i^o &= g^o(\boldsymbol{x}_i, \boldsymbol{A}) = \sigma((\hat{\boldsymbol{A}}\boldsymbol{X})_i \boldsymbol{W}^o) \\ \boldsymbol{r}_i^t &= g^t(\boldsymbol{x}_i, \boldsymbol{A}) = \sigma((\hat{\boldsymbol{A}}\boldsymbol{X})_i \boldsymbol{W}^t), \end{aligned} \tag{1}$$

where σ is the ReLU activation function for each GCN layer. $\boldsymbol{r}_i^o \in \mathbb{R}^d$ and $\boldsymbol{r}_i^t \in \mathbb{R}^d$ are the partial representations of hidden confounders of unit i with respect to the potential outcome and treatment, respectively. d is the embedding size. $\hat{\boldsymbol{A}} \in \mathbb{R}^{n \times n}$ is the normalized adjacency matrix with self-loop such that

$$\hat{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{-\frac{1}{2}} \tilde{\boldsymbol{A}} \tilde{\boldsymbol{D}}^{-\frac{1}{2}}, \quad \tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}_n, \tag{2}$$

where $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$ is an identity matrix and $\tilde{\boldsymbol{D}}_{ii} = \sum_{j=1}^{n} \tilde{\boldsymbol{A}}_{ij}$ is the degree matrix of $\tilde{\boldsymbol{A}}$. $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ is the observed feature matrix of the units, $(\hat{\boldsymbol{A}}\boldsymbol{X})_i$ denotes the ith row of the product of matrix multiplication $\hat{\boldsymbol{A}}\boldsymbol{X}$, and $\boldsymbol{W}^o$ and $\boldsymbol{W}^t \in \mathbb{R}^{m \times d}$ stand for the parameter matrix to be learned. Then these two partial representations of unit i will be later used for modeling the interference.

3.2 Modeling Interference

Regarding the networked observational data, the outcome of a unit can be affected by the treatments of its neighboring units, which is referred as *interference*. In this work we introduce an interference representation $\boldsymbol{h}_i$ to model such effect. Previously, many methods [34] directly aggregate the treatment assignments of a unit's neighbors into a one-dimensional variable (e.g., with mean pooling) and model the interference based on it. However, such solution is prone to failure when the observational data are high dimensional. To tackle this issue, a more principled way is desired to model the interference. Meanwhile, as mentioned previously, the magnitude of the spillover effects between neighboring units should differ. Such difference should also be reflected in the interference representation. As a summary, by reason of the foregoing, two questions need to be answered for modeling such spillover effect: *How do we model the interference that captures the influence of one unit's neighboring units on its own outcome? and How do we quantify the varied magnitude of the spillover effects between different neighboring units?*

To answer the first research question, we propose to model the interference representation $\boldsymbol{h}_i$ for each unit i by aggregating the treatment-partial representations of its neighbors. The main reason is that these treatment-partial representations embed units' information that is used for the treatment assignment prediction. And it is worth noting that we use the interference representation $\boldsymbol{h}_i$ as the proxy of the environment summary $\boldsymbol{e}_i$. In this way, the spillover effect over a unit that is resulted by the neighbors' treatment assignments can be well captured. Meanwhile, to answer the second research question of quantifying the varied magnitude of spillover effects between different neighboring units, we propose a masked-attention mechanism by leveraging the learned outcome-partial and treatment-partial representations $\boldsymbol{r}_i^o, \boldsymbol{r}_i^t$. More specifically, we aim to learn an attention function $a(.,.)$ that can assign different weights to different neighboring units during the information aggregation process of the spillover effect modeling. With the above considerations, we formulate the interference representation $\boldsymbol{h}_i \in \mathbb{R}^d$ for each unit i as follows:

$$\boldsymbol{h}_i = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \cdot \boldsymbol{r}_j^t, \tag{3}$$

where $\mathcal{N}(i)$ denotes the set of neighbors of unit i in the network $\boldsymbol{A}$, $\boldsymbol{r}_j^t$ is the treatment-partial representation of the jth unit. α_{ij} is the attention weight that represents the magnitude of the spillover effect of the jth unit on the ith unit. The coefficient is computed by the attention function $a(.,.) : \mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}$. The computation of the attention weight α_{ij} by the attention function $a(.,.)$ is formulated as follows:

$$\begin{aligned} \boldsymbol{r}_i &= \text{concat}\left(\boldsymbol{r}_i^o, \boldsymbol{r}_i^t\right), \quad \boldsymbol{r}_j = \text{concat}\left(\boldsymbol{r}_j^o, \boldsymbol{r}_j^t\right), \\ \alpha_{ij} &= a(\boldsymbol{r}_i, \boldsymbol{r}_j) = \frac{\exp(\text{ReLU}(\boldsymbol{a}^T[\boldsymbol{r}_i || \boldsymbol{r}_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{ReLU}(\boldsymbol{a}^T[\boldsymbol{r}_i || \boldsymbol{r}_k]))}, \end{aligned} \tag{4}$$

where concat(.,.) and || both stand for the vector concatenation operation. $\boldsymbol{a} \in \mathbb{R}^{4d}$ is the weight vector to be learned, and the term $\text{ReLU}(\boldsymbol{a}^T[\boldsymbol{r}_i || \boldsymbol{r}_j])$ can be interpreted as the unnormalized attention weight of the edge between the units i and j. More specifically, by concatenating the two partial confounder representations of each instance when learning the attention weights, we can gain a more comprehensive understanding of how the treatment of each unit affects the outcome of other neighboring units. Noting that we consider a more realistic situation to model the interference, where the attention weights are not symmetric. This means that the magnitude of the interference of the unit i on unit j does not necessarily equal to that of the unit j on unit i, i.e., $\alpha_{ji} \neq \alpha_{ij}$.

3.3 Predicting Potential Outcome and Treatment Assignment

Until now, we have obtained the outcome-partial representation $\boldsymbol{r}_i^o$ by Equation (1) and the interference representation $\boldsymbol{h}_i$ by Equation (3) for each unit i, where $\boldsymbol{r}_i^o$ encodes the information of confounders for inferring unit i's potential outcomes while $\boldsymbol{h}_i$ encodes the spillover effects resulted from treatment assignments of unit i's neighbors. We first combine (e.g., sum up) the outcome-partial representation $\boldsymbol{r}_i^o$ and the interference representation $\boldsymbol{h}_i$ together and obtain a final representation $\boldsymbol{z_i}$ through a **multilayer perceptron (MLP)**, then we develop an output layer to infer the potential outcomes for each unit i by using the obtained final representation $\boldsymbol{z_i}$.

Specifically, we develop a mapping function $f_z : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ to map the outcome-partial representation $\boldsymbol{r}_i^o$ of hidden confounders, the interference representation $\boldsymbol{h}_i$, into the final representation $\boldsymbol{z}_i$ for unit i. The formulation of function f_z is as follows:

$$\boldsymbol{z}_i = f_z\left(\boldsymbol{r}_i^o, \boldsymbol{h}_i\right). \tag{5}$$

Then, with the final representation $\boldsymbol{z}_i \in \mathbb{R}^d$ and the corresponding treatment assignment $t_i \in \{0, 1\}$, the potential outcome of unit i w.r.t. the treatment assignment t_i can be computed by the following function f:

$$f(\boldsymbol{z}_i, t_i) = \begin{cases} \hat{y}_i^{\,t_i=0} = f_0(\boldsymbol{z}_i) & \text{if } t_i = 0 \\ \hat{y}_i^{\,t_i=1} = f_1(\boldsymbol{z}_i) & \text{if } t_i = 1 \end{cases}, \tag{6}$$

where the functions f_0 and f_1 are both parameterized by L fully connected layers followed by an output layer for $t_i = 0$ and $t_i = 1$, respectively.

