IGNITE: A Minimax Game Toward Learning Individual Treatment Effects from Networked Observational Data

Ruocheng Guo[1] , **Jundong Li**[2,3] , **Yichuan Li**[1] ,
K. Selçuk Candan[1] , **Adrienne Raglin**[4] and **Huan Liu**[1]
[1]Computer Science and Engineering, Arizona State University, USA
[2]Electrical and Computer Engineering, University of Virginia, USA
[3]Computer Science & School of Data Science, University of Virginia, USA
[4]Army Research Laboratory , USA
rguo12@asu.edu, jundong@virginia.edu, yichuan1@asu.edu, candan@asu.edu,
adrienne.raglin2.civ@mail.mil, huan.liu@asu.edu

Abstract

Networked observational data presents new opportunities for learning individual causal effects, which plays an indispensable role in decision making. Such data poses the challenge of confounding bias. Previous work presents two desiderata to handle confounding bias. On the treatment group level, we aim to balance the distributions of confounder representations. On the individual level, it is desirable to capture patterns of hidden confounders that predict treatment assignments. Existing methods show the potential of utilizing network information to handle confounding bias, but they only try to satisfy one of the two desiderata. This is because the two desiderata seem to contradict each other. When the two distributions of confounder representations are highly overlapped, then we confront the undiscriminating problem between the treated and the controlled. In this work, we formulate the two desiderata as a minimax game. We propose IGNITE that learns representations of confounders from networked observational data, which is trained by a minimax game to achieve the two desiderata. Experiments verify the efficacy of IGNITE on two datasets under various settings.

1 Introduction

Networked observational data grants us a new source of learning individual treatment effects (ITEs), which plays a crucial role in rational decision making across a myriad of influential fintech related applications (e.g., economics, marketing, and advertising etc.). For example, in a social network with blog service, a blogger who aims to attract readers to adopt fintech products (e.g., investment apps like Robinhood[1] or online payment platforms like Alipay[2]) may want to determine which browsing device is more suitable to promote her articles. This requires us to learn the causal effect of browsing devices (treatments) on the number of readers who adopt fintech products after reading her articles (outcomes).

Learning ITEs from networked observational data requires controlling confounding bias. Confounding bias is the influence of confounders – the variables causally influencing treatment assignments and outcomes simultaneously. However, these confounders are extremely difficult to measure as they will induce the confounding bias even if observed features have been properly adjusted for [Pearl, 2009; Kallus and Zhou, 2018; Veitch *et al.*, 2019; Guo *et al.*, 2020a].

For example, measuring a blogger's writing style (confounders) can be extremely difficult, but it often causally influences (1) which type of browsing device is more frequently by readers used to read her blogs; and (2) readers' adoptions of fintech products resulting from her articles. Without being properly controlled, hidden confounding bias can result in overestimated or underestimated causal effects.

From existing methods, we find two desiderata that are used to handle confounding bias. On the group level, it is desirable to balance treatment groups with control groups w.r.t. the distributions of confounder representations [Shalit *et al.*, 2017; Yao *et al.*, 2018]. On the individual level, it is shown to be helpful to capture patterns of hidden confounders that can predict each individual's treatment assignment [Louizos *et al.*, 2017; Rosenbaum and Rubin, 1983]. Regarding networked observational data, a line of recent work [Veitch *et al.*, 2019; Guo *et al.*, 2020c] found that network information can be utilized to mitigate hidden confounding bias. For example, even though it is hard to measure the writing style of a blogger, we can partially catch it by considering her network patterns such as centrality measures and which community she is likely to belong to. This is often achieved through learning representations of hidden confounders from network structure among observational data. However, most of the existing methods can only satisfy one of the two desiderata. The main reason behind this is that the two desiderata seem to contradict each other. Balancing the distributions of confounder representations often makes it more difficult to discriminate the treated instances from the controlled ones. We notice that confounders' representations and predicted treatments are often computed by two separate components.

[1] https://robinhood.com/

[2] https://www.alipay.com/

This fact implies that we can develop a minimax game to optimize each component alternatively toward the two desiderata. In this game, we train a confounder representation function through playing against a discriminator function. The confounder representation function seeks to balance the distributions of confounders' representations. At the same time, the discriminator function aims to distinguish between instances under treatment and those under control.

We summarize the main contributions of this work as:

- We formulate the two desiderata of handling confounding bias as a minimax game.
- We propose the m**I**nimax **G**ame based **N**etwork **ITE** estimator (IGNITE). It learns ITEs from networked observational data. By playing the proposed minimax game, IGNITE balances confounder representations between the treated and the controlled and learns confounder representations to predict the observed treatments.
- Extensive experiments show that IGNITE consistently outperforms 9 state of the art baselines across datasets under various settings.

