

A Look into Causal Effects under Entangled Treatment in Graphs: Investigating the Impact of Contact on MRSA Infection

Jing Ma
University of Virginia
jm3mr@virginia.edu

Ritwick Mishra
University of Virginia
mbc7bu@virginia.edu

Chen Chen
University of Virginia
zrh6du@virginia.edu

Gregory Madden
University of Virginia
grm7q@hscmail.mcc.virginia.edu

Anil Vullikanti
University of Virginia
vsakumar@virginia.edu

Daniel Borrajo
J.P. Morgan AI Research
daniel.borrajo@jpmchase.com

Jundong Li
University of Virginia
jundong@virginia.edu

ABSTRACT

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a type of bacteria resistant to certain antibiotics, making it difficult to prevent MRSA infections. Among decades of efforts to conquer infectious diseases caused by MRSA, many studies have been proposed to estimate the causal effects of close contact (treatment) on MRSA infection (outcome) from observational data. In this problem, the treatment assignment mechanism plays a key role as it determines the patterns of missing counterfactuals — the fundamental challenge of causal effect estimation. Most existing observational studies for causal effect learning assume that the treatment is assigned individually for each unit. However, on many occasions, the treatments are pairwise assigned for units that are connected in graphs, i.e., the treatments of different units are entangled. Neglecting the entangled treatments can impede the causal effect estimation. In this paper, we study the problem of causal effect estimation with treatment entangled in a graph. Despite a few explorations for entangled treatments, this problem still remains challenging due to the following challenges: (1) the entanglement brings difficulties in modeling and leveraging the unknown treatment assignment mechanism; (2) there may exist hidden confounders which lead to confounding biases in causal effect estimation; (3) the observational data is often time-varying. To tackle these challenges, we propose a novel method NEAT, which explicitly leverages the graph structure to model the treatment assignment mechanism, and mitigates confounding biases based on the treatment assignment modeling. We also extend our method into a dynamic setting to handle time-varying observational data. Experiments on both synthetic datasets and a real-world MRSA dataset validate the effectiveness of the proposed method, and provide insights for future applications.

CCS CONCEPTS

- **Applied computing** → Law, social and behavioral sciences;
- **Mathematics of computing** → Causal networks.



This work is licensed under a Creative Commons Attribution International 4.0 License.

KDD '23, August 6–10, 2023, Long Beach, CA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0103-0/23/08.
<https://doi.org/10.1145/3580305.3599763>

KEYWORDS

Causal Inference; Graph; Network; Entangled Treatment; Instrumental Variable

ACM Reference Format:

Jing Ma, Chen Chen, Anil Vullikanti, Ritwick Mishra, Gregory Madden, Daniel Borrajo, and Jundong Li. 2023. A Look into Causal Effects under Entangled Treatment in Graphs: Investigating the Impact of Contact on MRSA Infection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, August 6–10, 2023, Long Beach, CA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3580305.3599763>

1 INTRODUCTION

In the past a few decades, a burgeoning body of studies [13, 21, 42] have been proposed for preventing infectious diseases such as Methicillin-resistant *Staphylococcus aureus* (MRSA). MRSA is a type of bacteria that is resistant to antibiotics, including methicillin and other penicillins. It can cause infections in the skin, respiratory tract, and urinary tract and can be spread through close contact with infected individuals or contaminated surfaces. In these scenarios, in-person contact relations are crucial for MRSA-related studies, and graphs are naturally used for modeling these relations. An important question that medical specialists are interested in is: “What is the causal effect of close contact (*treatment*) on the spread of MRSA (*outcome*) in a room-sharing network?” Inspiringly, an emerging field that aims to investigate causal effects rather than the statistical correlations between variables in graph data has attracted arising attention recently [12, 26]. In general, causal effect learning [20, 33] aims to estimate the causal effect of a certain treatment on an outcome for different units. On graph data, causal effect learning has great potential in many real-world applications such as epidemiology [10, 25]. The progress in this area provides us with effective tools for investigating contact impact on MRSA infection.

As discussed in [20], the fundamental challenge of causal effect learning is data missing—only one potential outcome (the one that corresponds to the treatment assignment) can be observed for each unit. For example, for a patient with frequent physical contact with others, the potential outcome for this individual with infrequent contact (i.e., counterfactual) is unavailable. As the treatment assignment mechanism (i.e., how the treatment is assigned to different

units) determines which part of the data is missing, treatment assignment plays an essential role in causal studies. Currently, most existing studies are based on the individualistic treatment assignment [20], where the treatment is assigned individually for each unit. However, in graphs, the treatment is often assigned in a pairwise manner to units that are connected. For example, the in-room contact in a room-sharing network is often not individually applied to each person. Instead, it often happens between a pair of people. In these scenarios, treatments are not individually applied to each unit (i.e., treatments cannot be determined only by each unit's own properties). This setting is referred to as *entangled treatment* [44]. Motivated by these scenarios, in this work, we study the problem of causal effect learning in graphs under entangled treatment.

