Deconfounding with Networked Observational Data in a Dynamic Environment

Jing Ma[1], Ruocheng Guo[2], Chen Chen[1], Aidong Zhang[1], Jundong Li[1,3*]
[1]University of Virginia, Charlottesville, VA, USA 22904
[2]Arizona State University, Tempe, AZ, USA 85287
[3]Global Infectious Disease Institute, University of Virginia, Charlottesville, VA, USA 22904
{jm3mr, aidong, jundong}@virginia.edu, rguo12@asu.edu, chenannie45@gmail.com

ABSTRACT

One fundamental problem in causal inference is to learn the individual treatment effects (ITE) – assessing the causal effects of a certain treatment (e.g., prescription of medicine) on an important outcome (e.g., cure of a disease) for each data instance, but the effectiveness of most existing methods is often limited due to the existence of hidden confounders. Recent studies have shown that the auxiliary relational information among data can be utilized to mitigate the confounding bias. However, these works assume that the observational data and the relations among them are static, while in reality, both of them will continuously evolve over time and we refer such data as time-evolving networked observational data. In this paper, we make an initial investigation of ITE estimation on such data. The problem remains difficult due to the following challenges: (1) modeling the evolution patterns of time-evolving networked observational data; (2) controlling the hidden confounders with current data and historical information; (3) alleviating the discrepancy between the control group and the treated group. To tackle these challenges, we propose a novel ITE estimation framework *Dynamic Networked Observational Data Deconfounder (DNDC)* which aims to learn representations of hidden confounders over time by leveraging both current networked observational data and historical information. Additionally, a novel adversarial learning based representation balancing method is incorporated toward unbiased ITE estimation. Extensive experiments validate the superiority of our framework when measured against state-of-the-art baselines. The implementation can be accessed in https://github.com/jma712/DNDC.

KEYWORDS

Causal inference, observational data, dynamic networks, treatment effect

ACM Reference Format:
Jing Ma, Ruocheng Guo, Chen Chen, Aidong Zhang, Jundong Li. 2021. Deconfounding with Networked Observational Data in a Dynamic Environment. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM'21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441818

*Corresponding Author.

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441818

1 INTRODUCTION

The increasing prevalence of observational data provides a rich source for estimating individual treatment effect (ITE) – assessing the causal effects of a certain treatment on an outcome for each data instance. ITE estimation has significant implications in many high-impact domains such as health care [1], education [11], and targeted advertising [32]. For example, to provide personalized recommendations for users, service providers need to decide whether the recommendation of a product (treatment assignment) will motivate a user to make a purchase (outcome) based on her profile. Most of existing works on ITE estimation [7, 13, 31, 36] ignore the influence of *hidden confounders* – the unobserved variables that causally affect both the treatment assignment and the outcome. For example, a user's purchasing preferences can be regarded as hidden confounders that causally impact the items recommended to her and her purchasing patterns [3, 30]. In other words, most existing works heavily rely on the *strong ignorability* assumption [27] that there does not exist any hidden confounders. However, without controlling the influence of hidden confounders, these methods may result in biased estimation of ITE [19].

To mitigate the bias induced by hidden confounders, recent studies [8–10] leverage auxiliary relational information (e.g., social connections, patient similarity) beyond the traditional i.i.d observational data for more accurate ITE estimation. Despite their empirical success, these works overwhelmingly assume that the observational data and the relations among them are static. In fact, both of them are naturally dynamic in many real-world scenarios [18]. For example, the purchasing preferences of users and their social connections are both evolving over time. We refer such data as *time-evolving networked observational data*. The prevalence of such data in a wide spectrum of domains brings about new opportunities to unravel the patterns of hidden confounders towards unbiased ITE estimation.

In this paper, we investigate a novel research problem of *deconfounding with networked observational data in a dynamic environment*. The causal graph of the problem setting is illustrated in Fig. 1, where the hidden confounders at a particular time stamp not only have causal relations to the observed variables at the same time stamp but also be causally affected by both the hidden confounders and the treatment assignment from previous time stamps [2], e.g., user purchasing preferences change over time, and are deeply influenced by their previous preferences and previously recommended

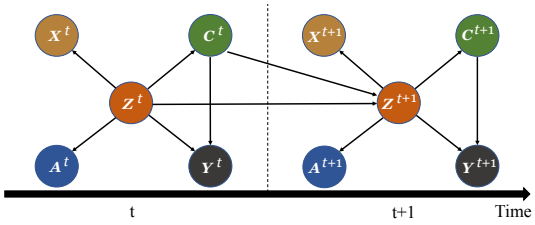


Figure 1: Causal graph for the studied problem. At time t, we use X^t, A^t, Z^t, C^t, Y^t to denote the features of observational data, relations among observational data, representations of hidden confounders, treatment assignment, and outcomes, respectively. The hidden confounders Z^{t+1} at $t+1$ causally affect the treatment assignment C^{t+1} and the outcome Y^{t+1} at that time. To infer Z^{t+1}, we can leverage the networked observational data X^{t+1} and A^{t+1} at $t+1$, previous hidden counfounders Z^t, and treatment assignment C^t. The black lines indicate the causal relations.

products to them, and the current user purchasing preferences would also influence the current profiles and social connections. In this work, we do not assume that the previous outcomes can causally affect the variables in the current time stamp, because in real world, it is a more common setting that the previous outcomes do not have causal relationship on the variables in the current time stamp, even if they might be correlated.

However, the problem remains difficult because of the following multifaceted challenges: (1) Most of the existing causal inference frameworks focus on static observational data. In dynamic environment, both modalities of networked observational data are evolving, how to systematically model the evolution patterns of different data modalities for unbiased ITE estimation requires deep investigation. (2) Previous studies have shown that hidden confounders can be approximated by the representations learnt from networked observational data [8, 9]. Since the hidden confounders at the current time stamp can be controlled by two sources of information – (i) the current networked observational data; and (ii) previous hidden confounders and treatment assignments, it is of vital importance to jointly model these two different sources. (3) Representation balancing has been widely adopted to control confounding bias for unbiased causal effect estimation [9, 14, 31], where *confounding bias* exists when the correlation between the treatment and the outcome is distorted by the existence of confounders. This problem becomes more important in our setting as uncontrolled confounding bias can accumulate over time, and degrade the precision of estimated causal effects in later time stamps. Thus, a more principled balancing method is often desired in dynamic environment.

To address the aforementioned challenges, we propose a novel causal inference framework *Dynamic Networked Observational Data Deconfounder (DNDC)*, which learns dynamic representations of hidden confounders over time by mapping the current observational data and historical information into the same representation space. Additionally, we propose a novel method based on adversarial learning to balance the representations of hidden confounders from the treated group and the control group. The main contributions of this work can be summarized as: 1) **Problem Formulation:** We formulate a new task of ITE estimation with networked observational data in dynamic environment and analyze its fundamental importance and challenges. 2) **Algorithm Design:** We propose a novel causal inference framework *DNDC* to tackle the challenges of the studied problem. *DNDC* leverages the evolving data of both observational variables and network structure, and learns dynamic representations of hidden confounders. A novel adversarial learning based representation balancing method is also incorporated toward unbiased ITE estimation. 3) **Experimental Evaluation:** Experimental results on real-world time-evolving networked observational data show that *DNDC* outperforms state-of-the-art methods.