2 Problem Statement

In this section, we start with technical preliminaries. Then the problem statement is presented.

In networked observational data, each instance is observed with its features x_i, treatment t_i, and outcome y_i. Each instances is connected with its neighbors by a underlying network represented by its adjacency matrix $\mathbf{A}$. Let n denote the number of instances, then $\mathbf{A} \in \{0,1\}^{n\times n}$. $\mathbf{A}_{i,j} = \mathbf{A}_{j,i} = 1$ (0) means there is an (no) edge between the i-th and the j-th instance. Thus, the tuple $(\{\mathbf{x}_i, t_i, y_i\}_{i=1}^n, \mathbf{A})$ denotes a networked observational dataset. In this work, we consider realistic scenarios where the weight of an edge (the importance of the edge between the two instances) may not be accurately measured. This implies that there exists an unseen weighted network $\tilde{\mathbf{A}}$ which represents the importance of edges in terms of their influence on confounders[3]. For simplicity, we adopt the setting of [Shalit *et al.*, 2017; Louizos *et al.*, 2017; Yao *et al.*, 2018] where the treatment takes binary values, i.e., $t \in \{0,1\}$. Then $t_i = 1$ ($t_i = 0$) means the i-th instance is under treatment (control).

To define the individual treatment effect (ITE), following [Rubin, 2005], we assume that, for each instance-treatment pair (i,t), there exists a potential outcome y_i^t. Thus, the observed outcome can be written as a function of the observed treatment and the potential outcomes, i.e., $y_i = t_i y_i^1 + (1-t_i)y_i^0$. The unobserved outcome $y_i^{1-t_i}$ is often referred to as the counterfactual outcome. Then the ITE of instance i is defined as:

$$\tau_i = y_i^1 - y_i^0, \tag{1}$$

which measures the improvement in outcome caused by the treatment for instance i. Then the average treatment effect (ATE) is defined as $\frac{1}{n}\sum_i^n \tau_i$. Following [Veitch *et al.*, 2019; Guo *et al.*, 2020c], we adopt a real-world setting where hidden confounders exist. This means the unconfoundedness assumption [Rubin, 2005; Pearl, 2009] does not hold as:

$$y^1, y^0 \not\perp\!\!\!\perp t|\mathbf{x}. \tag{2}$$

Instead, similar to [Veitch *et al.*, 2019; Guo *et al.*, 2020c], we assume the existence of latent confounders $\mathbf{h}$ such that:

$$y^1, y^0 \perp\!\!\!\perp t|\mathbf{h}. \tag{3}$$

This means controlling the influence of latent confounders $\mathbf{h}_i$ leads to unbiased estimates of ITEs. Note that we cannot observe the latent confounders from networked observational data. But we can approximate them by learning representations of them from networked observational data. Finally, we present the problem statement:

Learning ITEs from Networked Observational Data. *Given the networked observational data $(\{\mathbf{x}_i, t_i, y_i\}_{i=1}^n, \mathbf{A})$ with hidden confounders and unknown edge weights, we aim to develop a causal inference framework which maps each instance $(\mathbf{x}_i, t_i, y_i)$ along with the network information $\mathbf{A}$ to learn the ITE τ_i of each instance i.*

3 Methodology

This section presents the two desiderata of handling confounding bias and the description of the proposed framework.

3.1 Two Desiderata of Handling Confounding Bias

Hidden confounders pose the main challenge of learning ITEs from networked observational data. To handle confounding bias, existing methods present two desiderata.

First, on the group level, it is desirable to balance the distributions of confounders (or their representations) between the treated and the controlled. A variety of representation balancing methods for learning ITEs from observational data have been developed based on this principle [Shalit *et al.*, 2017; Yao *et al.*, 2018]. Let $\hat{\mathbf{h}}_i$ denote the approximated latent confounders' representation of instance i, the representation balancing methods that follow the first desideratum minimize a divergence metric (e.g., Wasserstein distance) between $P(\hat{\mathbf{h}}_i|\mathbf{x}_i, t_i = 1)$ and $P(\hat{\mathbf{h}}_i|\mathbf{x}_i, t_i = 0)$. The second desideratum, on the individual level, aims to capture the patterns of hidden confounders that are useful in predicting treatments. Following this idea, methods proposed in [Louizos *et al.*, 2017; Veitch *et al.*, 2019] learn a function that predicts the observed treatment of each individual based on the confounders' representations. Intuitively, this treatment prediction function mimics the treatment assignment mechanism that generates the data. Therefore, through learning the treatment prediction function, we can capture the information of hidden confounders that explains how the observed treatments are assigned. However, none of the existing methods can satisfy the two desiderata together because they seem to contradict each other. Intuitively, when the divergence between $P(\hat{\mathbf{h}}_i|\mathbf{x}_i, t_i = 1)$ and $P(\hat{\mathbf{h}}_i|\mathbf{x}_i, t_i = 0)$ becomes smaller, it becomes more difficult to distinguish between a treated instance and a controlled one by their confounders' representations. We introduce how to resolve this issue with a minimax game in the next section.