A few previous works [44, 45] have made preliminary explorations of this problem, but many challenges remain unaddressed: 1) As discussed in [44], treatment entanglement increases the risk of misspecification of the treatment effect estimator. If the entanglement through the graph is not considered, causal effect estimators tend to incorrectly attribute the observed treatment assignments to each unit's individual properties, and thus degrade the performance of causal effect estimation. To handle this entanglement problem, existing works [44] assume that the treatment assignment is determined by a pre-determined function over the graph (e.g., the treatment can be the node degree on the graph). However, on many occasions, this function is unknown. 2) Existing works [44, 45] rely on the unconfoundedness assumption [36] (or its weaker version) that there do not exist unobserved confounders (confounders are variables which causally influence both the treatment and the outcome. For example, patients' behavior habits are hidden confounders that influence their physical contact and infection risk. However, hidden confounders often exist in the real world and could lead to confounding biases. 3) Existing works are often limited to a static setting. However, the graph, treatment, outcome, and unit covariates are naturally dynamic in many real-world scenarios. For example, the patient data is evolving over time; the causal association across different timestamps also brings more difficulties in learning causal effects.

To address the aforementioned challenges, in this paper, we propose a novel framework NEAT to estimate causal effects under Network EntAngled Treatments. Specifically: 1) To handle the entangled treatment, for each node, we explicitly leverage its relevant graph topology to model the unknown treatment assignment with a learnable neural network module. 2) To tackle the hidden confounders, we take the graph structure regarding each node as an instrumental variable (IV) [16]. IV can eliminate the biases brought by hidden confounders in causal effect estimation. In the previous example, the room-sharing network is a valid IV if it is assumed to be independent of the patient's behavior habits, and its influence on the MRSA infection is fully mediated by the physical contact. A valid IV can provide a sort of randomization in the process of causal effect estimation and improve the estimation performance. 3) To learn causal effects in a dynamic setting, we generalize the setting and develop our framework to handle this problem across multiple timestamps.

Notice that our work differs from other two areas of causal effect learning on graphs: 1) interference: these works [27, 28] assume that the treatment of each unit could causally affect the outcome of

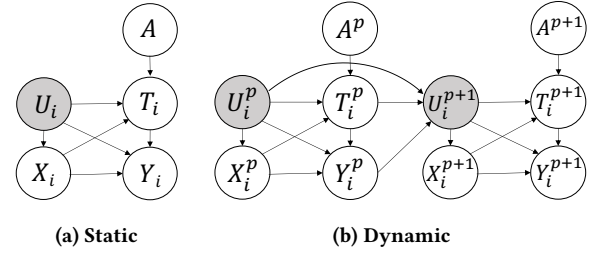


Figure 1: The causal graph of the studied problem in a static setting and in a dynamic setting. Each vertex in the causal graph represents a variable, and each arrow $A \rightarrow B$ represents a causal relation from A to B . The observable variables are shown in white while the unobserved ones are shown in grey.

other units; 2) network deconfounding [12, 26]: these works assume that hidden confounders are buried in the graph structure. These two lines of work and our paper study separate research problems with different assumptions and application scenarios. In this work, our contributions can be summarized as follows:

- **Problem.** Motivated by the MRSA clinical studies, we investigate the important problem of causal effect estimation under entangled treatment. We address the challenges of treatment entanglement, hidden confounders, and a time-evolving environment. To the best of our knowledge, this is the first work addressing these challenges of this problem.
- **Method.** We propose a novel method, NEAT, to address this problem. NEAT estimates causal effects with treatments entangled through a graph. This method leverages the graph topology w.r.t. each node to better model the treatment assignment and facilitate treatment effect estimation even with hidden confounders. This method works for both static and dynamic settings.
- **Experiment.** We conduct extensive experiments to evaluate our method on both synthetic and real-world graphs. Especially, we include real-world clinical data for MRSA infection. The results validate the effectiveness of our proposed method in different aspects.

2 PRELIMINARIES

2.1 Notations and Definitions

The observational data is denoted by $\{X, A, T, Y\}^{1, \dots, P}$, which corresponds to the node features (e.g., patients' covariates), graph adjacency matrices (e.g., room-sharing network), treatment assignments (e.g., close contact), and observed outcomes (e.g., MRSA test result), respectively, in P timestamps. We use $(\cdot)^p$ to denote the data in the p -th timestamp. When we focus on a static setting or a single timestamp, we drop this superscript for notation simplicity. We assume there are N units (nodes) with d_x covariates, with $X^p = \{X_i^p\}_{i \in [N]}$, and for each unit i , $X_i \in \mathbb{R}^{d_x}$. The graph structure connecting these units at each timestamp is an $N \times N$ binary matrix $A^p = \{A_{i,j}^p\}_{i,j \in [N]}$, where $A_{i,j}^p = 1$ when there is an edge from node i to node j , otherwise $A_{i,j}^p = 0$. The treatment is $T^p = \{T_i^p\}_{i \in [N]}$. In most studies, treatment is assumed to be a