Table 1: Notations.

Notation	Definition
$(\cdot)^t$	variables at time stamp t *
$(\cdot)^{<t}$, $(\cdot)^{\leq t}$	historical variables before time stamp t (not including/ including t)
$\boldsymbol{X}^t$, $\boldsymbol{x}_i^t$	features of all instances/the i-th instance
$\boldsymbol{C}^t$, $\hat{\boldsymbol{C}}^t$	true/predicted treatment assignment
c_i^t	treatment assignment for the i-th instance
$\boldsymbol{Y}^t$	observed outcome
Y_1^t, $\hat{Y}_1^t$	true/predicted potential outcome when get treated
$y_{1,i}^t$, $\hat{y}_{1,i}^t$	true/predicted potential outcome for the i-th instance when $c_i^t = 1$
Y_0^t, $\hat{Y}_0^t$	true/predicted potential outcome when not get treated (controlled)
$y_{0,i}^t$, $\hat{y}_{0,i}^t$	true/predicted potential outcome for the i-th instance when $c_i^t = 0$
$\boldsymbol{\tau}^t$, $\hat{\boldsymbol{\tau}}^t$	true/predicted ITE
τ_i^t, $\hat{\tau}_i^t$	true/predicted ITE of the i-th instance
$\boldsymbol{A}^t$	network structure among data
$\boldsymbol{Z}^t$	hidden confounders
$\boldsymbol{z}_i^t$	hidden confounders of the i-th instance
$\mathcal{H}^t$	historical data $\{\boldsymbol{X}^{<t}, \boldsymbol{A}^{<t}, \boldsymbol{C}^{<t}\}$ before time stamp t
$\tilde{\boldsymbol{H}}^t$	representation of historical information
$\boldsymbol{H}^t$	hidden state of GRU
$y_{F,i}^t$, $y_{CF,i}^t$	factual(observed)/counterfactual outcome for i-th instance
$\hat{y}_{F,i}^t$, $\hat{y}_{CF,i}^t$	predicted factual/counterfactual outcome
d_h, d_z	dimension of the representation of historical information and hidden confounders
T	# of time stamps
n^t	# of instances at time stamp t

2 PROBLEM DEFINITION

The time-evolving networked observational data is denoted as $\{\boldsymbol{X}^t, \boldsymbol{A}^t, \boldsymbol{C}^t, \boldsymbol{Y}^t\}_{t=1}^T$ across T different time stamps. Let $\boldsymbol{X}^t$ be the attributes (features) of observational data at time stamp t, such that $\boldsymbol{X}^t = \{\boldsymbol{x}_1^t, ..., \boldsymbol{x}_{n^t}^t\}$, where $\boldsymbol{x}_i^t$ represents the i-th instance (e.g., profile of each user), n^t denotes the number of instances, and $\boldsymbol{A}^t$ represents the adjacency matrix of the auxiliary network information among different data instances (e.g., user social connections). For simplicity, we assume the network is undirected and unweighted, but it can be naturally extended to directed and weighted networks. At time stamp t, the treatment assignment for these n^t instances is denoted by $\boldsymbol{C}^t = \{c_1^t, ..., c_{n^t}^t\}$, where c_i^t is either 0 or 1 (e.g., if a user receives the recommendation of a specific item or not). The observed outcome of all instances at time stamp t is denoted by $\boldsymbol{Y}^t = \{y_1^t, ..., y_{n^t}^t\}$ (e.g., if user buys

the item or not). $Z^t = \{z_1^t, ..., z_{n^t}^t\}$ stands for the hidden confounders (e.g., users' purchasing preferences). We use the superscript "$< t$" to denote the historical data before time stamp t, e.g., the instance features before time stamp t is referred to as $X^{<t} = \{X^1, X^2, ..., X^{t-1}\}$, and $C^{<t}, A^{<t}$ are defined similarly. Additionally, we use $\mathcal{H}^t = \{X^{<t}, A^{<t}, C^{<t}\}$ to denote all the historical data before time t. A detailed description of notations can be referred to Table 1. In this paper, we build our framework upon the well-adopted potential outcome framework [22, 28]. The *potential outcome* of the i-th instance under treatment c at time stamp t is denoted by $y_{c,i}^t \in \mathbb{R}$, which is the value of outcome that would be realized if instance i receives treatment c at time t. We represent the potential outcome of all instances at time stamp t by $Y_1^t = \{y_{1,1}^t, ..., y_{1,n^t}^t\}$ and $Y_0^t = \{y_{0,1}^t, ..., y_{0,n^t}^t\}$. Then we define the individual treatment effect (ITE) on time-evolving networked observational data as: $\tau_i^t = \tau^t(\boldsymbol{x}_i^t, \mathcal{H}^t, \boldsymbol{A}^t) = \mathbb{E}[y_{1,i}^t - y_{0,i}^t | \boldsymbol{x}_i^t, \mathcal{H}^t, \boldsymbol{A}^t]$.[1]

With ITE defined, the average treatment effect (ATE) is defined as $\tau_{ATE}^t = \frac{1}{n^t} \sum_{i=1}^{n^t} \tau_i^t$. With the above definitions, we can formally define the studied problem of learning individual treatment effect with time-evolving networked observational data as follows:

Definition 1. *(Learning ITE on Time-Evolving Networked Observational Data). Given the time-evolving networked observational data $\{\boldsymbol{X}^t, \boldsymbol{A}^t, \boldsymbol{C}^t, \boldsymbol{Y}^t\}_{t=1}^T$ across T different time stamps, the goal is to learn the ITE τ_i^t for each instance i at each time stamp t.*

It is worth noting that our work is different from the existing settings of *spillover effect* [26] or *treatment entanglement* [33], where the treatment on an instance may causally influence the outcomes of its neighbor units. In our setting, the network structures are exploited for controlling confounding bias. In particular, we assume that conditioning on the latent confounders, the treatment assignment and the outcome of an instance would not causally influence the treatment assignment or the outcome of other instances.

Most existing works [13, 31, 36] rely on the *strong ignorability* assumption [27], assuming that observed features are enough to eliminate the confounding bias, i.e., no hidden confounders exist.

Definition 2. *(Strong Ignorability Assumption). Given an instance's observed features, the potential outcome of this instance is independent of its treatment assignment: $y_{1,i}^t, y_{0,i}^t \perp\!\!\!\perp c_i^t | \boldsymbol{x}_i^t$.*

However, this assumption is often untenable due to the existence of hidden confounders in real-world scenarios [24]. Our method relaxes this assumption as there exist hidden confounders Z^t at each time stamp t which causally influence the treatment C^t and the potential outcome (Y_1^t and Y_0^t). Conditioning on Z^t, the treatment assignment is randomized, i.e., $y_{1,i}^t, y_{0,i}^t \perp\!\!\!\perp c_i^t | \boldsymbol{z}_i^t$. We aim to learn the representations of hidden confounders for bias elimination based on following assumption:

Assumption 1. *(Existence of Hidden Confounders) (i) The hidden confounders may not be accessible, but we assume that the instance features and network structures are both correlated with the hidden confounders, and can be considered as proxy variables. (ii) Hidden confounders at each time stamp are also influenced by the hidden confounders and treatment assignment from previous time stamps.*

[1] In this work, we follow [31] to define ITE in the form of the Conditional Average Treatment Effect (CATE).