[3] Without loss of generality, we assume that the observed network is undirected.

3.2 The Proposed Framework: IGNITE

We observe that confounders' representations and treatment predictions are often computed by two separate modules. This implies we can develop a minimax game where they are iteratively optimized toward satisfying the two desiderata. We propose IGNITE to learn ITEs from networked observational data. Here, we first introduce the components of IGNITE, then we formulate its loss function including the minimax game for handling confounding bias.

Components of IGNITE. IGNITE has three components: the confounder representation function, the treatment group's critic function, and the outcome inference function.

Confounder Representation Function. Here, we define the confounder representation function $g : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. This function maps the features and the adjacency matrix of the network structure into a d-dimensional representation space to approximate the confounders. To quantify the importance of each edge in its influence on the confounders, we extend the Graph Attention Network layers (GAT) [Veličković *et al.*, 2018]. The i-th instance's confounder representation is a function of its features and network structure. For the simplicity of notation, we formulate the confounder representation function g with a single GAT layer:

$$\hat{\mathbf{h}}_i = g(\mathbf{x}_i, \mathbf{A}) = \|_{k=1}^{K} \delta(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \mathbf{x}_j) \tag{4}$$

where $\|$ denotes concatenation. $\mathcal{N}_i$ is the set of neighbors of the i-th instance in the network $\mathbf{A}$. K is the number of attention heads. Each head of the attention mechanism is a weighted aggregation of information from the neighbors. $\mathbf{W}^k \in \mathbb{R}^{d \times m}$ is the weight matrix of the k-th attention head. δ is the ELU unit. We compute the normalized attention coefficients α_{ij}^k as:

$$\alpha_{ij}^k = \frac{\exp(\delta'(\mathbf{a}^T[\mathbf{W}^k\mathbf{x}_i \parallel \mathbf{W}^k\mathbf{x}_j]))}{\sum_{l \in \mathcal{N}_i} \exp(\delta'(\mathbf{a}^T[\mathbf{W}^k\mathbf{x}_i \parallel \mathbf{W}^k\mathbf{x}_l]))}, \tag{5}$$

where δ' denotes the LeakyReLU unit and $\mathbf{a} \in \mathbb{R}^{2d}$ denotes a weight vector. Stacking multiple GAT layers can help us capture Multi-hop relations.

Treatment Group Critic Function. The critic function $D : \mathbb{R}^d \to \mathbb{R}$ maps the confounders' representation of an instance to a real value. Larger value of $D(\hat{\mathbf{h}}_i)$ indicates that instance i is more likely to receive treatment. Following [Gulrajani *et al.*, 2017], we parameterize it with a neural network that consists of fully connected layers and LeakyReLU units.

Outcome Inference Function. We infer outcomes of an instance based on its confounders' representation. We define the output function $f : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}$. We parameterize the output function of each treatment with fully connected layers with ELU units (except the last layer). We can set $t = t_i$ or $1 - t_i$ to let the corresponding layers infer the factual or counterfactual outcome.

With these three components, given the features of the i-th instance $\mathbf{x}_i$, the treatment t, and the adjacency matrix $\mathbf{A}$, outcomes are inferred as $\hat{y}_i^t = f(g(\mathbf{x}_i, \mathbf{A}), t)$, where $\hat{y}_i^t$ is the inferred outcome of instance i under treatment t. After training, it can infer the ITE of instance i as $\hat{\tau}_i = \hat{y}_i^1 - \hat{y}_i^0$ and estimate the ATE as $\frac{1}{n}\sum_i \hat{\tau}_i$.

A Minimax Game for Handling Confounding Bias. Note that function g is used to compute confounders' representations $\hat{\mathbf{h}}_i$. Here, we formulate the two desiderata of handling confounding bias as a minimax game:

$$\min_g \max_D \mathcal{L}_{CB} = \frac{1}{n^1} \sum_{i:t_i=1} D(\hat{\mathbf{h}}_i) - \frac{1}{n^0} \sum_{i:t_i=0} D(\hat{\mathbf{h}}_i), \tag{6}$$

where n^1 and n^0 are the number of instances under treatment and control. In the maximization stage, the critic function D is trained to maximize the difference between the value it assigns for the treated instances and those for the controlled ones. In the minimization stage, the confounder representation function g is used to fool the treatment group critic D. This step balances the distributions of confounders' representations because it makes it more difficult to distinguish the confounders' representation of a treated instance from that of a controlled one. To avoid difficulty in training (e.g., vanishing gradients), we follow [Gulrajani *et al.*, 2017] to limit the functional space of the treatment group critic D to a subset of 1-Lipschitz functions. To achieve this, we add a gradient penalty term to the maximization stage. It is computed on n' randomly sampled pairs of treated and controlled instances:

$$\mathcal{L}_{GP} = -\frac{1}{n'} \sum_{i=1}^{n'} \lambda(|| \bigtriangledown_{\tilde{\mathbf{h}}_i} D(\tilde{\mathbf{h}}_i)||_2 - 1)^2, \tag{7}$$

where $\tilde{\mathbf{h}}_i = \epsilon\hat{\mathbf{h}}_j + (1-\epsilon)\hat{\mathbf{h}}_k$, (j, k) is one of the n' randomly sampled pairs. Each pair contains a treated instance and a controlled one. $|| \cdot ||_2$ denotes L_2 norm and $\epsilon \sim U[0, 1]$. We set the parameter $\lambda = 10$ as in [Gulrajani *et al.*, 2017]. In addition, we aim to achieve accurate inference of factual outcomes. We minimize the mean squared error on the inferred factual outcomes:

$$\mathcal{L}_{FO} = \frac{1}{n} \sum_i (\hat{y}_i^{t_i} - y_i)^2, \tag{8}$$

Finally, we present the objective functions of the proposed minimax game in two stages:

$$\begin{aligned} \max_D \mathcal{L}_D &= \beta(\mathcal{L}_{CB} + \mathcal{L}_{GP}), \\ \min_g \mathcal{L}_g &= \mathcal{L}_{FO} + \beta(\mathcal{L}_{CB}), \end{aligned} \tag{9}$$

where $\beta \geq 0$ is a hyperparameters controlling the trade-off between the objectives. IGNITE is trained with backpropagation by iteratively optimizing $\mathcal{L}_D$ and $\mathcal{L}_g$.

4 Experiments

In this section, we investigate the two following research questions: **RQ1.** In learning ITEs from networked observational data, is the proposed minimax game more effective in handling confounding bias than representation balancing, treatment prediction or a combination of them? **RQ2.** How does the hyperparameter β affect the performance of the proposed framework, IGNITE?

4.1 Dataset Description

It is extremely challenging to collect ground truth of ITEs because each instance can only be observed with one of the potential outcomes. For instance, we can only observe y_i^1 of the i-th blogger if she has more readers using mobile devices. Thus, we follow previous work [Veitch *et al.*, 2019; Guo *et al.*, 2020c] to create semi-synthetic datasets. To mimic real-world situations, we consider hidden confounders and unobserved edge weights. We include the steps to reproduce the semi-synthetic datasets from the publicly available datasets (BlogCatalog and Flickr).

BlogCatalog (BC) is a social network with blog service. Each instance is a blogger. Each edge signifies the friendship between two bloggers. The features are the keywords of each blogger's articles. We extend the BlogCatalog dataset [Li *et al.*, 2019a] by synthesizing (a) the outcomes – the number of readers who adopt fintech products after reading each blogger's work; and (b) the treatment assignments – whether work of a blogger is browsed more on desktops or on mobile devices. The following assumptions are made: (1) Readers either read on mobile devices or desktops. A blogger is treated (controlled) if her blogs are more popular on mobile devices (desktops). (2) A blogger's articles are either more popular on mobile devices or desktops. (3) A blogger's treatment and outcomes can be influenced by her topics and her neighbors' topics. To synthesize treatments and outcomes, we train an LDA topic model on a large corpus. Then the centroids of the two treatment groups are defined as: (i) the topic distribution of a randomly selected blogger is the centroid of the treatment group, denoted by $\bar{r}^1$; (ii) the centroid of the controlled, $\bar{r}^0$, is the average topic distribution of all the bloggers. Then the treatments and outcomes are generated based on the similarity between the topic distributions of bloggers and the two centroids. Let $r(\mathbf{x}_i)$ denote the topic distribution of the i-th blogger, we model the readers' preference of browsing devices on the blogger's content:

$$Pr(t=1|\mathbf{x}_i, \tilde{\mathbf{A}}) = \frac{\exp(p_i^1)}{\exp(p_i^1)+\exp(p_i^0)}, \tag{10}$$

where p_i^t is calculated as:

$$p_i^t = \kappa_1 r(\mathbf{x}_i)^T \bar{r}^t + \kappa_2 (\tilde{\mathbf{A}} r(\mathbf{x}_j))^T \bar{r}^t, \tag{11}$$