Meanwhile, we also develop a treatment prediction function $f^t : \mathbb{R}^d \rightarrow (0, 1)$, which maps the treatment-partial representation $\boldsymbol{r}_i^t$ into the estimated probability $P(t_i = 1|\boldsymbol{r}_i^t)$, i.e., the probability that the unit i receives treatment $t_i = 1$. The treatment prediction function f^t is parameterized by K fully connected layers with a sigmoid activation function.

3.4 Loss Function of *SPNet*

Now, we design a loss term to combine all the essential components for ITE estimation, such that the learned partial representations that capture the hidden confounding bias and spillover effect can be well trained in an end-to-end manner.

Loss for Potential Outcome Prediction. First, we use **Mean Square Error (MSE)** loss function to minimize the error between the inferred factual outcome $\hat{y}_i^{t_i}$ by Equation (6) and observed factual outcome y_i for each unit and denote the factual outcome loss as $\mathcal{L}_y$,

$$\mathcal{L}_y = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i^{t_i} - y_i\right)^2. \tag{7}$$

Loss for Treatment Assignment Prediction. Second, we design another loss term to minimize the difference of the predicted treatment assignment by f^t and the true treatment assignment t_i. We view it as a classification problem using cross-entropy loss function and denote it as $\mathcal{L}_t$,

$$\begin{aligned} \mathcal{L}_t = -\frac{1}{n}\sum_{i=1}^{n} & t_i \log\left(P\left(t_i = 1|\boldsymbol{r}_i^t\right)\right) \\ & +(1 - t_i)\log\left(1 - P\left(t_i = 1|\boldsymbol{r}_i^t\right)\right). \end{aligned} \tag{8}$$

Loss for Representation Balancing. Since the distribution of the factual outcome and the counterfactual outcome could be rather different, we follow References [19, 43] to measure the difference between the representation distributions of hidden confounders for the treatment group and the control group. Here, we denote the representation balancing loss as $\mathcal{L}_b$.

The Overall Loss Function. The final loss function for *SPNet* is formulated as follows:

$$\mathcal{L} = \mathcal{L}_y + \alpha\mathcal{L}_t + \beta\mathcal{L}_b + \gamma\|\Theta\|_2^2, \tag{9}$$

where α and β are two non-negative hyperparameters to control the tradeoff between the corresponding term and other terms. The last term $\gamma\|\Theta\|_2^2$ is imposed on all the model parameters Θ to prevent overfitting.

4 ITE IDENTIFIABILITY UNDER INTERFERENCE

In this section, we provide and prove a formal theorem about the identifiability of individual treatment effect under network interference from networked observational data:

Theorem 1. *(Identifiability of ITE under interference) If we can recover $p(Z_i|E_i, X_i)$ and $p(Y_i|Z_i, T_i)$ from the networked observational data, then we can recover the ITE under network interference.*

Here X_i, Z_i, E_i, T_i denote the random variables for the covariates, the confounders, the environment summary, and the treatment assignment of unit i, respectively. In our framework, $p(Z_i|E_i, X_i)$ is modeled with the partial representation learning and interference modeling components. E_i is the environment summary variable, we assume a summary function [34] that can output the value of E_i to describe the treatment assignments and covariates of other units (e.g., neighbors of unit i) that influence unit i given the network structure $\boldsymbol{A}$. And $p(Y_i|Z_i, T_i)$ is modeled by the potential outcome prediction component. The proof of the theorem generally follows similar assumptions as the previous work [34, 36], but it is worth noting that we allow hidden confounders to exist in our setting, whereas the two literature mentioned above assume no existence of hidden confounders.

4.1 Preconditions

Before proving Theorem 1, we introduce a conjecture for modeling the interference in networked environment:

Conjecture 1. *There exists an environment summary variable $E_i = F_i(\boldsymbol{T}_{\mathcal{N}_i}, \boldsymbol{X}_{\mathcal{N}_i})$ for any unit i, which represents the aggregation of its neighbors' covariates and treatment assignments.*

In this conjecture, $F_i : \{0, 1\}^{|\mathcal{N}_i|} \times \mathcal{X}^{\times|\mathcal{N}_i|} \to \mathcal{E}_i$ represents the aggregation function for unit i, $\mathcal{N}_i$ denotes the set of unit i's neighbors, and $\boldsymbol{T}_{\mathcal{N}_i}, \boldsymbol{X}_{\mathcal{N}_i}$ denote the treatment assignments and covariates of unit i's neighbors, respectively. By the way, we infer the environment summary variable E_i by using GCN model and attention mechanism built in the proposed model *SPNet* based on the covariates and treatment assignments of unit i's neighbors determined by the network structure $\boldsymbol{A}$.

Based on the above definitions of environment summary variable and the corresponding aggregation function, we introduce another three necessary assumptions for the identifiability of ITE under network interference:

Assumption 1. *If the responses of the aggregation function F_i with respect to two different value assignments of $(\boldsymbol{T}_{\mathcal{N}_i}, \boldsymbol{X}_{\mathcal{N}_i})$ are identical for unit i, then the value of the potential outcome of unit i under these two value assignments is the same.*

Assumption 2. *Assuming that complete confounders (containing all observed and hidden confounders) can be captured from latent space of observational data. Then we refer the variable of representation of the complete confounders as Z_i for unit i.*

Assumption 3. *Given the variable of the representation of complete confounders Z_i, the treatment assignment T_i and environment summary variable E_i are independent to the potential outcome for unit i, i.e., $Y_i^{T_i=1}, Y_i^{T_i=0} \perp T_i, E_i|Z_i$. This assumption is a relaxed version of the widely used*

unconfoundedness assumption in causal inference. Notice that it is relaxed to allow the existence of hidden confounders.

4.2 Derivation of Proof

Finally, we formally present detailed derivation of the proof of Theorem 1 for the identifiability of individual treatment effect under network interference.