Based on the above assumption and some common assumptions in causal inference described in Section 4, we now show the identification result of our framework. For simplicity, we drop the instance index i for notations $\boldsymbol{z}^t, \boldsymbol{x}^t, y^t, c^t$:

Theorem 2.1. *(Identification of ITE) If we recover $p(\boldsymbol{z}^t | \boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t)$ and $p(y^t | \boldsymbol{z}^t, c^t)$, then the proposed DNDC can recover the ITE under the causal graph in Fig. 1.*

3 THE PROPOSED FRAMEWORK

We propose a framework *DNDC* for ITE estimation in time-evolving networked observational data. The overall framework, as illustrated by Fig. 2, consists of three essential components: confounder representation learning, potential outcome and treatment prediction, and representation balancing. Firstly, *DNDC* learns representations of hidden confounders over time by mapping the current networked observational data and historical information into the same representation space. Later on, the learnt representations are leveraged for the potential outcome prediction and the treatment prediction. Additionally, to balance the representations of hidden confounders from the treated group and the control group, we develop a novel adversarial learning based balancing method. Next, we will elaborate on these components in the following subsections.

3.1 Confounder Representation Learning

As mentioned previously, the hidden confounders can be causally related to the current features of data instances and the relations among them, as well as the previous hidden confounders and treatment assignment. Therefore, the proposed *DNDC* first aims to learn the representations of hidden confounders by taking advantage of the aforementioned information. As the confounders can be related to the relations among observational data in addition to the feature information, we propose to use graph convolutional networks (GCNs) [15] to capture the influence of these two different data modalities in learning the representations of hidden confounders:

$$\boldsymbol{z}_i^t = g((\boldsymbol{[X^t, \tilde{H}^{t-1}]})_i, \boldsymbol{A}^t) = \hat{\boldsymbol{A}}^t \mathrm{ReLU}((\hat{\boldsymbol{A}}^t \boldsymbol{[X^t, \tilde{H}^{t-1}]})_i \boldsymbol{U}_0) \boldsymbol{U}_1, \quad (1)$$

where $g(\cdot)$ denotes the transformation function parameterized by GCNs. Here, we stack two GCN layers to capture the non-linear dependency between the current hidden confounders and the input. $\boldsymbol{U}_0, \boldsymbol{U}_1$ denote the parameters of two GCN layers. $\tilde{\boldsymbol{H}}^{t-1} \in \mathbb{R}^{n^t \times d_h}$ denotes the historical information before time stamp t, which compresses previous hidden confounders and treatment assignment. $\boldsymbol{z}_i^t \in \mathbb{R}^{d_z}$ denotes the representation of hidden confounders for instance i at time t. Also, $[.,.]$ denotes the concatenation operation and $(\cdot)_i$ represents the i-th row of the matrix. The matrix $\hat{\boldsymbol{A}}^t$ is the normalized adjacency matrix computed from $\boldsymbol{A}^t$ beforehand with the renormalization trick [15].

To characterize the evolution patterns of time-evolving networked observational data, we make use of a Gated Recurrent Unit (GRU) [5] based memory unit to catch the temporal dependency between the current information and the historical information, in both of which the hidden confounders and treatment assignment are encoded. Specifically, in the GRU, the current information ($\boldsymbol{Z}^t$, $\boldsymbol{X}^t$, $\boldsymbol{C}^t$) and previous hidden state $\boldsymbol{H}^{t-1}$ are first compressed into a new hidden state $\boldsymbol{H}^t \in \mathbb{R}^{n^t \times d_h}$ by the GRU layer: $\boldsymbol{H}^t = \mathrm{GRU}(\boldsymbol{H}^{t-1}, [\boldsymbol{Z}^t, \boldsymbol{X}^t, \boldsymbol{C}^t])$. It should be noted that the

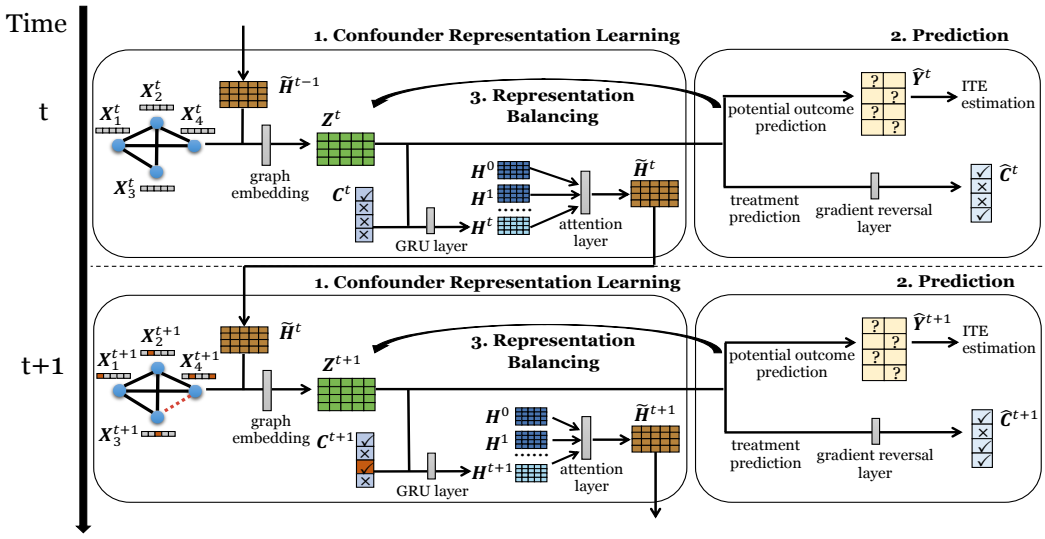


Figure 2: An illustration of the proposed framework *DNDC*.

confounders' representation $\boldsymbol{Z}^t$ does not take the treatment assignment $\boldsymbol{C}^t$ at time stamp t as its input because it stands for the *pre-treatment* status of an instance. At the same time, $\boldsymbol{C}^t$ is a part of the hidden state $\boldsymbol{H}^t$ of the GRU and will contribute to learning the historical information $\tilde{\boldsymbol{H}}^t$.