binary value, but in this work, we allow it to be a d_t -size vector (e.g., a vector that describes patients' close contact patterns). The observed outcomes are denoted by $Y^P = \{Y_i^P\}_{i \in [N]}$. For each unit i at timestamp p , $Y_i^P \in \mathbb{R}$. In this paper, we use bold letters (e.g., X^P) to denote variables for all units, and use unbold letters (e.g., X_i^P) to denote variables for a single unit. For simplicity, we use the same notation for both variables and data. The subscript $(\cdot)_i$ denotes the index of a unit. If it is not necessary to emphasize the index of a specific unit, we drop the subscript to denote a random unit. The causal graph for this study is shown in Fig. 1; in this case, not all the confounders can be directly observed or measured, thus they can often lead to biased treatment effect estimation. The hidden confounders are denoted by $U^P = \{U_i\}_{i \in [N]}^P$.

This work is based on the well-known Neyman-Rubin potential outcome framework [37]. The potential outcome is defined as the outcome which would have been realized when the treatment assignment had been set to a certain value. We denote the potential outcomes under treatment $T = t$ as $Y^P(t) = \{Y_i(t)\}_{i \in [N]}^P$. Consider a baseline treatment as $T = t_0$, for a treatment $T = t$, the treatment effect conditioned on covariates X in a static setting is defined as:

$$\tau(X) = \mathbb{E}[Y_i(t) - Y_i(t_0)|X]. \quad (1)$$

In a dynamic setting, we denote the historical information before timestamp p as $M^P = \{X, A, T, Y\}^{1, \dots, p-1}$. Similar to the above, we denote the historical information regarding unit i before timestamp p as M_i^P . When estimate causal effects at timestamp p , only the data no later than timestamp p can be used. We define the treatment effect at timestamp p as:

$$\tau(X^P, M^P) = \mathbb{E}[Y_i^P(t) - Y_i^P(t_0)|X^P, M^P]. \quad (2)$$

Similar as [38], we define the treatment effect for each unit i at timestamp p as $\tau_i^P = \tau(X_i^P, M^P)$.

We define the entangled treatment as follows:

DEFINITION 2.1. (Entangled treatment) *The treatment here can be a function $\mathcal{T}(\cdot)$ over the graph structure, the observed features, and the hidden confounders:*

$$T = \mathcal{T}(A, X, U). \quad (3)$$

In a dynamic setting, the treatment is also a function over historical information

$$T^P = \mathcal{T}(A^P, X^P, U^P, M^P). \quad (4)$$

Notice that as the treatment function has the graph structure as an input, the treatments across different units are no longer individualistic (i.e., T_i cannot be determined only based on variables of unit i). A typical example of the treatment function $\mathcal{T}(\cdot)$ is the degree of each node. But under many real-world circumstances, $\mathcal{T}(\cdot)$ is an unknown function.

The problem we study in this work is formally defined as:

DEFINITION 2.2. (Causal effect estimation under entangled treatments) *Given the observational data $\{X, A, T, Y\}^{1, \dots, P}$, we aim to estimate the treatment effect $\tau(X^P, M^P)$ for different units at each timestamp p with treatments entangled in the graph.*

2.2 Assumptions

We assume that the outcome is generated by treatment, features, historical information, and hidden confounders as follows:

$$Y^P = \mathcal{Y}(T^P, X^P, M^P) + g(U^P), \quad (5)$$

where $\mathcal{Y}$ and g are unknown and (nonlinear) functions. We assume $E[g(U^P)] = 0$. In this work, we take the graph structure as an instrumental variable for IV analysis. An implicit assumption of our work is that the graph information of each node i can be represented as a variable A_i , and its samples in observational data are sufficient for us to capture the patterns it influences the treatment assignment. The following assumptions make the graph structure as a valid IV.

Assumption 1. (Relevance) *Given X^P, M^P for any random unit, the treatment is relevant to the graph structure, i.e., $A^P \not\perp\!\!\!\perp T^P|X^P, M^P$.*

Assumption 2. (Exclusion restriction) *For any random unit, the causal effect of A^P on Y^P is fully mediated by T^P , i.e., $Y^P(T, A) = Y^P(T, A')$, $\forall A \neq A'$. Here, $Y^P(T, A)$ denotes the potential outcome for treatment T and graph A at timestamp p .*

Assumption 3. (Instrumental unconfoundedness) *There is no unblocked backdoor path from A^P to Y^P , i.e., $A^P \perp\!\!\!\perp Y^P(A)|X^P, M^P$ for any random unit. Here, $Y^P(A)$ denotes the potential outcome for graph A at timestamp p .*

Inspired by recent IV studies [16], we use the above assumptions to effectively leverage the graph structure as an IV for treatment effect estimation. More analysis can be found in Appendix A.