Proof. Given the above assumptions and observational data, we have the distribution of the potential outcome of any unit i under network interference as follows (we use $T_i = 1$ as an example):

$$\begin{aligned} & p(Y_i^{T_i=1}|E_i, X_i) \\ & \overset{(i)}{=} \int_{Z_i} p(Y_i^{T_i=1}|E_i, X_i, Z_i) p(Z_i|E_i, X_i) dZ_i \\ & \overset{(ii)}{=} \int_{Z_i} p(Y_i^{T_i=1}|E_i, Z_i) p(Z_i|E_i, X_i) dZ_i \\ & \overset{(iii)}{=} \int_{Z_i} p(Y_i^{T_i=1}|E_i, Z_i, T_i = 1) p(Z_i|E_i, X_i) dZ_i \\ & \overset{(iv)}{=} \int_{Z_i} p(Y_i|Z_i, T_i = 1) p(Z_i|E_i, X_i) dZ_i, \end{aligned} \tag{10}$$

where equation (i) is the straightforward expectation over $p(Z_i|E_i, X_i)$, equation (ii) can be inferred from Assumption 2, and equation (iii) is derived by Assumption 3 based on the conditional independence rule $Y_i^{T_i=1}, Y_i^{T_i=0} \perp T_i, E_i|Z_i$. On the one hand, a unit's observed outcome needs to account for the whole network; on the other hand, we assume that aggregating a unit's neighbouring information can summarize the influence of the whole network on its potential outcome in Assumption 1, thus equation (iv) can be derived based on the Assumption 1 and the widely used consistency assumption [24]. Thus, if our framework *SPNet* can correctly model $p(Z_i|E_i, X_i)$ and $p(Y_i|Z_i, T_i)$, then the distribution $p(Y_i^{T_i}|E_i, X_i)$ can be recovered, and thus the individual treatment effect can be identified under network interference. Noting that our framework *SPNet* approximates the aggregation function F_i using the GCN model and masked-attention mechanism based on the treatment assignments and covariates of unit i's neighbors according to the network structure $\boldsymbol{A}$ and then infers the representation of complete confounder Z_i by GCN as well as MLP, and finally the potential outcome is modeled by stacking multiple MLP layers based on the learned representation of complete confounder. The framework transforms the two probability distribution estimation into the parameter learning problem of neural network. One can see that modeling process of the proposed framework *SPNet* coincides with the above proof for ITE identifiability under interference. □

5 EXPERIMENTAL EVALUATIONS

In this section, we perform empirical experimental evaluations on multiple semi-synthetic datasets to assess the performance of our proposed framework *SPNet*. Before presenting the detailed experimental results, we first introduce the used datasets, evaluation metrics, and experimental settings. Then we report the experimental results of ITE estimation performance, robustness of *SPNet*, ablation study, and hyperparameter study.

5.1 Dataset Description

BlogCatalog (BC). BlogCatalog [47] is a social blog directory website, which contains a friendship network between bloggers. In this dataset, each unit is a blogger, and each edge denotes the friendship between two bloggers. The features are the bag-of-words representation of

the keywords of the blogs posted by bloggers. We follow the assumptions of References [19, 49] to synthesize treatment assignment and potential outcome: (a) treatment (blogger's blogs are viewed more on mobile devices or desktops) and (b) outcome (the opinion of readers about the bloggers). Additionally, we follow Reference [19] by assuming a blogger and her/his neighbors' topics can causally affect her treatment assignment and outcome. More specifically, we train the topic distribution $r(\boldsymbol{x}_i)$ for each blogger using the LDA topic model [6]. Then we randomly select a blogger's topic distribution as the centroid of the treatment group and control group, which are denoted as $\boldsymbol{r}_1^c$ and $\boldsymbol{r}_0^c$, respectively. We then model the readers' browsing device preferences (i.e., propensity score) on the ith blogger's contents as follows:

$$\begin{aligned} P(t_i = 1|\boldsymbol{x}_i, \boldsymbol{A}) &= \frac{\exp(p_1^i)}{\exp(p_1^i) + \exp(p_0^i)} \\ p_1^i &= k_1 r(\boldsymbol{x}_i)^T \boldsymbol{r}_1^c + k_2 \sum_{j \in \mathcal{N}(i)} r(\boldsymbol{x}_j) \boldsymbol{r}_1^c \\ p_0^i &= k_1 r(\boldsymbol{x}_i)^T \boldsymbol{r}_0^c + k_2 \sum_{j \in \mathcal{N}(i)} r(\boldsymbol{x}_j) \boldsymbol{r}_0^c, \end{aligned} \tag{11}$$

where k_1 and k_2 control the magnitude of the confounding bias, which are determined by a blogger's topic and her neighbors' topics. Then the factual outcome and counterfactual outcome of the ith unit with spillover effect considered are simulated as

$$\begin{aligned} y_i^F &= C(p_0^i + t_i p_1^i) + \sum_{j \in \mathcal{N}(i)} \boldsymbol{S}_{ij} P(t_j = 1|\boldsymbol{x}_j, \boldsymbol{A}) + \epsilon \\ y_i^{CF} &= C(p_0^i + (1 - t_i) p_1^i) + \sum_{j \in \mathcal{N}(i)} \boldsymbol{S}_{ij} P(t_j = 1|\boldsymbol{x}_j, \boldsymbol{A}) + \epsilon, \end{aligned} \tag{12}$$

where C is the scaling weight and $\boldsymbol{S}_{ij}$ is the magnitude of spillover effect of the treatment of the j-unit on the outcome of the ith unit, which are sampled from a uniform distribution: $\boldsymbol{S}_{ij} \sim U(0, s)$. Here, we adjust the value of s to control the magnitude of spillover effect in the network; the larger the value of s, the stronger the interference in the network. ϵ is the noise term sampled as $\epsilon \sim \mathcal{N}(0, 1)$.

Flickr. Flickr [47] is an online community where users can share images and videos and follow each other. In this dataset, each unit is a user and each edge represents the friendship between two users. The features of a unit is the set of its interest tags. We also study the causal effect of viewing images and videos on mobile devices or desktops (treatment) on readers' opinions of the user (outcome). Here, we follow the same setting and assumptions of BlogCatalog to synthesize the treatments and outcomes for Flickr.

PeerRead. PeerRead [27] is a dataset of computer scientific peer reviews for papers. Each unit in the dataset denotes an author, the edges stand for the co-author relationships between authors. The features of each unit are the bag-of-word representations extracted from their paper title and abstract. The treatment of a unit in the dataset is whether the unit's (author's) papers contain some keywords, and the outcome denotes the citation number of their papers. We follow the same setting as BlogCatalog and Flickr to synthesize the treatments and outcomes for each unit.

The detailed statistics of these three semi-synthetic datasets are shown in Table 1. We conduct the procedure of the data simulation over 10 runs for each parameter setting. In addition to some conventional statistical information, we also report the average number of treated units in each setting for each dataset and the network density for each dataset by $\eta = \frac{2|E|}{|V|(|V|-1)}$, where $|V|$ is the number of vertexes and $|E|$ is the number of edges in the graph.

Table 1. Detailed Statistics of the Used Three Semi-synthetic Datasets BlogCatalog (BC), Flickr, and PeerRead

Datasets	# Instances	# Edges	# Features	η	s	# Treated	ATE Mean	STD
BC	5,196	173,468	8,189	1.2×10^{-2}	0.2	2379.2	17.768	3.587
					0.5	2633.8	24.250	3.989
					0.8	2122.4	28.434	6.294
Flickr	7,575	239,738	12,047	8.3×10^{-3}	0.2	3839.8	9.135	1.962
					0.5	3611.7	7.988	1.791
					0.8	3683.8	8.387	1.960
PeerRead	7,601	13,691	1,080	4.7×10^{-4}	0.2	3663.1	2.558	0.622
					0.5	3626.5	2.080	0.714
					0.8	3717.4	2.463	0.882

The column s denotes the magnitude of spillover effect of the network. η denotes the network density. ATE Mean stands for the mean of average treatment effect, and STD stands for the standard deviation of ATE over 10 different simulation runs.