To weigh the importance of the historical information from different time stamps, attention mechanism [20, 34] is applied among different hidden states of GRU. The attention weight $\alpha_{t,s}$ that models the importance of the hidden states of GRU from time stamp s on those of time stamp t ($s < t$) can be calculated with different attention score functions on $\boldsymbol{h}^t$ and $\boldsymbol{h}^s$, e.g., the widely used bilinear [20] function or the scaled dot product [34] function. At each time stamp, the attention weights are normalized by softmax function: $\boldsymbol{\alpha}_t = \text{softmax}([\alpha_{t,1}, \alpha_{t,2}, ...\alpha_{t,t-1}])$. With these attention weights, we can obtain a context vector $\boldsymbol{v}^t$ at each time stamp t, which will capture the historical dependency by a linear combination of previous hidden states of GRU: $\boldsymbol{v}^t = \sum_{s=1}^{t-1} \alpha_{t,s}\boldsymbol{h}^s$. In order to incorporate the current and historical information, we concatenate $\boldsymbol{h}^t$ and $\boldsymbol{v}^t$ and feed it into a MLP layer to generate the representation for the historical information till time stamp t as: $\tilde{\boldsymbol{h}}^t = \text{MLP}([\boldsymbol{h}^t, \boldsymbol{v}^t])$, where $\tilde{\boldsymbol{h}}^t \in \mathbb{R}^{d_h}$ is the compressed vector of historical information for an instance at time stamp t. For all instances, they form a matrix $\tilde{\boldsymbol{H}}^t$. As shown in Fig. 2, at the next time stamp, we will feed $\tilde{\boldsymbol{H}}^t$ as input to the confounding representation phase to capture the evolution of confounders over time.

3.2 Prediction with Hidden Confounders

Potential outcome prediction. The proposed *DNDC* infers the potential outcome with two functions $f_1, f_0 : \mathbb{R}^{d_z} \to \mathbb{R}$, corresponding to the two cases when treatment is taken or not, i.e., $\hat{y}^t_{1,i} = f(\boldsymbol{z}^t_i, c^t_i = 1) = f_1(\boldsymbol{z}^t_i)$, $\hat{y}^t_{0,i} = f(\boldsymbol{z}^t_i, c^t_i = 0) = f_0(\boldsymbol{z}^t_i)$. Here, we use two MLPs to model f_1 and f_0. In this way, for each instance, both of its *factual outcome* $y^t_{F,i}$ (observed outcome) and *counterfactual outcome* $y^t_{CF,i}$ (unobserved outcome with the contrary treatment) are estimated. The loss function of the potential outcome inference module is defined using the mean squared error:

$$\mathscr{L}_y = \mathbb{E}_{t\in[T],i\in[n^t]}[(\hat{y}^t_{F,i} - y^t_{F,i})^2]. \tag{2}$$

Treatment prediction. The proposed *DNDC* also contains a treatment predictor, which takes $\boldsymbol{Z}^t$ as the input, and the actual treatment assignment $\boldsymbol{C}^t$ as the target. This treatment predictor is implemented by a MLP with softmax operation as the last layer. The loss function of the treatment predictor is:

$$\mathscr{L}_c = -\mathbb{E}_{t\in[T],i\in[n^t]}[(c^t_i \log(\hat{s}^t_i) + (1 - c^t_i)\log(1 - \hat{s}^t_i))], \tag{3}$$

where $\hat{s}^t_i$ is the output of the softmax layer, which can be considered as the predicted propensity score for instance i at time stamp t. Specifically, the propensity score of an instance i refers to its probability to be treated (in the treated group) such that $P(c^t_i = 1|\boldsymbol{x}^t_i, \boldsymbol{A}^t, \mathcal{H}^t)$, and in our setting we approximate it with $\hat{s}^t_i = \text{softmax}(\text{MLP}(\boldsymbol{Z}^t_i))$.

3.3 Representation Balancing

Recent studies [31] theoretically show that balancing the representations of treated and control groups would help mitigate the confounding bias and minimize the upper bound of the outcome inference error. Motivated by this, in our work, we study the problem of learning a balanced representation for ITE estimation from time-evolving networked observational data and develop a novel adversarial learning based balancing method.

Adversarial Learning based Balancing. We propose to use a gradient reversal layer [6] to solve the representation balancing problem. Specifically, the gradient reversal layer does not change the input during the forward-propagation phase, but when back-propagation happens, the gradient reversal layer reverses the gradient by multiplying it by a negative scalar. Intuitively, during back-propagation, the gradient reversal layer enables us to (1) train the treatment predictor by minimizing the treatment prediction loss $\mathscr{L}_c$; and (2) achieve representation balancing via maximizing $\mathscr{L}_c$ w.r.t. the model parameters of the confounder representation learning. In particular, we add the gradient reversal layer before

the treatment predictor to ensure the confounder representation distribution of the treated and that of the controlled are similar at the group level. At the same time, we still utilize the observed treatment assignment as a supervision signal to learn the confounder representation of each instance. In this way, balanced representations are learned for potential outcome prediction and treatment prediction. As the model will minimize the loss of the treatment prediction, the adversarial learning process will benefit from both the treatment predictor and the distribution balancing.

3.4 Loss Function for the Proposed *DNDC*

By putting all the aforementioned components together, we obtain the final loss function of the proposed *DNDC* framework:

$$\mathscr{L}\{\{\boldsymbol{x}_i^t, y_i^t, c_i^t\}_1^{n^t}, \boldsymbol{A}^t\}_1^T = \mathscr{L}_y + \beta\mathscr{L}_c + \gamma||\boldsymbol{\theta}||^2, \tag{4}$$

where β, γ are the hyperparameters to control the effect of different parts, $\boldsymbol{\theta}$ is the set of parameters in this model and $\gamma||\boldsymbol{\theta}||^2$ is included to prevent overfitting. To show how the proposed adversarial learning based balancing method works in the training process, we write the gradient updates that happen while minimizing Eq. (4) as:

$$\begin{aligned}
\boldsymbol{\theta}_z &\leftarrow \boldsymbol{\theta}_z - \mu(\frac{\partial\mathscr{L}_y}{\partial\boldsymbol{\theta}_z} - \omega\frac{\partial\beta\mathscr{L}_c}{\partial\boldsymbol{\theta}_z} + 2\gamma\boldsymbol{\theta}_z),\\
\boldsymbol{\theta}_c &\leftarrow \boldsymbol{\theta}_c - \mu(\frac{\partial\beta\mathscr{L}_c}{\partial\boldsymbol{\theta}_c} + 2\gamma\boldsymbol{\theta}_c),\\
\boldsymbol{\theta}_y &\leftarrow \boldsymbol{\theta}_y - \mu(\frac{\partial\mathscr{L}_y}{\partial\boldsymbol{\theta}_y} + 2\gamma\boldsymbol{\theta}_y),
\end{aligned} \tag{5}$$

where $\boldsymbol{\theta}_z$, $\boldsymbol{\theta}_c$, and $\boldsymbol{\theta}_y$ are the model parameters of the hidden confounder representation learning, the treatment prediction, and the potential outcome prediction, respectively. When updating $\boldsymbol{\theta}_z$, the gradient backpropagated from the treatment predictor is reversed by multiplying with a negative constant $-\omega$. The positive real scalar μ stands for the learning rate of the optimization process.