3 THE PROPOSED FRAMEWORK

In this section, we introduce the proposed NEAT framework for causal effect learning under entangled treatment on the graph. Fig. 2 shows an illustration of the proposed framework. Specifically, this framework contains three modules: node representation learning, entangled treatment modeling, and outcome prediction.

3.1 Overall Pipeline

The whole framework is designed in a classical two-stage IV study pipeline [1, 16]. Generally, in this pipeline, the first stage predicts the treatment with IVs, and the second stage estimates the potential outcomes based on the treatment predicted by the first stage. The key intuition behind this design is that, as the IVs are unconfounded, the predicted treatment from the first stage can provide more randomization, and thus it can help mitigate the confounding bias brought by hidden confounders.

In our framework, in the first stage, we train a treatment modeling module to predict treatment assignments for each node at each timestamp. In this module, we leverage the graph structure as an IV, and utilize it to capture the patterns of entangled treatment in the graph. Simultaneously, we learn a representation for each node to encode its properties, including its current features and historical information. In the second stage, we predict potential outcomes based on the original node features, the learned node representations, and the predicted treatment. In this two-stage IV framework, the biases brought by hidden confounders can be effectively eliminated.

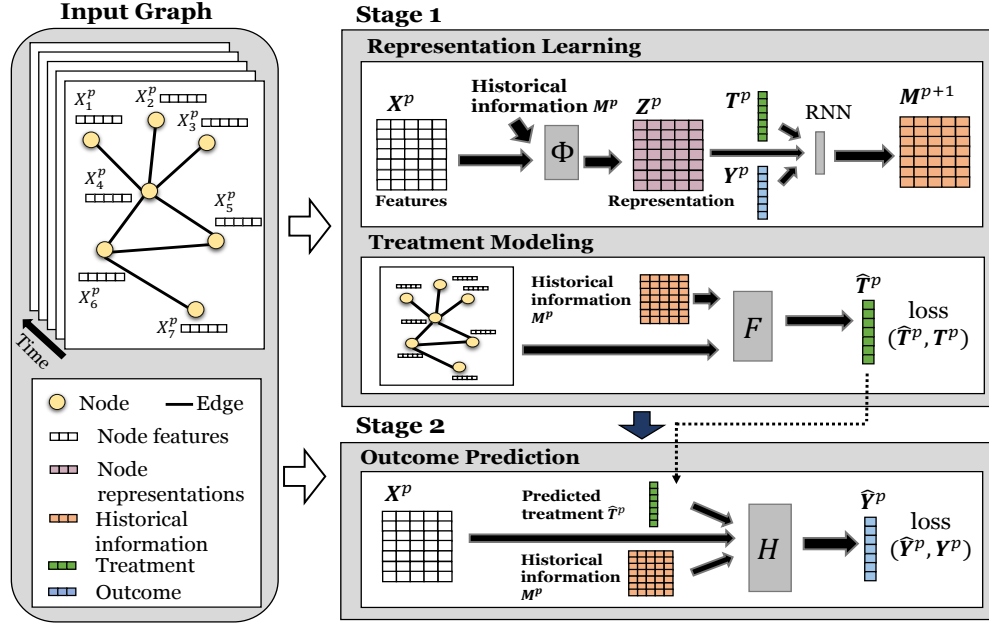


Figure 2: The proposed framework NEAT. It contains three components: node representation learning, entangled treatment modeling, and outcome prediction.

3.2 Node Representation Learning

The treatment effects are often different for nodes with different properties. For example, close contact may influence patients of different ages differently. To model such heterogeneity, we capture the properties of each node through node representation learning. For each node i , we learn a representation Z_i to encode its properties based on its node features X_i :

$$Z_i = \phi(X_i). \quad (6)$$

Here, $\phi(\cdot)$ is implemented by a neural network module with learnable parameters.

Dynamic setting. In a time-evolving environment, as illustrated in Fig. 1 (b), the current properties of each node can be influenced by the historical data in previous timestamps. To capture the time-evolving properties and model the causal mechanism in a dynamic setting, for each node i , we embed the historical information before each timestamp p into a representation M_i^p with a recurrent neural network (RNN) [7, 18]. M_i^p is then incorporated into Z_i^p . At each timestamp, we update the historical embedding as:

$$M_i^p = \text{RNN}(M_i^{p-1}, T_i^{p-1}, Y_i^{p-1}, Z_i^{p-1}, X_i^{p-1}). \quad (7)$$