where $t \in \{0,1\}$. $\kappa_1 \geq 0$ ($\kappa_2 \geq 0$) signifies the strength of the confounding bias resulting from a blogger's (her neighbors') topics. When $\kappa_1 = \kappa_2 = 0$ the treatment assignment is random and the greater the value κ_1 and κ_2 are, the more significant the bias of device preference is. $\tilde{\mathbf{A}}$ denotes the weighted adjacency matrix, where each entry $\tilde{\mathbf{A}}_{ij}$ denotes the importance of an edge with related to the influence on confounding bias. To emphasize the fact that in many real-world networks the edge weights are unknown, we only let the unweighted adjacency matrix $\mathbf{A}$ be observed in the data. However, the unobserved weighted adjacency matrix $\tilde{\mathbf{A}}$ is the one that influences the values of treatments and outcomes. Thus, an ideal causal inference approach needs to catch the weights of each edge. If $\mathbf{A}_{ij} = 1$, we sample $\tilde{\mathbf{A}}_{ij} = \tilde{\mathbf{A}}_{ji} \sim U(0.8, 1.2)$; otherwise, we set $\tilde{\mathbf{A}}_{ij} = \tilde{\mathbf{A}}_{ji} = 0$. Outcomes of a blogger are simulated as:

$$y^t(\mathbf{x}_i) = C(p_i^0 + t p_i^1) + \epsilon, \tag{12}$$

where C is a scaling factor and $\epsilon \sim \mathcal{N}(0,1)$. We set $C = 5, \kappa_1 = 10, \kappa_2 \in \{0.5, 1, 2\}$. 50 LDA topics are learned from the training corpus. Then we reduce the vocabulary by taking the union of the most probable 100 words from each topic, which results in 2,173 bag-of-word features.

Dataset	Instances	Edges	Features	κ_2	Average ATE ± STD
BC	5,196	173,468	8,189	0.5	6.079 ± 2.962
				1	9.012 ± 3.602
				2	20.003 ± 8.132
Flickr	7,575	239,738	12,047	0.5	5.130 ± 0.892
				1	7.576 ± 0.715
				2	13.445 ± 2.093

Table 1: Statistics of the Datasets

Flickr is an image and video sharing service. Each instance refers to a user and each edge represents the social relationship between two users. The features of each user represent a list of tags of interest. We adopt the same settings and assumptions as we do for the BC datasets. Thus, we study the ITE of being viewed on mobile devices on the number of readers' adoptions of fintech products recommended by the user's images and videos. We learn 50 topics from the training corpus using LDA and concatenate the top 25 words of each topic which reduces the feature dimension to 1,210. We set the parameters the same as the BC datasets.

In Table 1, we present the statistics of the semi-synthetic datasets. The average and standard deviation of ATE are calculated over the 10 runs under each setting of parameters. The ATE varies because the true edge weights are randomly sampled from the uniform distribution $U(0.8, 1.2)$.

4.2 Experimental Settings

We randomly split the data into training (60%), validation (20%), and test sets (20%), which is repeated ten times for each simulated dataset. We train IGNITE with Adam [Kingma and Ba, 2014] optimizer with weight decay set to 10^{-4}. We iteratively optimize the two objectives in Eq. (9). Grid search finds the optimal set of hyperparameters for IGNITE and the baselines. For IGNITE, we search learning rate in $\{5 \times 10^{-3}, 10^{-3}, 5 \times 10^{-4}, 10^{-4}\}$, the number of GAT layers and fully connected layers of the functions g, D and f in $\{1, 2, 3\}$, the number of hidden units of the GAT layers and the fully connected layers in $\{16, 32, 64, 128\}$, the number of attention heads in $\{2, 4, 8\}$, β in $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Then, we list the baselines:

- **Network Deconfounder (ND)** [Guo *et al.*, 2020c] learns confounders' representations using GCN layer(s) [Kipf and Welling, 2016]. It minimizes the Wasserstein distance between the two confounder representation distributions.
- **GATD** is a variant of ND with GAT layer(s) [Veličković *et al.*, 2018] for fair comparison. **GATD+ and GATDT.** To show the advantage of the proposed minimax game over a simple combination of representation balancing and treatment prediction, we further create two variants of GATD. GATD+ balances confounder representations

and predicts treatments based on these representations. GATDT predicts treatments to handle confounding bias.

- **CNE** [Veitch *et al.*, 2019] learns confounders' representations by predicting observed outcomes, treatments and edges. It does not utilize observed features. CNE uses AIPW [Robins *et al.*, 1994], therefore, only infers ATE.
- **CNE-.** We create a variant of CNE w/o AIPW, which can infer both ATE and ITEs.
- **Counterfactual Regression (CFR) [Shalit *et al.*, 2017]** is a ITEs estimator for i.i.d. data. It minimizes errors on inferred factual outcomes and balances representation distributions. We report the optimal results of the three CFR models: representation balancing with Wasserstein distance, that with Maximum Mean Discrepancy and no representation balancing.
- **CEVAE [Louizos *et al.*, 2017]** is a deep latent-variable model for learning ITEs. It learns the joint distribution of features, latent confounders, treatments, and outcomes to infer ITEs.
- **Causal Forest [Wager and Athey, 2018]** is an ensemble model trained by predicting observed treatments.