5.2 Evaluation Metrics

We adopt two widely used metrics in causal inference to evaluate the effectiveness of our proposed framework, *SPNet*, for learning individual treatment effect: (1) For measuring the accuracy of the unit-level treatment effect, we report Rooted Precision in Estimation of Heterogeneous Effect ($\sqrt{\epsilon_{PEHE}}$), and (2) for the population-level treatment effect we report Mean Absolute Error on **average treatment effect (ATE)** (ϵ_{ATE}) [20, 32]. The definition of the two metrics is as follows:

$$\begin{aligned} \sqrt{\epsilon_{PEHE}} &= \sqrt{\frac{1}{n}\sum_{i=1}(\tau_i - \hat{\tau}_i)^2} \\ \epsilon_{ATE} &= \left|\frac{1}{n}\sum_{i=1}\hat{\tau}_i - \frac{1}{n}\sum_{i=1}\tau_i\right|, \end{aligned} \tag{13}$$

where $\tau_i = y_i^{t_i=1} - y_i^{t_i=0}$ and $\hat{\tau}_i = \hat{y}_i^{t_i=1} - \hat{y}_i^{t_i=0}$ are the ground truth ITE and the inferred ITE, respectively. Lower values of them denote better estimation performance.

5.3 Experimental Settings

We compare our proposed framework, *SPNet*, with the following state-of-the-art models for ITE estimation.

Counterfactual Regression (CFR). *CFR* [42] is a representation learning–based method to map the original features to latent space to capture hidden confounders by minimizing the error of inferred factual outcomes and the imbalance between latent representations of treatment group and control group. Here we consider two types of balancing penalties: Wasserstein-1 distance (CFR-Wass) and the maximum mean discrepancy (CFR-MMD).

Treatment-agnostic Representation Networks (TARNet). *TARNet* [42] is a variant of the CFR framework, which does not have a built-in representation balancing component.

Causal Effect Variational Autoencoder (CEVAE). *CEVAE* [32] is based on the Variational Autoencoders [29]. It follows the causal structure of inference with proxies and builds deep latent variable model to estimate the unknown latent space summarizing the confounders and the causal effect.

Causal Forest. *Causal Forest* [50] is a nonparametric causal inference method that extends the Breiman's random forest algorithm [7] for estimating heterogeneous treatment effects in subgroups with the strong ignorability assumption.

Bayesian Additive Regression Trees (BART). *BART* [20] is a widely used Bayesian nonparametric modeling procedure for causal inference, which is also based on the strong ignorability assumption.

Network Deconfounder (NetDeconf). *NetDeconf* [19] is a framework that aims to exploit the network information to control the confounding bias to learn ITEs from networked observational data. Specifically, it relaxes the strong Ignorability assumption and captures the hidden confounders using graph convolutional network in latent space. NetDeconf is different from the above methods as it leverages network structure to infer the confounders.

Linked Causal Variational Autoencoder (LCVA). *LCVA* [38] is a generative model using variational autoencoder architecture. It incorporates an encoder neural network to learn the latent attributes and a decoder network to reconstruct the inputs and then capture the spillover effect between units for ITE estimation.

Heterogeneous Partial Interference (HGPI). *HGPI* [37] partition the units into a large number of disjoint clusters based on observables and develop an **augmented inverse propensity weighted (AIPW)** estimator for estimating heterogeneous treatment and spillover effects under conditional exchangeability.

Flame-Network. *Flame-Network* [3] is a causal estimator based on a matching method and estimates the average treatment effect using the FLAME [51] algorithm. The method matches units almost exactly on counts of the subgraphs for each unit's neighborhood graph and captures the network interference by utilizing the subgraph information.

For each dataset, we run the experiments 10 times and the average performance is reported. For each run, we split the dataset into training set (60%), validation set (20%), and test set (20%). Regarding the hyperparameters of the proposed framework, we utilize the grid search strategy to find the optimal hyperparameters combination based on the results on the validation set. More specifically, we set the learning rate as 0.01 and the dimension of the representation space d as 400. We also vary the number of GCN layers and hidden layers of fully connected neural networks in {1,2,3}; α, β, γ range in $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. For the simulation procedure, we set $C = 5$, $k_1 = 10$, and $k_2 = 1$, s varies from $\{0.2, 0.5, 0.8\}$ to control the magnitude of network interference. The optimizer Adam [28] is adopted to train the model.

5.4 Performance of ITE Estimation

In this subsection, we compare the proposed framework *SPNet* against the aforementioned baseline methods on the task of ITE estimation. The experimental results on ITE estimation task are shown in Table 2. Noting that the causal estimand of interest for HGPI and Flame-networks is the average treatment effect over population, thus the results in terms of $\sqrt{\epsilon_{PEHE}}$ for these two models are not available. By analyzing the experimental results, we have the following observations:

- The proposed ITE estimation framework *SPNet* clearly outperforms traditional baseline methods, including CFR, CEVAE, Causal Forest, and BART, which ignore the network information for learning ITEs.
- *SPNet* and NetDeconf have better performance than the other baseline methods for estimating ITEs. The main reason is that *SPNet* and NetDeconf consider the auxiliary network information to capture the influence of hidden confounders toward unbiased ITE

Table 2. ITE Estimation Performance Comparison for Different Methods on BlogCatalog, Flickr, and PeerRead (Mean ±Std)