Here, we learn the representation for each node i at timestamp p with a transformation function $\Phi(\cdot)$:

$$Z_i^p = \Phi(X_i^p, M_i^p). \quad (8)$$

3.3 Entangled Treatment Modeling

The treatment function $\mathcal{T}(\cdot)$ in Eq. (3) or Eq. (4) is often not pre-determined. To better estimate treatment effects from observational data, we capture the treatment assignment patterns by training a module $F(\cdot)$ to model the conditional distribution of treatment T_i^p

given A^p, X_i^p, M_i^p . The treatment modeling module is trained in the first stage together with node representation learning:

$$\hat{T}_i^p = F(A^p, X_i^p, M_i^p) = f(A^p, \Phi(X_i^p, M_i^p)). \quad (9)$$

Treatment Entanglement. As the treatments of different units are entangled through the graph structure, to effectively capture the patterns of treatment assignment, we explicitly leverage the graph structure in the treatment modeling module. As a feasible implementation, we design this module $F(\cdot)$ based on graph neural networks (GNNs) [22, 46]. Here we use one-layer graph convolutional network (GCN) [22] to predict the treatment as follows:

$$\hat{T}_i^p = \sigma(\hat{A}^p([X^p, Z^p])W_0), \quad (10)$$

where $\sigma(\cdot)$ is an activation function such as Softmax. $\hat{A}^p$ is the normalized adjacency matrix calculated from the graph A^p beforehand with the renormalization trick [22]. Here $[\cdot, \cdot]$ stands for the concatenation operation. W_0 denotes the parameters in GCNs.

Loss for treatment modeling. The loss for treatment prediction is denoted by $\mathcal{L}_t$. Generally, $\mathcal{L}_t$ is defined as:

$$\mathcal{L}_t = \sum_{p=1}^P \sum_{i=1}^N l_t(\hat{T}_i^p, T_i^p) = \sum_{p=1}^P \sum_{i=1}^N l_t(F(A^p, X_i^p, M_i^p), T_i^p), \quad (11)$$

where $l_t(\cdot)$ is a loss term to measure the prediction error of treatment modeling. Noticeably, in this work, we do not restrict the data type of treatment. To handle different types of treatment, we design a different implementation for this module. More specifically, for discrete treatments (e.g., whether a patient has frequent close contact), we implement treatment prediction $F(\cdot)$ as a classification model with the cross-entropy loss function; for continuous treatments (e.g., values that describe the patient's contact patterns) we

implement this module as a prediction task with mean square error (MSE) loss.

3.4 Outcome Prediction

We train an outcome prediction module $H(\cdot)$ in the second stage, which predicts Y_i^p based on M_i^p, X_i^p , and $\hat{T}_i^p$:

$$\hat{Y}_i^p = \int H(\hat{T}_i^p, M_i^p, X_i^p) dF(\hat{T}_i^p | A^p, X_i^p, M_i^p). \quad (12)$$

We denote the loss function for outcome prediction by:

$$\mathcal{L}_y = \sum_{p=1}^P \sum_{i=1}^N l_y(\hat{Y}_i^p, Y_i^p), \quad (13)$$

where $l_y(\cdot)$ is a loss function (e.g., MSE) to measure the prediction error of the outcome. For each node i , the potential outcome w.r.t. treatment $T = t$ is predicted by $\hat{Y}_i(t) = H(t, M_i, X_i)$. We thereby estimate the treatment effect for each node i as:

$$\hat{\tau}_i = \hat{Y}_i(t) - \hat{Y}_i(t_0). \quad (14)$$

3.5 Implementation Details

In node representation learning, we implement $\Phi(\cdot)$ with a multi-layer perceptron (MLP) and use a Gated Recurrent Unit (GRU) [7] for RNN. In entangled treatment modeling, we implement $F(\cdot)$ with a GCN layer. For discrete treatments, we use Softmax as the final layer, and take the output logits to model the probability of treatment values. For continuous treatments, we model them with a mixture of Gaussian distribution with component weights $w_k(A^p, X^p, M^p)$ and parameters $(\mu_k(A^p, X^p, M^p), \sigma_k(A^p, X^p, M^p))$ for each component k . In outcome prediction, we use an MLP module to implement $H(\cdot)$, and use MSE loss for $l_y(\cdot)$. We use two optimizers to train the first and the second stage, respectively.

3.6 Discussion

Many graph learning techniques (e.g., GCNs) mainly focus on local graph information (generally, k -layer GCNs can handle neighbors within k hops), but if the treatment assignment is affected by a wider range on the graph (e.g., the length of the longest path which contains node i), it would be more difficult to capture and handle such information. However, it is worth noting that the proposed framework should not be limited to the specific implementation as introduced above. Instead, we can replace each component with a different implementation to achieve better specifications if relevant prior knowledge is given.