4 THEORY

Before the formal proof of Theorem 1, as there are some common assumptions used in most works as well as ours for ITE estimation, we first present them under our setting:

Assumption 2. *(Consistency). If the treatment history is $\boldsymbol{C}^{\leq t}$, then the observed outcome $\boldsymbol{Y}^{\leq t}$ equals to the potential outcome under treatment $\boldsymbol{C}^{\leq t}$.*

Assumption 3. *(Positivity). If the probability $p(\boldsymbol{z}^t) \neq 0$, then the probability of any treatment assignment $\boldsymbol{c}^t$ at time stamp t is in the range of $(0, 1)$, i.e., $0 < p(\boldsymbol{c}^t|\boldsymbol{z}^t) < 1$.*

Assumption 4. *(SUTVA). The potential outcomes for any instance are not influenced by the treatment assignment of other instances, and, for each instance, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

In most existing works, the identification of ITE is based on above three assumptions, along with the strong ignorability assumption. In this paper, we relax the strong ignorability assumption and allow the existence of the hidden confounders which could be captured from the time-evolving networked observational data. Based on above premise, we study on the identification of ITE in such data:

Theorem 1. (Identification of ITE) If we recover $p(\boldsymbol{z}^t|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t)$, $p(y^t|\boldsymbol{z}^t, c^t)$, then the proposed *DNDC* can recover the ITE under the causal graph in Fig. 1.

Table 2: Detailed statistics of the datasets.

Dataset	Flickr	BlogCatalog	PeerRead
# of instances	7,575	5,196	6,867 ~ 7,601
# of links	236,582	170,626	11,819
	~ 240,374	~ 173,524	~ 13,684
# of features	12,047	8,189	1,087
# of time stamps	25	20	17
ratio (%) of treated	48.72 ± 1.42	46.52 ± 1.58	56.52 ± 3.36
Avg ATE ± STD	14.35 ± 21.10	20.45 ± 16.63	60.12 ± 83.57

Proof. Under these above assumptions, we can prove the identification of ITE:

$$\tau^t \overset{(i)}{=} \mathbb{E}_y[y_1^t - y_0^t|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t] \tag{6}$$

$$\overset{(ii)}{=} \mathbb{E}_z[\mathbb{E}_y[y_1^t - y_0^t|\boldsymbol{x}^t, \boldsymbol{z}^t, \mathcal{H}^t, \boldsymbol{A}^t]|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t] \tag{7}$$

$$\overset{(iii)}{=} \mathbb{E}_z[\mathbb{E}_y[y_1^t - y_0^t|\boldsymbol{z}^t]|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t] \tag{8}$$

$$\overset{(iv)}{=} \mathbb{E}_z[\mathbb{E}_y[y_1^t|\boldsymbol{z}^t, c^t = 1] - \mathbb{E}_y[y_0^t|\boldsymbol{z}^t, c^t = 0]|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t] \tag{9}$$

$$\overset{(v)}{=} \mathbb{E}_z[\mathbb{E}_y[y_F^t|\boldsymbol{z}^t, c^t = 1] - \mathbb{E}_y[y_F^t|\boldsymbol{z}^t, c^t = 0]|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t], \tag{10}$$

where $\tau^t = \tau^t(\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t)$, we drop the instance index i for simplification. The equation (i) is the definition of ITE in our setting, equation (ii) is a straightforward expectation over $p(\boldsymbol{z}^t|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t)$, equation (iii) can be inferred from the structure of the causal graph shown in Fig. 1, and the SUTVA assumption is implicitly used in the causal graph, equation (iv) is based on the assumption that $\boldsymbol{z}^t$ contains all the hidden confounders, as well as the positivity assumption, and equation (v) can be inferred from the consistency assumption. Thus, if our framework can correctly model $p(\boldsymbol{z}^t|\boldsymbol{x}^t, \mathcal{H}^t, \boldsymbol{A}^t)$ and $p(y^t|\boldsymbol{z}^t, c^t)$, then the ITEs can be identified under the causal graph in Fig. 1. □

5 EVALUATION

In this section, we conduct extensive experiments to validate the effectiveness of the proposed framework *DNDC*. Considering that the counterfactual outcome and hidden confounders are often unavailable in most real-world scenarios, it is notoriously hard to collect the ground-truth ITEs on real-world observational datasets. Thus, we follow existing literature [8, 9] to create semi-synthetic datasets under different settings.

5.1 Datasets and Simulation

5.1.1 Datasets. The datasets used in the experiments are based on three real-world attributed networks, Flickr, BlogCatalog, and PeerRead. The key statistics of these datasets are shown in Table 2, including the number of instances, links, features, and time stamps, as well as the ratio of the treated instances, and the average ATE and its standard deviation over 10 experiments.

Flickr. Flickr[2] is an image and video based social network, where each node represents a user and each edge stands for the friendship between two users. At each time stamp, we randomly inject/remove 0.1 ~ 1.0% edges, and perturb 0.1% node features (based on the noises sampled from $N(0, 0.01^2)$), yielding a dynamic network across 25 time stamps. The features are a list of tags showing users'

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

interest. We train a 50-topic LDA model [25] on the features and select the top-25 most frequent words from each topic to create hidden confounding. We create a semi-synthetic dataset with the following assumptions: (1) *Treatment.* A user has more viewers from either mobile devices (treated) or desktops (controlled). (2) *Outcome.* A user receives some reviews from the viewers of her posts. (2) *Confounders.* Viewers of a user have their preferences of devices which are influenced by the *topics* of the user's posts. These topics of users causally influence both the devices chosen by their viewers and the reviews they get. (4) *Historical influence.* The topics of a user's posts can be influenced by her previously observed treatment (more viewers from mobile devices or desktops) and the social network, e.g., if a user finds more viewers of her posts are from mobile devices, then she may consider to post more about the topics (e.g., sports) that are preferred by users on mobile devices. To study the ITE of peoples' device preference on the reviews, we simulate the confounders, treatment assignments, and outcome. The detailed simulation process is described in Section 5.1.2.

BlogCatalog. BlogCatalog[3] is a social network website where bloggers can share their blogs, where each node represents a blogger and each edge stands for a social relationship between two bloggers. The node features are the bag-of-word representations of the blogger's blogs. As this dataset is also a static data, we follow the same process as Flickr to generate a time-evolving attributed network across 20 different time stamps. We further simulate the treatment assignment, confounders, and outcome in the same way as Flickr.

PeerRead. PeerRead[4] is a dataset of computer scientific peer reviews for papers, and has been used in previous research of causal inference [35]. This dataset contains a real-world dynamic network of coauthor relations over time. We select 17 time stamps of dynamic network which contains 6867 ∼ 7601 authors. In this dataset, each node refers to an author, and each edge represents their co-author relationship. The features are the bag-of-word representations of their paper titles and abstracts. The confounders are their research areas. The treatment is whether the authors' papers contain buzzy words in their titles and abtracts, which are words in a preset dictionary {"deep", "neural", "network", "model"}. The outcome denotes the citation numbers of authors. To study the ITE of buzzy words on the authors' citation, we use the real-world treatment, and simulate the confounders and potential outcomes in the same way as Flickr.