BlogCatalog						
s	0.2		0.5		0.8	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
CFR-Wass	24.047 ± 4.648	17.502 ± 3.353	29.983 ± 5.347	20.245 ± 3.889	31.109 ± 7.991	22.098 ± 5.980
CFR-MMD	22.534 ± 4.327	15.838 ± 3.114	30.276 ± 5.298	20.303 ± 3.762	31.313 ± 7.618	21.734 ± 5.784
TARNet	23.033 ± 4.639	16.818 ± 3.423	30.456 ± 5.176	21.305 ± 3.659	32.347 ± 7.818	22.547 ± 5.982
CEVAE	18.743 ± 3.812	8.559 ± 2.978	24.942 ± 4.379	11.726 ± 3.121	29.812 ± 6.348	17.196 ± 4.215
Causal Forest	15.943 ± 2.675	3.277 ± 1.817	21.267 ± 2.913	5.568 ± 1.681	26.175 ± 4.587	12.916 ± 2.786
BART	12.150 ± 1.934	7.480 ± 2.432	14.457 ± 1.896	8.452 ± 2.312	18.131 ± 1.771	12.725 ± 2.578
LCVA	22.307 ± 5.731	6.212 ± 2.128	23.086 ± 6.107	7.435 ± 2.764	27.314 ± 5.875	10.315 ± 2.302
HGPI	–	3.740 ± 2.627	–	6.413 ± 3.637	–	14.639 ± 8.521
Flame-Network	–	4.141 ± 3.197	–	6.192 ± 4.096	–	9.631 ± 5.901
NetDeconf	6.927 ± 0.927	2.551 ± 0.325	9.094 ± 0.892	4.140 ± 0.501	9.534 ± 1.031	4.204 ± 0.674
SPNet (ours)	**3.491 ± 0.351**	**1.122 ± 0.256**	**6.122 ± 0.377**	**1.677 ± 0.243**	**6.958 ± 0.504**	**2.448 ± 0.247**
Flickr						
s	0.2		0.5		0.8	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
CFR-Wass	16.743 ± 3.996	3.612 ± 1.329	15.512 ± 3.510	4.121 ± 0.994	17.390 ± 3.806	6.347 ± 1.125
CFR-MMD	17.438 ± 4.127	3.594 ± 1.237	15.341 ± 3.469	4.329 ± 1.118	17.654 ± 3.764	6.502 ± 1.092
TARNet	17.414 ± 4.358	3.682 ± 1.223	15.647 ± 3.585	5.788 ± 1.201	16.737 ± 3.978	6.172 ± 0.972
CEVAE	16.638 ± 4.215	5.919 ± 1.078	14.277 ± 3.419	4.396 ± 0.978	17.276 ± 3.917	8.551 ± 1.341
Causal Forest	18.791 ± 3.975	7.508 ± 0.918	16.304 ± 3.364	6.025 ± 1.024	19.620 ± 3.875	10.927 ± 1.242
BART	10.324 ± 2.186	5.748 ± 0.863	9.601 ± 1.974	4.840 ± 0.737	14.285 ± 2.216	8.343 ± 1.257
LCVA	21.023 ± 4.903	6.950 ± 1.521	18.919 ± 3.765	6.576 ± 1.319	22.165 ± 4.844	8.907 ± 1.428
HGPI	–	7.076 ± 2.630	–	4.952 ± 2.429	–	10.316 ± 3.013
Flame-Network	–	5.351 ± 2.850	–	3.328 ± 1.917	–	8.636 ± 4.222
NetDeconf	6.971 ± 0.833	0.930 ± 0.385	7.051 ± 0.869	0.910 ± 0.418	11.461 ± 1.223	1.791 ± 0.503
SPNet (ours)	**5.250 ± 0.511**	**0.717 ± 0.103**	**5.562 ± 0.537**	**0.860 ± 0.127**	**7.562 ± 0.487**	**1.237 ± 0.187**
PeerRead						
s	0.2		0.5		0.8	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
CFR-Wass	4.399 ± 1.560	2.064 ± 0.717	2.862 ± 1.317	1.270 ± 0.697	3.619 ± 1.406	1.636 ± 0.708
CFR-MMD	4.402 ± 1.459	2.065 ± 0.873	2.683 ± 1.262	1.152 ± 0.671	3.406 ± 1.396	1.524 ± 0.712
TARNet	4.401 ± 1.332	2.063 ± 0.799	2.707 ± 1.307	1.252 ± 0.665	3.601 ± 1.519	1.576 ± 0.698
CEVAE	4.339 ± 1.178	1.864 ± 0.745	2.831 ± 1.203	1.397 ± 0.697	3.430 ± 1.318	1.530 ± 0.710
Causal Forest	3.578 ± 0.448	0.660 ± 0.353	2.255 ± 0.371	0.462 ± 0.289	2.935 ± 0.496	0.567 ± 0.332
BART	5.169 ± 0.892	1.972 ± 0.531	2.736 ± 0.671	1.202 ± 0.473	3.539 ± 0.745	1.448 ± 0.519
LCVA	3.417 ± 0.818	1.025 ± 0.675	2.948 ± 0.935	0.805 ± 0.593	3.124 ± 0.741	0.953 ± 0.507
HGPI	–	1.017 ± 0.217	–	0.898 ± 0.180	–	1.039 ± 0.208
Flame-Network	–	1.791 ± 0.664	–	1.347 ± 0.901	–	1.533 ± 1.179
NetDeconf	3.900 ± 0.271	1.084 ± 0.209	2.355 ± 0.236	0.511 ± 0.157	3.045 ± 0.197	0.788 ± 0.128
SPNet (ours)	**3.355 ± 0.213**	**0.626 ± 0.118**	**2.090 ± 0.168**	**0.225 ± 0.091**	**2.698 ± 0.121**	**0.314 ± 0.102**

Bold indicates the best performance, "–" means not available result.

estimation. This observation demonstrates that the auxiliary network information is helpful for controlling the confounding bias in causal inference.

- *SPNet* is also superior to NetDeconf in terms of ITE estimation, because it also models the inherent interference between different units while NetDeconf only leverages network information for confounder representation learning.
- Although *LCVA* accounts for the interference when estimating ITE, it does not render satisfactory performance, especially on BlogCatalog and Flickr due to the following reasons: First, LCVA assumes that the spillover effect between different neighboring units are the same,

which is untenable in reality. Second, LCVA is effective only when the dimension of covariates and edge density of network are low (e.g., PeerRead) due to its weak anti-noise ability.

- HGPI and Flame-Network does not perform well on the task of estimating treatment effect although they take into account interference, here are the reasons: First, the basic estimators of these two methods (i.e., IPW, Matching-based estimator) cannot capture the complex interactions between units in the networked environment. Second, HGPI assumes that interference is restricted to be only among units within the same cluster, that ignores the interference between different clusters; Flame-Network rely on the counts of the subgraph to capture the interference and cannot capture the different magnitude of spillover effects between different node pairs. Third, HGPI and Flame-Network are not able to handle the confounding bias resulting from hidden confounders in observational data.
- The proposed framework *SPNet* suffers the least in terms of $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} when the interference of network grows with the increase of the value of s, because *SPNet* can model and quantify the magnitude of the interference and infer ITEs more precisely.

Additional discussions. As we can see in Table 2, the performance improvement of the proposed method SPNet on the PeerRead dataset is not as significant as the improvement on the other two datasets, BlogCatalog and Flickr in general. We analyze that this is caused by the property of the PeerRead dataset with the following aspects: First, we can see that in Table 1 the dimension of covariates in PeerRead is much lower than that in BlogCatalog and Flickr, which means the mapping relationship between inputs (covariates) and outputs (potential outcomes) is easier to capture when estimating causal effect even if the estimator is simple. Second, the ATE in PeerRead is much lower than that in BlogCatalog and Flickr, which means that for the difference between the true causal effect and the causal effect estimated by the estimator, it is easier to present a smaller difference on PeerRead, compared to the other two datasets. Last but most importantly, we can see that the network density of PeerRead is significantly lower than that of BlogCatalog and Flickr (by one or two orders of magnitude), which means that the influence of network structural information on the causal estimands of interest as well as the network interference on PeerRead is much smaller than that on the other two datasets. Anyway, our proposed model can still be superior over the baseline methods even on PeerRead dataset.

5.5 Robustness of *SPNet*

Individual treatment effect estimation is frequently used in decision-making across a wide range of high-impact domains, including health care, economics, and so on. As a result, using a robust estimator against noise is critical; otherwise, incorrect decisions will be made as a result of an imprecise ITE estimator caused by noise, leading to serious consequences.

The most common way to introduce noise into the networked data to change the prediction of a model is to disrupt the network structure, i.e., the adjacent matrix. Here we investigate the robustness of the *SPNet* model by perturbing the adjacency matrix of the networked observational data. To be specific, we poison the network structure through the following two modes: (1) Add (randomly adding edges into the clean graph) and (2) Flip (randomly flipping the edges in the clean graph; flipping means to add an edge between two nodes that are not originally connected or to remove an edge between two nodes that are originally connected).