4 EXPERIMENTS

In this section, we validate the effectiveness of our proposed method by conducting extensive evaluations. More specifically, our experiments are designed to answer the following research questions: (1) **RQ1**: How does the proposed framework perform under treatment entanglement compared with state-of-the-art baselines? (2) **RQ2**: How does the proposed framework perform under different levels of treatment entanglement and hidden confounders? (3) **RQ3**: How does each component of the proposed framework contribute to the final treatment effect estimation? (4) **RQ4**: How does the proposed framework perform under different parameter settings?

Table 1: Detailed statistics of the datasets.

Dataset	Random	Transaction	Social	MRSA
# of nodes	30,000	186,509	52,406	11,044
# of edges	208,193	61,572	107,394	31,403
# of features	32	21	16	8
# of timestamps	12	15	10	13

4.1 Dataset and Simulation

In our experiment, we use four datasets with dynamic graph data, including synthetic, semi-synthetic, and real-world data. As it is notoriously hard to obtain the true causal models and counterfactuals from the real world, on the first three datasets, we follow regular practice to evaluate our method on data with simulated causal models. Nevertheless, we encourage our simulation to be as close to reality as possible, thus, our synthetic and semi-synthetic datasets are based on graphs that are generated by real-world relational information and node features. Based on these graph data, we simulate the time-varying hidden confounders, treatment assignments, and outcomes.

4.1.1 Simulation. We describe the way we simulate different variables as follows. More details of simulation are in Appendix B.

Hidden confounders. In a static setting, we simulate the hidden confounders as:

$$U_i \sim \mathcal{N}(0, \mu \mathbf{I}). \quad (15)$$

Here, $\mathbf{I}$ denotes an identity matrix of size d_u (i.e., the dimension of hidden confounders). We set $\mu = 20$ by default.

Features. If the node features are available in the dataset, we directly use them. Otherwise, we simulate them by:

$$X_i = \psi(U_i) + \epsilon_x, \quad (16)$$

where $\psi(\cdot)$ is a linear function $\mathbb{R}^{d_u} \rightarrow \mathbb{R}^{d_x}$. Here, d_x is the dimension of node features. ϵ_x is a noise vector in Gaussian distribution.

Treatment. We simulate the treatment with function $\mathcal{T}$:

$$T_i = \text{BI}((1 - \lambda)\Theta_{t,x}^\top X_i + \lambda \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (\Theta_{t,x}^\top X_j) + \Theta_{t,u}^\top U_i + \epsilon_t), \quad (17)$$

where $\Theta_{t,x}, \Theta_{t,u}$ are parameter vectors with dimension d_x and d_u , respectively. Each parameter in $\Theta_{t,*}$ is in Gaussian distribution $\mathcal{N}(0, 0.5^2)$. $\mathcal{N}_i$ is the set of neighbors of node i in the graph. We use only one-hop neighbors by default. $\lambda \in [0, 1]$ is the parameter that controls the strength of treatment entanglement, i.e., the larger λ is set, the stronger the graph influences the treatment assignments. $\text{BI}(\cdot)$ is a function that maps the input to a binary value. A regular implementation is to transform the input to a probability using a Sigmoid function, and then sample the output with Bernoulli distribution. Noticeably, we do not restrict the treatment to be a binary value. Continuous treatment can be simulated without the $\text{BI}(\cdot)$ function; and high-dimensional treatment with dimension d_t can be simulated by replacing the parameter vector $\Theta_{t,x}$ with a parameter matrix $\Theta_{t,x}$ with dimension $d_x \times d_t$ (similarly for $\Theta_{t,u}$). $\epsilon_t \sim \mathcal{N}(0, 0.01^2)$ is a random Gaussian noise.

Potential outcome. We simulate the potential outcomes as follows:

$$Y_i(t) = t \cdot \Theta_y^\top X_i + \Theta_0^\top X_i + \beta \Theta_u^\top U_i + \epsilon_y, \quad (18)$$

where Θ_y and Θ_0 are parameter vectors of dimension d_x , and Θ_u is of dimension d_u . $\beta \geq 0$ is a parameter that controls the strength of the hidden confounder. $\epsilon_y \sim \mathcal{N}(0, 0.1^2)$ is a noise.

Dynamic setting. In a dynamic setting, we simulate the historical data and hidden confounders over time as:

$$M_i^p = \sum_{r=1}^R (W_u^r U_i^{R-r} + W_x^r X_i^{R-r} + W_t^r T_i^{R-r} + W_y^r Y_i^{R-r}), \quad (19)$$

$$U_i^p = \psi_u(M_i^p) + \epsilon_u \quad (20)$$

where R is the number of previous timestamps which influence the current one. We set $R = 3$ by default. Generally, the historical information at each timestamp encodes the previous hidden confounders, node features, treatments, and outcomes. Parameters W_u^r , W_x^r , W_t^r , and W_y^r control these four types of influence from timestamp $R - r$. We generate time-varying hidden confounders with a transformation over the historical information. Here, $\psi_u(\cdot)$ is a linear transformation function. $\epsilon_u \sim \mathcal{N}(0, I)$ is a noise. We use the same way as Eq. (16) to simulate features. The treatments and outcomes are also generated similarly as above description in Eq. (17) and Eq. (18), but the historical information M_i^p is incorporated by concatenating it with X_i^p as input.