5.1.2 Data Simulation. We incorporate the effect of historical influence (as a p-order autoregressive term [21]) and network information to simulate the confounders $\boldsymbol{z}_i^t$ at time stamp t as:

$$\boldsymbol{z}_i^t = \left(\frac{1}{\sum_{k=1}^{3}\lambda_k}\right)\left(\lambda_1 \boldsymbol{\psi}_i^t + \lambda_2 \sum_{u\in N(i)} f(\boldsymbol{x}_u^t) + \lambda_3 f(\boldsymbol{x}_i^t)\right) + \epsilon^t, \quad (11)$$

$$\psi_{i,j}^t = \frac{1}{p}\left(\sum_{r=1}^{p}\alpha_{r,j} z_{i,j}^{t-r} + \sum_{r=1}^{p}\beta_r c_i^{t-r}\right), \quad (12)$$

where $\boldsymbol{z}_i^t$ denotes the hidden confounders of instance i at time stamp t. $\boldsymbol{\psi}_i^t$ denotes the historical information. $z_{i,j}^t$ and $\psi_{i,j}^t$ represents the j-th dimension of $\boldsymbol{z}_i^t$ and $\boldsymbol{\psi}_i^t$, respectively. $N(i)$ denotes the neighboring nodes of node i. The function $f(\cdot)$ maps the bag-of-words features of instances to their LDA topics [25]. Here, we train a LDA model with 50 topics with the whole training corpus. The parameters λ_1, λ_2 and λ_3 control the impact of historical information, current network structure, and current features on the confounders. In the experiments, we set $\lambda_1 = 0.3, \lambda_2 = 0.3, \lambda_3 = 0.3$ by default. $\epsilon^t \sim \mathcal{N}(0, 0.001^2)$ is the random noise, $\alpha_{r,j} \sim \mathcal{N}(1-(r/p), (1/p)^2)$ controls the impact of historical information from the previous p time stamps, where p is set to 3 by default, $\beta_r \sim \mathcal{N}(0, 0.02^2)$ controls the impact of previous treatment assignment.

To synthesize observed treatment assignment, we select two points $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ in the LDA topic space as the centroids for the treated and controlled groups. We simulate the treatment as follows:

$$c_i^t \sim Bernoulli\left(\frac{exp(p_{i,1}^t)}{exp(p_{i,0}^t) + exp(p_{i,1}^t)}\right), \quad (13)$$

$$p_{i,0}^t = \boldsymbol{r}_0^{'}\boldsymbol{z}_i^t,\ p_{i,1}^t = \boldsymbol{r}_1^{'}\boldsymbol{z}_i^t. \quad (14)$$

Then we synthesize the potential outcomes as below by setting $c = 1$ or $c = 0$:

$$y_{c,i}^t = S(p_{i,0}^t + c \cdot p_{i,1}^t) + \eta^t \quad (15)$$

where S is a scaling factor, and is specified as $S = 20$. $\eta^t \sim \mathcal{N}(0, 0.5^2)$ is a random noise term.

5.2 Evaluation Metrics

We adopt two widely used evaluation metrics – Rooted Precision in Estimation of Heterogeneous Effect (PEHE) [13] and Mean Absolute Error (ATE) [38] to measure the quality of the estimated individual treatment effects at different time stamps:

$$\sqrt{\epsilon_{PEHE}^t} = \sqrt{\frac{1}{n^t}\sum_{i\in[n^t]}(\hat{\tau_i^t} - \tau_i^t)^2}. \quad (16)$$

$$\epsilon_{ATE}^t = |\frac{1}{n^t}\sum_{i\in[n^t]}\hat{\tau_i}^t - \frac{1}{n^t}\sum_{i\in[n^t]}\tau_i^t|. \quad (17)$$

We take the average over all time stamps for evaluation.

5.3 Experiment Settings

Baselines. To investigate the effectiveness of our framework in learning ITEs from time-evolving networked observational data, we compare our framework with multiple state-of-the-art methods:

- **Causal Forest (CF)** [36]. Based on the strong ignorability assumption [27], CF learns ITE as an extension of Breiman's random forest [4]. We set the number of trees as 100.
- **Bayesian Additive Regression Trees (BART)** [13]. BART is a Bayesian regression tree based ensemble model that is widely used in learning ITE. It is also based on the strong ignorability assumption.
- **Counterfactual Regression (CFR)** [31]. CFR also learns representation for the confounders based on the strong ignorability assumption. Balancing techniques including Wasserstein-1 distance and maximum mean discrepancy are adopted and we refer these two variants as CFR-Wass and CFR-MMD.
- **Causal Effect Variational Autoencoder (CEVAE)** [19]. CEVAE is a deep latent-variable model for learning ITE, which learns representation of confounders as Gaussian distributions through propagating information from original features, observed treatments, and factual outcome.

[3] https://www.blogcatalog.com/
[4] https://github.com/allenai/PeerRead

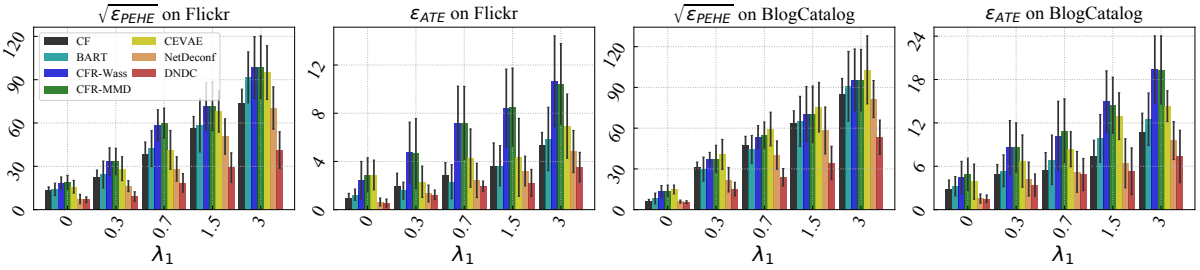


Figure 3: Performance comparison under different settings of historical information influence.

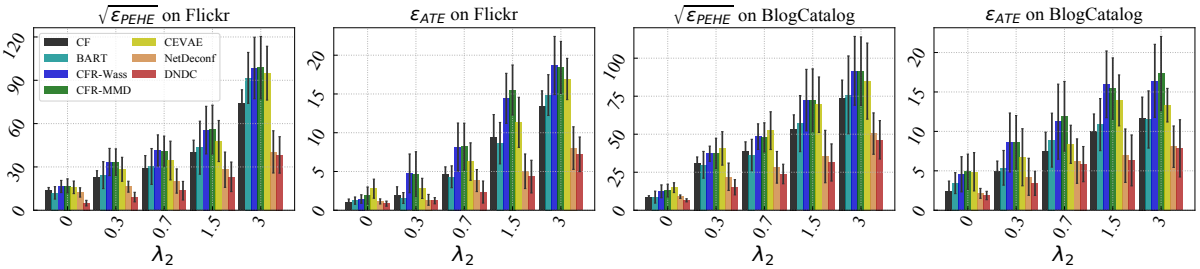


Figure 4: Performance comparison under different settings of network structure influence.