For the evaluation metrics, the Rooted Precision in Estimation of Heterogeneous Effect ($\sqrt{\epsilon_{PEHE}}$) and Mean Absolute Error on ATE (ϵ_{ATE}), are used. They are defined as:

$$\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n}\sum_{i=1}(\hat{\tau}_i - \tau_i)^2}, \epsilon_{ATE} = |\frac{1}{n}\sum_{i=1}(\hat{\tau}_i) - \frac{1}{n}\sum_{i=1}(\tau_i)|, \qquad (13)$$

where $\hat{\tau}_i$ and $\tau_i = y_i^1 - y_i^0$ denote the inferred ITE and the ground truth ITE for the i-th instance.

4.3 Experimental Results

Effectiveness. Here, we compare the effectiveness of IGNITE with the baselines in the task of learning ITEs from networked observational data. Table 2 shows the results evaluated on the BC and Flickr datasets with $C = 1, \kappa_1 = 10$ and $\kappa_2 \in \{0.5, 1, 2\}$. We summarize the observations made from these experimental results as follows:

- IGNITE outperforms the baselines consistently in almost all cases. One-tailed T-tests show that the boldfaced results are significantly better than others with a significant level of 0.05.
- IGNITE shows consistent superior performance than GATD+. This verifies that the proposed minimax game does a better job in satisfying the two desiderata than a simple combination of representation balancing and treatment prediction.
- The fact that IGNITE outperforms GATD and GATDT implies that the proposed minimax game handles confounding bias better than doing representation balancing or treatment prediction alone.
- We observe that GATD+ fails to outperform GATD and GATDT in a majority of cases. This implies that a naïve combination of representation balancing and treatment prediction may not achieve the two desiderata together. Instead, it may perform worse than representation balancing or treatment prediction alone.
- GATD outperforms ND under various settings. This is because GAT layers can capture the unobserved edge importance. Note that the unobserved edge importance plays may have a significant influence on the values of treatments and outcomes.
- The improvement of IGNITE over CNE and CNE- results from two aspects. First, the proposed minimax game shows better efficacy in dealing with confounding bias than treatment prediction alone. Second, the GAT layer(s) capture unobserved edge weights and incorporate observed features.
- Compared to the methods for i.i.d. data – CFR, CEVAE, and CF, IGNITE achieves better performance because it is trained by the proposed minimax game for handling confounding bias and it utilizes the network information to recognize patterns of latent confounders.

Parameter Study. Then we investigate how the variation in values of the important hyperparameter β affects the performance of IGNITE. β controls the trade-off between more accurate outcome inference and better confounding bias handling. We set β to $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The following settings are applied: learning rate is 5×10^{-3}, the number of epochs is 300, the number of GAT layer is 2 and the numbers of fully connected layers for D and f are 2 and 1, the number of attention head is 8, the number of hidden units of each attention head and each fully connected layer of D and f are 128, 64 and 32. Due to space limit, we only show the results of this parameter study on the BC datasets in Table 3 as we have similar observations on the Flickr datasets. We observe that IGNITE maintains reasonably consistent performance in terms of both evaluation metrics when $\beta \in [10^{-4}, 10^{-1}]$. In addition, IGNITE often achieves the optimal performance when $\beta \in [10^{-3}, 10^{-2}]$.

5 Related Work

Here, we introduce the related work in the causal inference literature. Limited by space, we could not present the work related to minimax games and graph neural networks.

Causal Inference with Network Data. Researchers aim to utilize networks to approximate hidden confounders using observational studies. Shalizi et al. [Shalizi and McFowland III, 2016] propose a two-stage approach to estimate causal effects in networks based on predefined generative models. To avoid misspecified generative models, Veitch et al. [Veitch *et al.*, 2019] propose causal network embedding (CNE) which learns node embeddings from pure network data to represent confounders. However, CNE relies on treatment prediction alone to handle confounding bias. In addition, CNE requires observable edge weights, only infers ATE and cannot effectively use the observed features. The Network Deconfounder [Guo *et al.*, 2020c] learns representations of confounders from both features and network structures. It handles confounding bias through representation balancing. None of the existing methods can satisfy the two desiderata together. Networks can propagate the treatment received by an instance to interfere the outcomes or treatments of its neighbors. This phenomena can be referred to