We present the results of *SPNet* and *NetDeconf* on BlogCatalog ($s = 0.2$) under different perturbation mode and rate in terms of $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}, we omit *LCVA*, which also takes graph structure into account due to its poor performance. And we have similar results on other datasets. We set the perturbation rate as {10%, 20%,30%}, and the experimental results are shown in Table 3. One can see that the performance of the proposed model *SPNet* is stable, and it does not suffer

Table 3. The Results (Mean ± Std) of *SPNet* and NetDeconf on BlogCatalog ($s = 0.2$) under Different Perturbation Modes and Rates (%)

SPNet (Ours)				
Mode / Rate (%)	Add		Flip	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
10	3.631 ± 0.402	1.102 ± 0.283	3.724 ± 0.397	1.220 ± 0.375
20	3.945 ± 0.521	1.375 ± 0.328	3.968 ± 0.445	1.201 ± 0.381
30	4.308 ± 0.693	1.718 ± 0.491	4.231 ± 0.724	1.616 ± 0.478
NetDeconf				
Mode / Rate (%)	Add		Flip	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
10	7.707 ± 1.011	2.871 ± 0.436	7.848 ± 0.901	2.927 ± 0.518
20	7.981 ± 1.094	3.171 ± 0.522	7.880 ± 0.923	3.212 ± 0.578
30	8.757 ± 1.309	4.152 ± 0.698	8.592 ± 1.287	5.050 ± 0.724

much from the graph noise, and we have the following observations: (1) When the perturbation rate is small (e.g., 10%), the performance of *SPNet* is close to that under clean graph; (2) the proposed model *SPNet* suffers least from the graph noise with the increasing of perturbation rate; and (3) even the perturbation rate is relatively high (e.g., 30%), the performance of *SPNet* is competitive and still better than other state-of-the-art models for ITE estimation (see Table 2). The above observations fully demonstrate the robustness of the proposed model against noise, which is crucial in practical applications.

5.6 Ablation Study

We further analyze the impact of different components in the proposed framework *SPNet* for learning ITE from networked observational data. Specifically, we conduct ablation study by deriving the following three variants of *SPNet* and compare their performance with the original *SPNet*:

(i) ***SPNet w/o Interference Modeling*:** This variant does not model the spillover effect between different units, which means that it only uses the combination of two partial representations to infer potential outcome without utilizing network information to model interference between neighboring units. Here, we use *SPNet w/o IM* to denote this variant.

(ii) ***SPNet w/o Masked Attention*:** This variant omits the masked-attention mechanism, which means that the different magnitude of spillover effect between different neighboring units cannot be quantified. It assumes that the magnitude of interference is the same everywhere in the network. Here, we use *SPNet w/o MA* to denote this variant.

(iii) ***SPNet w/o Treatment Prediction*:** This variant does not add the observed treatment loss $\mathcal{L}_t$ to train the representations. Here, we use *SPNet w/o TP* to denote this variant.

The comparison results of the three variants and the original *SPNet* are shown in Figure 2. By analysing the results, we can have the following observations:

- *SPNet w/o IM* does not achieve the desired performance and generally performs the worst, as it ignores the interference between neighboring units, which demonstrates that modeling interference is crucial for ITE estimation from networked observational data.
- *SPNet w/o MA* also does not render satisfactory performance, as it cannot model the different magnitude of spillover effects between different neighboring units. This implies that modeling interference is beneficial for ITE estimation, but treating the magnitude of spillover effects between different pairs of units as identical is erroneous.

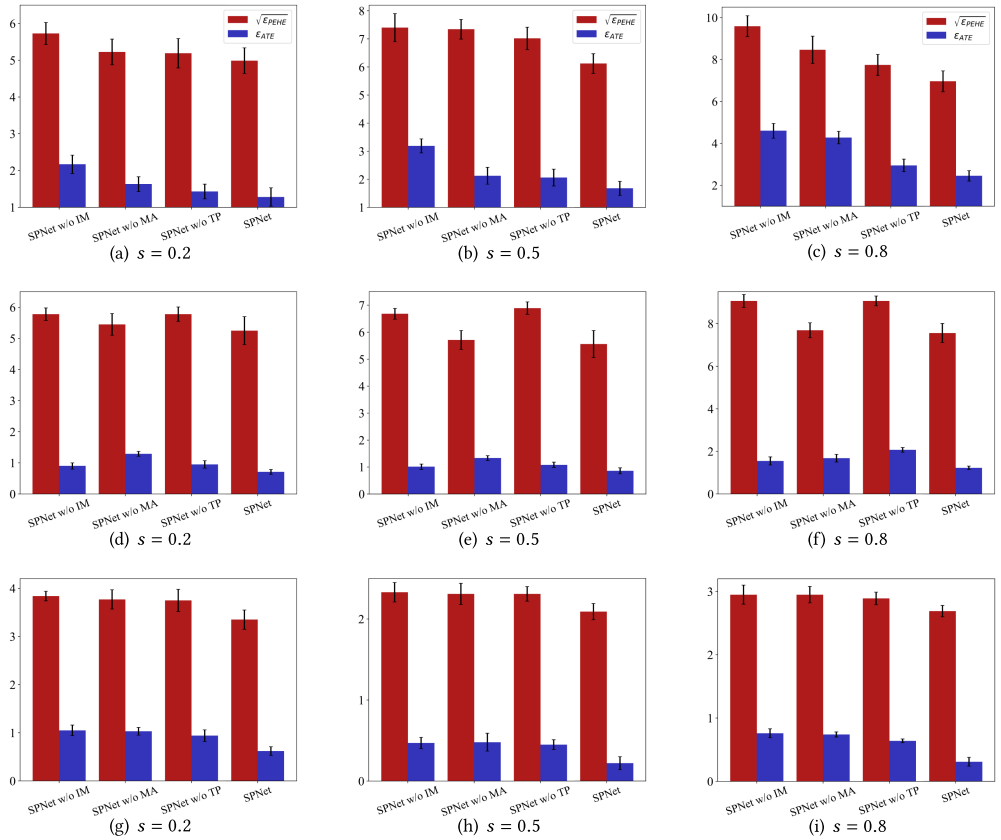

Fig. 2. Ablation study of *SPNet*. ((a)–(c)) For BlogCatalog, ((d)–(f)) for Flickr, and ((g)–(i)) for PeerRead.

- *SPNet w/o TP* also performs worse than the original model *SPNet*, which demonstrates the necessity of treatment prediction for learning better representations of hidden confounders for ITE estimation.

5.7 Hyperparameter Study

We further investigate the impact of the two critical hyperparameters α and β, which control the contribution of treatment prediction and representation balancing for ITE estimation on the networked observational data. We report the parameter analysis result on BlogCatalog in terms of $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}. We vary these two hyperparameters in the range of {0.001,0.01, 0.1, 1, 10}, and the hyperparameter study results are shown in Figure 3. Generally speaking, *SPNet* is not sensitive to these two parameters, as its performance is stable when the two hyperparameters vary in a wide range. When α and β range in {0.001, 0.01, 0.1}, the performance is relatively better. And we have similar experimental results and observations on other datasets.