4.1.2 Datasets. We further introduce more details about each dataset used in this paper. More details of data statistics are shown in Table 1, including the number of nodes, edges, features, and timestamps.

Random graph. This dataset contains synthetic graphs generated by the Erdős-Rényi (E-R) model [11] at each timestamp. We use NetworkX [14] to generate these graphs. Based on these graphs, we simulate other variables as described in Section 4.1.1.

Real-world graphs. We use two real-world dynamic graphs with each node representing a real person and each edge representing a certain type of connection between them. Based on the type of connection, these two datasets are referred as **Transaction** and **Social**, respectively. We use the covariates of people in these datasets as node features, and simulate the treatments and outcomes as described in Section 4.1.1. More details of these datasets can be found in Appendix B.

MRSA. This dataset contains real-world hospital data for studying Methicillin-resistant Staphylococcus aureus (MRSA) infection. We construct a dynamic graph for the room-sharing relations between patients. At each timestamp, each node is a patient, and an edge exists between a pair of patients if and only if they have shared at least one room during this timestamp. The patient information such as medicine usage and length of stay are taken as node features. We investigate the causal effect of the number of in-room contacts (treatment) on MRSA infection test results (outcome). We consider there exist hidden confounders such as patients' behavior habits. In this dataset, we do not use any simulated data, and do not evaluate our causal effect estimation based on simulated counterfactuals. Instead, we use the domain knowledge regarding MRSA to confirm our findings.

4.2 Baselines

In the experiments, we compare our method with some state-of-the-art baselines. These baselines can be divided into the following three main categories:

- **Individual units.** These methods are based on the assumption that different units are independent. They estimate the treatment effect by adjusting for confounders based on unit covariates. We adopt the widely-used methods including S-Learner (SL) [23], causal forest (CF) [47], and counterfactual regression (CFR) [39].
- **Network deconfounder.** These methods assume that there is a graph connecting different units. They mitigate confounding biases by using the graph structure as a proxy for hidden confounders. We use the network deconfounder (NetDeconf) [12] and the dynamic network deconfounder (DNDC) [26].
- **DeepIV.** This method [16] uses instrumental variables to mitigate the confounding biases. For each node i , we take the i -th row in the adjacency matrix as its IV.

We use the implementation released in the EconML package¹ for S-Learner, causal forest, and DeepIV.

4.3 Evaluation Metrics

We adopt two widely-adopted metrics for treatment effect estimation, including Rooted Precision in Estimation of Heterogeneous Effect (PEHE) [17] and Mean Absolute Error (ATE) [49] at each timestamp p :

$$\sqrt{\epsilon_{PEHE}^p} = \sqrt{\frac{1}{N} \sum_{i \in [N]} (\hat{\tau}_i^p - \tau_i^p)^2}, \quad (21)$$

$$\epsilon_{ATE}^p = \left| \frac{1}{N} \sum_{i \in [N]} \hat{\tau}_i^p - \frac{1}{N} \sum_{i \in [N]} \tau_i^p \right|. \quad (22)$$

For all the experiments, we calculate the average values of these metrics over all timestamps, and still denote them by $\sqrt{\epsilon_{PEHE}}$ and ϵ_{ATE} for simplicity.

4.4 Setup

For all datasets, we randomly split them into 60%/20%/20% training/validation/test data. By default, we focus on the dynamic setting and set the number of training epochs as 2000, the learning rate as 0.004, the dimension for node representation and history embedding as 32 and 20, respectively, $\lambda = 0.5$, $\beta = 0.5$. We report the mean and standard deviation of performance over ten repeated executions on test data. More details of experiment setup are in Appendix B.

4.5 RQ1: Performance of Different Methods

To demonstrate the effectiveness of the proposed method, in Table 2, we show the treatment effect estimation performance of our method and the baselines in both static and dynamic settings. We observe that in both settings, the proposed method NEAT outperforms other baselines in different metrics. We attribute the improvement to two key factors: 1) We explicitly incorporate the graph structure to model the treatment assignment. During this process, we can better utilize the observational data for treatment effect estimation. Among the baselines, SL, CF, and CFR do not consider the graph which connects different units; NetDeconf and DNDC can leverage graph structure, but they use the graph as a proxy to infer the hidden confounders. These methods, however, do not fit in well in

¹<https://github.com/microsoft/EconML>

Table 2: Performance of treatment effect estimation for different methods.