- **Network Deconfounder (NetDeconf)** [9]. NetDconf relaxes the strong ignorability assumption by assuming that the hidden confounders of observational data can be controlled with auxiliary relational information among data.

Setup. All the aforementioned baselines are designed for static data and thus we run these algorithms at each time stamp independently. On the other hand, only our proposed *DNDC* can well capture the temporal dependency for ITE estimation. The data instances (nodes) are randomly split into 60-20-20% of training/validation/test data. We evaluate the average $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} over all the time stamps, and all the results are averaged over 10-time repeated executions. Unless otherwise specified, we set our learning rate as $5e{-}4$, $d_h = 64$, $d_z = 64$, $\beta = 1.0$, $\gamma = 0.01$, and we use Adam as our optimizer. For all the baselines based on confounder representation learning such as CEVAE and NetDeconf, we also set the dimension of the learnt representation as d_z, same as our proposed method. As described in the Section 5.1.2, in the experiments, we use three hyperparameters λ_1, λ_2, and λ_3 to control the influence of the historical information, current network structure, and current feature information on the current confounders, respectively.

5.4 ITE Estimation by Varying the Influence of Historical Information

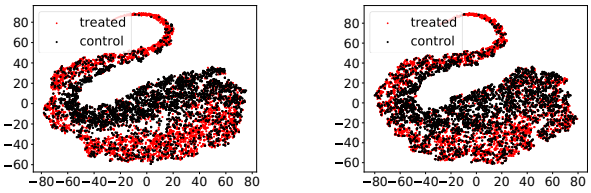
Here we investigate how *DNDC* performs when the level of influence from historical information (including previous hidden confounders and treatment assignment) varies. We fix other parameters $\lambda_2 = 0.3$ and $\lambda_3 = 0.3$, and modify λ_1 in Eq. (11) (in Section 5.1.2) to control the influence of the historical information on the hidden confounders at the current time stamp. We compare the performance of ITE prediction of *DNDC* and other baselines. Due to the space limit, we only show the results on Flickr and BlogCatalog in Fig. 3 as we have similar observations on PeerRead. Generally speaking, we observe that our proposed *DNDC* consistently outperforms all the baselines with lower $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE}. When $\lambda_1 = 0$, the historical information will not affect the current hidden confounders. In this case, our model and NetDeconf achieve the best performance by utilizing the auxiliary relational information among data to approximate the hidden confounders. When λ_1 increases, the current ITE inference relies more on the historical information, while other baselines are not able to catch the historical influence on ITE at the current time stamp. Thus their performance drops dramatically, while the performance of our proposed *DNDC* is stably better as it leverages the historical information. One-tail Student's t-tests are also conducted to confirm that our method performs significantly better than other baselines with a significant level of 0.05.

5.5 ITE Estimation by Varying the Influence of Network Structure

To verify the effectiveness of the proposed *DNDC* in utilizing the relational information embedded in a network, we fix other parameters $\lambda_1 = 0.3$ and $\lambda_3 = 0.3$, and compare *DNDC* with baselines by varying the values of λ_2. Similarly, we only show the results on Flickr and BlogCatalog to save space. As can be observed from the comparison results shown in Fig. 4, when $\lambda_2 = 0$, the hidden confounders of each instance is independent of others', thus NetDeconf loses its superiority without using the relational information among data. When λ_2 increases, its performance is much less jeopardized compared to that of other baselines due to the increasing impact from network structure on hidden confounders. Causal Forest also achieves good performance because its hypothesis implicitly models the propagation mechanism on the network over time, which to some degree leverages the neighbor features to catch the hidden confounders. Our *DNDC* achieves the best performance, we attribute the improvement to two key factors: 1) when λ_2 is small, *DNDC* can achieve better ITE estimation by capturing the historical influence on the hidden confounders at the current time stamp; 2)

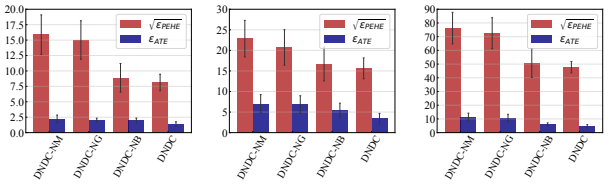
Table 3: Performance comparison with different representation balancing methods.

Dataset		Wass	MMD	Gradient Reverse
Flickr	$\sqrt{\epsilon_{PEHE}}$	9.125 ± 1.566	9.531 ± 1.573	**8.131 ± 1.342**
	ϵ_{ATE}	1.839 ± 0.368	1.952 ± 0.433	**1.413 ± 0.351**
BlogCatalog	$\sqrt{\epsilon_{PEHE}}$	16.115 ± 2.857	17.035 ± 4.243	**15.132 ± 2.542**
	ϵ_{ATE}	4.815 ± 1.367	5.250 ± 1.345	**3.414 ± 1.272**
PeerRead	$\sqrt{\epsilon_{PEHE}}$	49.062 ± 4.452	49.643 ± 4.834	**47.716 ± 4.014**
	ϵ_{ATE}	5.482 ± 1.347	5.648 ± 1.617	**4.451 ± 1.379**



(a) No balancing **(b) With balancing**

Figure 5: Representation distributions with or without gradient reverse layer.



(a) Flickr **(b) BlogCatalog** **(c) PeerRead**

Figure 6: Ablation study for different variants of *DNDC*.

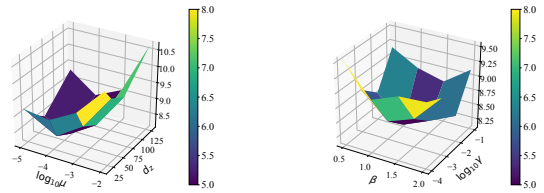
when λ_2 increases, the confounder representation learning component captures the series of information hidden in the network structure, which leads to a more accurate and stable ITE estimation.

5.6 The Impact of Representation Balancing

To evaluate the impact of the proposed adversarial learning based representation balancing method, we compare the performance of our balancing method with other two commonly used representation balancing methods: Wasserstein-1 (Wass) distance [29] and maximum mean discrepancy (MMD) [31]. Table 3 shows the results of ITE estimation performance with these different representation balancing techniques and our method consistently outperforms other baselines. Fig. 5 shows a specific example of the representation distributions with/without the gradient reverse layer. We observe that with the gradient reverse layer, the representation distributions of treated and control group become closer.