6 RELATED WORK

In this section, we review related works from two aspects: (1) causal inference for i.i.d. observational data and (2) causal inference for network environment.

Causal inference for i.i.d. observational data. Due to the expensive costs and potential ethical issues of randomized experiments, causal inference from observational data has attracted a

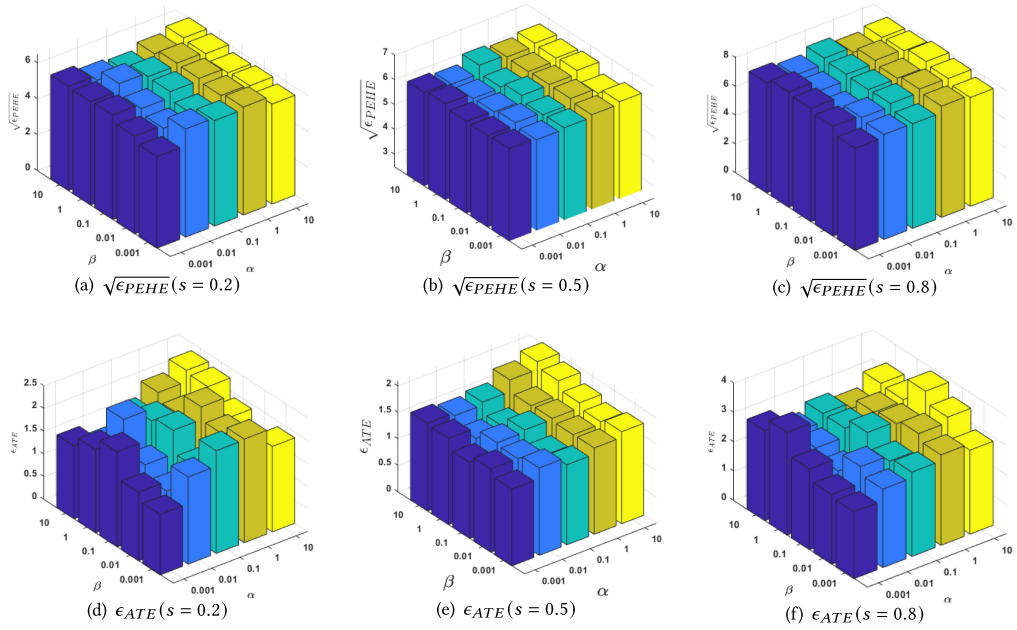

Fig. 3. Hyperparameter analysis on BlogCatalog.

surge of interests in recent years, and most of existing studies focus on i.i.d. data. BART [20] applies Bayesian nonparametric additive regression trees to infer the potential outcomes for estimating ITEs from observational data. CFR [42] casts counterfactual inference as a type of domain adaptation problem and learns the ITE using neural network to learn balanced representations by minimizing the distribution difference of control group and treated gruop. Causal Forest [50] utilizes random forest to estimate heterogeneous treatment effects in subgroups. The methods mentioned above rely on the strong ignorability assumption that essentially ignores the influence of hidden confounders. By relaxing the strong ignorability assumption, CEVAE [32] proposes to map the original observed features to latent space to capture the hidden confounders based on variational autoencoder [29]. Atan et al. [2] propose Deep-Treat to reduce the bias by learning representations and constructing effective treatment policies using deep neural networks on the transformed data for causal effects estimation. Yao et al. [54] propose a local similarity preserved individual treatment effect estimation method based on deep representation learning, which can capture the hidden confounders and preserve local similarity of data. Kallus et. al [26] develop a functional interval estimator that predicts bounds on the ITEs by taking the form of a weighted kernel estimator whose weights vary adversarially.

Causal inference for network environment. With the prevalence of networked observational data in many real-world scenarios, leveraging network information among different units to provide better causal effects estimation has emerged as an important subarea in causal inference. For example, NetDeconf [19] leverages GCNs to map the features of units and their network structure into an embedding space to better control the influence of hidden confounders for ITE estimation. CONE [18] further utilizes graph attention network for counterfactual evaluation of treatment assignment functions in networked observational data. IGNITE [17] infers ITE from networked observational data by balancing distributions of confounder representations and predicting treatment assignment based on a minimax strategy. Ma et al. [33] propose a DNDC framework to estimate the ITE in a dynamic environment by utilizing the network information as well as

the temporal dependencies. However, these methods do not account for network interference. LCVA [38] adopts variational auto-encoder to model the interference for ITE estimation from networked observational data. However, it does not identify the different magnitudes of spillover effects between different neighboring units and could be ineffective when applied to large-scale networks. In addition, *CauseIS* [13] and *CMatch* [12] are proposed to design network experiments to minimize interference and infer unbiased causal effect, while these works are mainly for the analysis of experimental design. In Reference [37], the authors proposed to create disjoint clusters for the units, which are based on the observables of units, where an AIPW estimator is proposed to estimate treatment and spillover effects under conditional exchangeability for the heterogeneous groups. *Flame-Network* [3] is a matching-based causal estimator that calculates the average treatment effect. It achieves precise matching of units based on the subgraph counts of each unit's neighborhood graph and leverages this subgraph information to account for network interference. However, the above two methods ignore the different magnitudes of interference between different node pairs and the confounding bias. The model proposed in Reference [15] is a propensity score-based causal effect estimation method, in which the method aggregates the intervention assignment of neighbors through a function and forms a joint treatment with each unit's own intervention, such that the method is prone to fail when the covariate dimension is high, because it is only a one-dimensional scalar variable that determines the interference, and the method also fails to address the confounding bias caused by hidden confounders. In this work, the model SPNet we proposed assumes first-order full interference (i.e., spillover effects may exist between any neighboring nodes) and quantifies the magnitude of interference between different node pairs, which is more realistic. And we adopt graph neural network model, effectively leveraging the node covariates and auxiliary graph structure information to capture hidden confounders and provide more unbiased, accurate causal effect estimates.

7 CONCLUSION AND FUTURE WORK

This article aims to learn individual treatment effect from networked observational data via modeling network interference. The proposed model, *SPNet*, utilizes observed features and auxiliary network information to control confounding bias and develops a twin-channel graph convolutional network to map original features into partial representations for modelling the interference along with a masked attention mechanism. Empirically, extensive experiments on multiple semi-synthetic datasets demonstrate the superiority of *SPNet* over the existing state-of-the-art methods in learning ITEs on networked observational data.

For future work, we can focus on two interesting aspects. First, most of the existing work for learning individual treatment effect is for homogeneous network, and thus we would like to investigate how to model the interference for ITE estimation from heterogeneous network [8, 45], which contains multiple types of nodes and edges. Second, another noteworthy research direction is to model interference for learning ITE from signed networks [22, 46], which contain positive and negative edges between nodes, the modeling form would be different due to the existence of negative links comparing with the unsigned networks.

REFERENCES

[1] Peter M. Aronow, Cyrus Samii, et al. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* 11, 4 (2017), 1912–1947.

[2] Onur Atan, James Jordon, and Mihaela Van der Schaar. 2018. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[3] Usaid Awan, Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. 2020. Almost-matching-exactly for treatment effect estimation under network interference. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3252–3262.