	BC						Flickr					
	$\kappa_2 = 0.5$		$\kappa_2 = 1$		$\kappa_2 = 2$		$\kappa_2 = 0.5$		$\kappa_2 = 1$		$\kappa_2 = 2$	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
IGNITE	**4.415**	**0.506**	**6.163**	**0.971**	**10.998**	**2.514**	**6.938**	**1.242**	**10.725**	2.006	**18.864**	**2.643**
GATD+	5.132	0.666	8.442	2.159	17.167	10.74	7.731	1.394	13.201	2.903	27.105	7.088
GATD	5.170	1.070	7.989	1.779	16.574	5.942	7.605	1.688	13.092	2.436	26.846	7.196
GATDT	5.165	1.055	8.017	1.863	16.578	5.940	7.602	1.681	13.075	2.452	26.781	7.099
ND	5.386	2.070	10.403	4.811	20.286	10.350	7.337	2.000	14.006	3.046	28.379	5.817
CNE	–	7.314	–	13.212	–	24.298	–	8.103	–	16.058	–	33.94
CNE-	10.323	8.194	18.839	14.991	33.607	26.531	14.109	9.001	26.536	17.275	54.906	35.262
CFR	10.073	5.000	15.229	9.631	36.680	16.481	9.826	3.619	16.859	7.240	45.150	12.787
CEVAE	6.812	3.129	12.055	2.700	24.128	14.576	11.836	2.678	22.171	3.493	48.840	7.360
CF	5.941	3.349	10.413	3.336	19.145	16.812	8.406	1.938	14.485	**1.821**	31.111	6.520

Table 2: Results on the two datasets with $\kappa_2 \in \{0, 1, 2\}$ measured by the two evaluation metrics $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}, the smaller the better.

		β	10^{-4}	10^{-3}	10^{-2}	10^{-1}
BC	$\kappa_2 = 0.5$	$\sqrt{\epsilon_{PEHE}}$	4.422	4.439	**4.415**	4.566
		ϵ_{ATE}	0.526	0.56	**0.506**	0.642
	$\kappa_2 = 1$	$\sqrt{\epsilon_{PEHE}}$	6.196	**6.163**	6.166	6.177
		ϵ_{ATE}	1.139	**0.971**	0.993	1.124
	$\kappa_2 = 2$	$\sqrt{\epsilon_{PEHE}}$	11.934	**10.998**	12.046	12.385
		ϵ_{ATE}	2.183	**2.514**	2.675	3.134

Table 3: Parameter study results on the BC datasets with $\kappa_2 \in \{0.5, 1, 2\}$ in terms of $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}, the smaller the better.

as contagion [Shalizi and Thomas, 2011], treatment entanglement [Toulis *et al.*, 2018], or spillover effect [Arbour *et al.*, 2016; Rakesh *et al.*, 2018]. Different from them, we follow [Veitch *et al.*, 2019; Guo *et al.*, 2020b] to assume that conditioning on latent confounders decouples each individual's treatment and outcome from those of the others.

Causal Inference with Proxy Variables. When hidden confounders exist, observed proxy variables can be utilized to approximate them. [Pearl, 2012; Miao *et al.*, 2018; Louizos *et al.*, 2017; Veitch *et al.*, 2019]. Most of the existing work assumes that the observed data is i.i.d. and generated by latent confounders. Theoretically, in [Pearl, 2012; Kuroki and Pearl, 2014], authors showed that causal effects can be identified by proxy variables. Miao et al. [Miao *et al.*, 2018] showed that it is feasible to restore the causal effects when the size of the latent confounders is known. Louizos et al. [Louizos *et al.*, 2017] showed that ITE (CATE) can be identified given the joint distribution $P(\mathbf{x}, t, y, \mathbf{z})$ and proposed a deep latent-varibale model to estimate ITEs. Recently, results in [Veitch *et al.*, 2019; Guo *et al.*, 2020c] show that network information can help mitigate confounding bias.

Learning Individual Treatment Effects from i.i.d. Data. Learning ITEs from i.i.d. observational data has attracted great attention. Causal Forest (CF) [Wager and Athey, 2018] is a method that recursively partitions the original feature space through treatment prediction. Its hypothesis is that within each subspace, the instances are very similar in terms of their estimated propensity score. Therefore, we can think the treatment assignment in each subspace is random and the instances in the same subspace share the same ITE. So, CF infers ITEs via applying the naive estimator in each subspace. CFR [Shalit *et al.*, 2017] is a pioneer method for learning ITEs by representation learning. Both theoretical analysis and empirical results indicate that balancing the distributions of the treated and controlled instances in the representation space can improve the performance in learning ITE. However, the methods mentioned above rely on the unconfoundedness assumption, which is often untenable in observational data. Louizos et al. [Louizos *et al.*, 2017] proposed to consider observed features as proxy variables of hidden confounders and use a deep latent-variable model to learn representation of confounders via variational inference. A comprehensive review of these methods can be found in [Guo *et al.*, 2020a]. In [Cheng *et al.*, 2019], a review of the datasets and metrics for evaluation of ITE estimation is presented. However, this line of work does not consider to utilize network information for learning causal effects.