5.7 Ablation Study

To further investigate the impact of different components of *DNDC*, we conduct ablation study by comparing *DNDC* against the following variants: (1) *No GRU*: This variant omits the GRU and attention layers, which means that no historical information is utilized in learning the confounders. As this variant does not benefit from the memory of previous time stamps, we denote it by *DNDC-NM*. (2) *No GCNs*: In this variant, we replace the GCN layers with a simple MLP. We denote this variant by *DNDC-NG*. (3) *No balancing*: This variant does not use any representation balancing techniques and is denoted by *DNDC-NB*. Fig. 6 reports the ITE estimation performance of different variants of our proposed framework. We can see that *DNDC-NM* and *DNDC-NG* cannot render satisfactory performance as they cannot leverage historical information or network structure for learning representations of hidden confounders. The performance of *DNDC-NB* is degraded by the imbalance of distributions between the treated and the control group, while *DNDC* performs better with balancing method because the distribution balancing helps mitigate the confounding bias. In short, all three components contribute to the superior performance of *DNDC*.



(a) μ and d_z **(b) β and γ**

Figure 7: $\sqrt{\epsilon_{PEHE}}$ with different values of learning rate μ, embedding size d_z, β and γ.

5.8 Hyperparameter Study

To investigate how the values of model hyperparameters affect the performance of *DNDC*, we assess its performance under different settings of the learning rate $\mu \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, representation dimension $d_z \in \{16, 32, 64, 128\}$, $\beta \in \{0.5, 1.0, 1.5, 2.0\}$ and $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Leaving out all the similar observations, we only show the ITE estimation performance $\sqrt{\epsilon_{PEHE}}$ with different values of learning rates μ, embedding size d_z, β and γ on Flickr in Fig. 7, where we observe that, the model achieves the best performance when μ is around 10^{-4}, $d_z \in [32, 64]$, $\beta = 1.0$ and $\gamma = 0.01$. Similar observations can be observed on other datasets, as well as ϵ_{ATE}. Generally speaking, our model is not very sensitive to the model parameters in a wide range.

6 RELATED WORK

Treatment effect learning for static i.i.d data. Most observational studies [4, 13] focus on i.i.d data in the static setting. Specifically, recent years witnessed a surge of research interests in estimating ITEs by representation learning [39]. Among them, [31] shows that balancing the representation distribution of the treated and control group can help upper bound the error of counterfactual outcome estimation. However, these methods have two limitations: 1) they rely on the strong ignorability assumption and ignore the influence of hidden confounders; 2) they fail to utilize the evolving process of the observed variables and the relations among individuals. [19] relaxes the strong ignorability assumption by taking observed features as proxy variables of hidden confounders using variational inference, but still limited for static i.i.d data.

Treatment effect learning from networked data. Instead of making the strong ignorability assumption, [17, 23] theoretically show that observed proxy variables can be exploited to capture the hidden confounders and estimate the treatment effect. Recently,

[35] starts to utilize the network information as proxy variables to mitigate the confounding bias, but this work has some limitations: (1) it only considers network structure without leveraging features of instances; (2) it relies on the doubly robust estimator which can only estimate average treatment effect (ATE) over the whole population. Recently, [8, 9] propose to unravel patterns of hidden confounders from the network structure along with the observed features by learning representations of hidden confounders, and use the representation for potential outcome prediction. Nonetheless, these works focus on a static setting and are unable to provide accurate ITE estimation over time when the data is evolving.

7 CONCLUSION

In this paper, we study a problem of ITE estimation from networked observational data in a dynamic environment and develop a novel framework *DNDC*. We specify this framework by learning representations of hidden confounders over time for potential outcome prediction and treatment prediction. Additionally, we also incorporate a novel adversarial learning based balancing technique into *DNDC* toward unbiased ITE estimation. The proposed framework is evaluated on multiple semi-synthetic datasets extended from real-world data of attributed networks. Extensive experimental evaluations demonstrate the superiority of our framework over existing ITE estimation frameworks. The future work may include exploiting more available models [12, 16], as well as expanding the framework to more general settings such as multiple treatments [37] and heterogeneous data [40].

ACKNOWLEDGEMENTS

This material is, in part, supported by the National Science Foundation (NSF) under grants #2006844, #2008208, #1934600, #1938167 and #1955151, and COVID-19 Rapid Response grant from UVA Global Infectious Diseases Institute.

REFERENCES

[1] Andrew Anglemyer, Hacsi T Horvath, and Lisa Bero. 2014. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews* 4 (2014).
[2] Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. 2020. Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*.
[3] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *the ACM Conference on Recommender Systems*.
[4] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001).
[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17, 1 (2016).
[7] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020).
[8] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Counterfactual evaluation of treatment assignment functions with networked observational data. In *SIAM International Conference on Data Mining*.
[9] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *ACM International Conference on Web Search and Data Mining*.
[10] Ruocheng Guo, Yichuan Li, Jundong Li, K Selçuk Candan, Adrienne Raglin, and Huan Liu. 2020. IGNITE: A minimax game toward learning individual treatment effects from networked observational data. In *International Joint Conferences on Artificial Intelligence*.
[11] Jan-Eric Gustafsson. 2013. Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement* 24, 3 (2013).
[12] Ehsan Hajiramezanali, Arman Hasanzadeh, Krishna Narayanan, Nick Duffield, Mingyuan Zhou, and Xiaoning Qian. 2019. Variational graph recurrent neural networks. In *Advances in Neural Information Processing Systems*.
[13] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011).
[14] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning*.
[15] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[16] Srijan Kumar, Xikun Zhang, and Jure Leskovec. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
[17] Manabu Kuroki and Judea Pearl. 2014. Measurement bias and effect restoration in causal inference. *Biometrika* 101, 2 (2014).
[18] Jundong Li, Harsh Dani, Xia Hu, Jiliang Tang, Yi Chang, and Huan Liu. 2017. Attributed network embedding for learning in a dynamic environment. In *ACM International Conference on Information and Knowledge Management*.
[19] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*.
[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).
[21] Terence C Mills and Terence C Mills. 1991. *Time series techniques for economists*.
[22] Jersey Neyman. 1923. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* 10 (1923).
[23] Judea Pearl. 2012. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504* (2012).
[24] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3 (2009).
[25] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155, 2 (2000).
[26] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In *ACM International Conference on Information and Knowledge Management*.
[27] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983).
[28] Donald B Rubin. 2005. Bayesian inference for causal effects. *Handbook of Statistics* 25 (2005).
[29] Ludger Rüschendorf. 1985. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields* 70, 1 (1985).
[30] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).
[31] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*.
[32] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. 2015. Causal inference via sparse additive models with application to online advertising. In *AAAI Conference on Artificial Intelligence*.
[33] Panos Toulis, Alexander Volfovsky, and Edoardo M Airoldi. 2018. Propensity score methodology in the presence of network entanglement between treatments. *arXiv preprint arXiv:1801.07310* (2018).
[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.
[35] Victor Veitch, Dhanya Sridhar, and David M Blei. 2019. Using text embeddings for causal inference. *arXiv preprint arXiv:1905.12741* (2019).
[36] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018).
[37] Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *J. Amer. Statist. Assoc.* 114, 528 (2019).
[38] Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30, 1 (2005).
[39] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*.
[40] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. 2017. Causal discovery from nonstationary/heterogeneous data: skeleton estimation and orientation determination. In *International Joint Conference on Artificial Intelligence*.