[4] Guillaume Basse and Avi Feller. 2018. Analyzing two-stage experiments in the presence of interference. *J. Am. Statist. Assoc.* 113, 521 (2018), 41–55.
[5] Howard Bauchner, Robert Vinci, Sharon Bak, Colleen Pearson, and Michael J. Corwin. 1996. Parents and procedures: A randomized controlled trial. *Pediatrics* 98, 5 (1996), 861–867.
[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3, Jan (2003), 993–1022.
[7] Leo Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.
[8] Shiyu Chang, Wei Han, Jiliang Tang, Guo-Jun Qi, Charu C. Aggarwal, and Thomas S. Huang. 2015. Heterogeneous network embedding via deep architectures. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. 119–128.
[9] Zhixuan Chu, Stephen L. Rathbun, and Sheng Li. 2021. Graph infomax adversarial learning for treatment effect estimation with networked observational data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 176–184.
[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in Neural Information Processing Systems* 29, (2016), 3837–3845.
[11] Johannes A. N. Dorresteijn, Frank L. J. Visseren, Paul M. Ridker, Annemarie M. J. Wassink, Nina P. Paynter, Ewout W. Steyerberg, Yolanda van der Graaf, and Nancy R. Cook. 2011. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *Br. Med. J.* 343 (2011).
[12] Zahra Fatemi and Elena Zheleva. 2020. Minimizing interference and selection bias in network experiment design. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 176–186.
[13] Zahra Fatemi and Elena Zheleva. 2020. Network experiment design for estimating direct treatment effects. In *Proceedings of the KDD Workshop on Mining and Learning with Graphs (MLG'20)*, Vol. 8.
[14] Amy Finkelstein, Annetta Zhou, Sarah Taubman, and Joseph Doyle. 2020. Health care hotspottingâĂŤa randomized, controlled trial. *New Engl. J. Med.* 382, 2 (2020), 152–162.
[15] Laura Forastiere, Edoardo M. Airoldi, and Fabrizia Mealli. 2021. Identification and estimation of treatment and interference effects in observational studies on networks. *J. Am. Statist. Assoc.* 116, 534 (2021), 901–918.
[16] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Comput. Surv.* 53, 4 (2020), 1–37.
[17] Ruocheng Guo, Jundong Li, Yichuan Li, K. Selçuk Candan, Adrienne Raglin, and Huan Liu. 2021. Ignite: A minimax game toward learning individual treatment effects from networked observational data. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence*. 4534–4540.
[18] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Counterfactual evaluation of treatment assignment functions with networked observational data. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 271–279.
[19] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 232–240.
[20] Jennifer L. Hill. 2011. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* 20, 1 (2011), 217–240.
[21] Guanglei Hong. 2015. *Causality in a Social World: Moderation, Mediation and Spill-over*. John Wiley&Sons.
[22] Qiang Huang, Tingyu Xia, Huiyan Sun, Makoto Yamada, and Yi Chang. 2020. Unsupervised nonlinear feature selection from high-dimensional signed networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4182–4189.
[23] Michael G. Hudgens and M. Elizabeth Halloran. 2008. Toward causal inference with interference. *J. Am. Statist. Assoc.* 103, 482 (2008), 832–842.
[24] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
[25] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, PMLR, 3020–3029.
[26] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2019. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2281–2290.
[27] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 1647–1661.
[28] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, Yoshua Bengio and Yann LeCun (Eds.).

[29] Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Prceedings of the 2nd International Conference on Learning Representations (ICLR'14)*, Yoshua Bengio and Yann LeCun (Eds.).
[30] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations (ICLR'17)*, Toulon, France, April 24-26, 2017, Conference Track Proceedings.
[31] Lan Liu and Michael G. Hudgens. 2014. Large sample randomization inference of causal effects in the presence of interference. *J. Am. Statist. Assoc.* 109, 505 (2014), 288–301.
[32] Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems* 30, (2017), 6446–6456.
[33] Jing Ma, Ruocheng Guo, Chen Chen, Aidong Zhang, and Jundong Li. 2021. Deconfounding with networked observational data in a dynamic environment. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 166–174.
[34] Yunpu Ma and Volker Tresp. 2021. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3700–3708.
[35] Yunpu Ma, Yuyi Wang, and Volker Tresp. 2020. Causal inference under networked interference. arXiv:2002.08506. Retrieved from https://arxiv.org/abs/2002.08506
[36] Elizabeth L. Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J. Van der Laan. 2022. Causal inference for social network data. *Journal of the American Statistical Association* (2022), 1–15.
[37] Zhaonan Qu, Ruoxuan Xiong, Jizhou Liu, and Guido Imbens. 2021. Efficient treatment effect estimation in observational studies under heterogeneous partial interference. arXiv:2107.12420. Retrieved from https://arxiv.org/abs/2107.12420
[38] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1679–1682.
[39] Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5 (1974), 688.
[40] Donald B. Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Am. Statist. Assoc.* 100, 469 (2005), 322–331.
[41] Usman Shahid and Elena Zheleva. 2019. Counterfactual learning in networks: an empirical study of model dependence. In *Beyond Curve Fitting: Causation, Counterfactuals, and Imagination-based AI. AAAI Spring Symposium (AAAI-WHY 2019)*, Standford, CA., Association for the Advancement of Artificial Intelligence.
[42] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. 3076–3085.
[43] Cosma Rohilla Shalizi and Edward McFowland III. 2023. Estimating causal peer influence in homophilous social networks by inferring latent locations. *Journal of the American Statistical Association* 118, 541 (2023), 707–718.
[44] Cosma Rohilla Shalizi and Andrew C. Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* 40, 2 (2011), 211–239.
[45] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2016. A survey of heterogeneous information network analysis. *IEEE Trans. Knowl. Data Eng.* 29, 1 (2016), 17–37.
[46] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. A survey of signed network mining in social media. *Comput. Surv.* 49, 3 (2016), 1–37.
[47] Lei Tang and Huan Liu. 2011. Leveraging social media networks for classification. *Data Min. Knowl. Discov.* 23, 3 (2011), 447–478.
[48] Panos Toulis, Alexander Volfovsky, and Edoardo M. Airoldi. 2018. Propensity score methodology in the presence of network entanglement between treatments. arXiv:1801.07310. Retrieved from https://arxiv.org/abs/1801.07310
[49] Victor Veitch, Yixin Wang, and David Blei. 2019. Using embeddings to correct for unobserved confounding in networks. *Adv. Neural Inf. Process. Syst.* 32 (2019).
[50] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[51] Tianyu Wang, Marco Morucci, M Usaid Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. 2021. Flame: A fast large-scale almost matching exactly approach to causal inference. *J. Mach. Learn. Res.* 22, 1 (2021), 1477–1517.
[52] Christopher Winship and Stephen L. Morgan. 1999. The estimation of causal effects from observational data. *Annu. Rev. Sociol.* 25, 1 (1999), 659–706.
[53] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.

[54] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. *Adv. Neural Inf. Process. Syst.* 31 (2018).
[55] Bo-Heng Zhang, Bing-Hui Yang, and Zhao-You Tang. 2004. Randomized controlled trial of screening for hepatocellular carcinoma. *J. Cancer Res. Clin. Oncol.* 130, 7 (2004), 417–422.

Received 12 September 2022; revised 29 March 2023; accepted 26 September 2023