6 Conclusion

In this work, we study the problem of learning ITEs from networked observational data. To mitigate confounding bias, previous work presents two desiderata: representation balancing and treatment prediction. None of the existing methods aim to satisfy them simultaneously. To overcome this issue, we propose a novel framework, IGNITE, which is optimized to satisfy the two desiderata together. We propose a minimax game to train IGNITE. The confounder representation function is trained by playing against the treatment group critic function. Empirical results corroborate the efficacy of IGNITE in mitigating confounding bias.

Future work includes: (1) causal inference with complex data (e.g., dynamic networks [Sarkar *et al.*, 2019; Marin *et al.*, 2017], temporal sequences [Guo *et al.*, 2018], and complex treatment variables [Li *et al.*, 2019b]) and (2) inferring the causal effect of inputs and design choices of deep learning algorithms for interpretation [Moraffah *et al.*, 2020].

Acknowledgements

This material is based upon work supported by ARL and the National Science Foundation (NSF) Grant #1909555.

References

[Arbour *et al.*, 2016] David Arbour, Dan Garant, and David Jensen. Inferring network effects from observational data. In *KDD*, pages 715–724. ACM, 2016.

[Cheng *et al.*, 2019] Lu Cheng, K Selçuk Candan, and Adrienne Raglin. A practical data repository for causal learning with big data. In *BenchCouncil*, 2019.

[Gulrajani *et al.*, 2017] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, pages 5767–5777, 2017.

[Guo *et al.*, 2018] Ruocheng Guo, Jundong Li, and Huan Liu. Initiator: Noise-contrastive estimation for marked temporal point process. In *IJCAI*, pages 2191–2197, 2018.

[Guo *et al.*, 2020a] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *CSUR*, 2020.

[Guo *et al.*, 2020b] Ruocheng Guo, Jundong Li, and Huan Liu. Counterfactual evaluation of treatment assignment functions with networked observational data. In *SDM*, pages 271–279, 2020.

[Guo *et al.*, 2020c] Ruocheng Guo, Jundong Li, and Huan Liu. Learning individual treatment effects from networked observational data. In *WSDM*, 2020.

[Kallus and Zhou, 2018] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *NeurIPS*, pages 9289–9299, 2018.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Kuroki and Pearl, 2014] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

[Li *et al.*, 2019a] Jundong Li, Liang Wu, Ruocheng Guo, Chenghao Liu, and Huan Liu. Multi-level network embedding with boosted low-rank matrix approximation. In *ASONAM*, pages 49–56, 2019.

[Li *et al.*, 2019b] Yichuan Li, Ruocheng Guo, Weiying Wang, and Huan Liu. Causal learning in question quality improvement. In *BenchCouncil*, 2019.

[Louizos *et al.*, 2017] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, pages 6446–6456, 2017.

[Marin *et al.*, 2017] Ericsson Marin, Ruocheng Guo, and Paulo Shakarian. Temporal analysis of influence to predict users' adoption in online social networks. In *SBP*, pages 254–261. Springer, 2017.

[Miao *et al.*, 2018] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[Moraffah *et al.*, 2020] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Ragliny, and Huan Liu. Causal interpretability for machine learning–problems, methods and evaluation. *SIGKDD Exploration*, 2020.

[Pearl, 2009] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[Pearl, 2012] Judea Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

[Rakesh *et al.*, 2018] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. Linked causal variational autoencoder for inferring paired spillover effects. In *CIKM*, pages 1679–1682. ACM, 2018.

[Robins *et al.*, 1994] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89(427):846–866, 1994.

[Rosenbaum and Rubin, 1983] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

[Sarkar *et al.*, 2019] Soumajyoti Sarkar, Ruocheng Guo, and Paulo Shakarian. Using network motifs to characterize temporal network evolution leading to diffusion inhibition. *SNAM*, 9(1):14, 2019.

[Shalit *et al.*, 2017] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, pages 3076–3085. JMLR. org, 2017.

[Shalizi and McFowland III, 2016] 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.

[Shalizi and Thomas, 2011] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2):211–239, 2011.

[Toulis *et al.*, 2018] Panos Toulis, Alexander Volfovsky, and Edoardo M Airoldi. Propensity score methodology in the presence of network entanglement between treatments. *arXiv preprint arXiv:1801.07310*, 2018.

[Veitch *et al.*, 2019] Victor Veitch, Yixin Wang, and David M Blei. Using embeddings to correct for unobserved confounding. In *NeurIPS*, 2019.

[Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *ICLR*, 2018.

[Wager and Athey, 2018] 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.

[Yao *et al.*, 2018] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *NeurIPS*, pages 2633–2643, 2018.