Method	Static						Dynamic					
	Random		Transaction		Social		Random		Transaction		Social	
	$\sqrt{\epsilon_{PEHE}}$	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}
SL	67.2 $\pm$ 3.0	7.3 $\pm$ 0.5	40.9 $\pm$ 1.4	7.1 $\pm$ 0.3	48.3 $\pm$ 2.5	9.2 $\pm$ 0.7	69.4 $\pm$ 3.1	7.7 $\pm$ 0.4	55.8 $\pm$ 1.8	8.4 $\pm$ 0.6	45.3 $\pm$ 1.4	6.5 $\pm$ 0.3
CF	33.7 $\pm$ 2.1	7.0 $\pm$ 0.2	30.9 $\pm$ 1.8	6.9 $\pm$ 0.3	23.6 $\pm$ 1.1	5.9 $\pm$ 0.4	36.2 $\pm$ 2.4	7.4 $\pm$ 0.6	39.6 $\pm$ 1.2	6.2 $\pm$ 0.4	31.4 $\pm$ 1.0	5.8 $\pm$ 0.4
CFR	28.1 $\pm$ 2.4	6.3 $\pm$ 0.5	34.4 $\pm$ 2.3	5.6 $\pm$ 0.9	27.3 $\pm$ 2.0	5.2 $\pm$ 0.5	33.3 $\pm$ 2.7	6.7 $\pm$ 0.4	30.0 $\pm$ 2.6	5.9 $\pm$ 0.4	27.7 $\pm$ 2.2	6.0 $\pm$ 0.5
NetDeconf	35.6 $\pm$ 3.0	6.2 $\pm$ 0.3	28.6 $\pm$ 2.0	5.8 $\pm$ 0.7	30.5 $\pm$ 2.7	6.3 $\pm$ 0.4	34.0 $\pm$ 2.5	6.8 $\pm$ 0.7	29.4 $\pm$ 1.5	6.1 $\pm$ 0.5	32.9 $\pm$ 2.2	5.8 $\pm$ 0.8
DNDC	32.9 $\pm$ 2.4	6.8 $\pm$ 0.3	29.8 $\pm$ 2.2	6.0 $\pm$ 0.5	33.2 $\pm$ 3.1	6.6 $\pm$ 0.7	29.9 $\pm$ 2.2	6.2 $\pm$ 0.4	28.9 $\pm$ 1.8	5.7 $\pm$ 0.5	35.8 $\pm$ 3.0	6.4 $\pm$ 0.7
DeepIV	31.0 $\pm$ 2.3	5.9 $\pm$ 0.4	26.7 $\pm$ 1.9	5.4 $\pm$ 0.6	21.4 $\pm$ 1.6	5.1 $\pm$ 0.3	32.2 $\pm$ 3.1	5.8 $\pm$ 0.5	30.2 $\pm$ 1.9	5.8 $\pm$ 0.4	24.1 $\pm$ 1.8	5.6 $\pm$ 0.6
NEAT	22.4 $\pm$ 1.8	5.2 $\pm$ 0.3	18.8 $\pm$ 1.4	4.6 $\pm$ 0.4	17.9 $\pm$ 1.2	4.1 $\pm$ 0.5	20.1 $\pm$ 1.4	5.0 $\pm$ 0.2	22.5 $\pm$ 1.0	5.3 $\pm$ 0.3	18.2 $\pm$ 1.6	5.0 $\pm$ 0.4

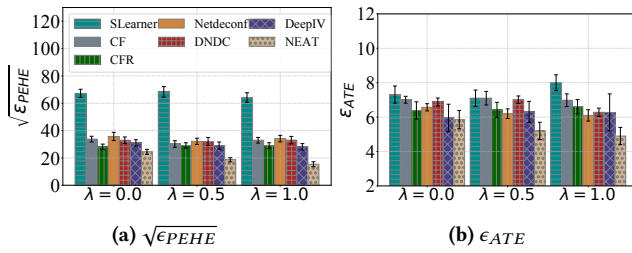


Figure 3: Treatment effect estimation performance under different levels of treatment entanglement on Random dataset.

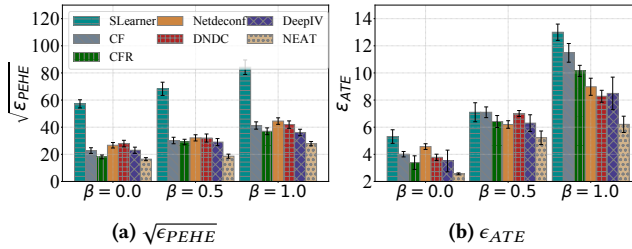


Figure 4: Treatment effect estimation performance under different levels of hidden confounders on Random dataset.

the problem setting studied in this paper. 2) We utilize the graph structure as an instrumental variable to eliminate the confounding biases. Among the baselines, SL, CF, and CFR are based on the unconfoundedness assumption; NetDeconf and DNDC assume the hidden confounders can be inferred from the graph structure. These assumptions cannot be satisfied in our datasets. DeepIV also takes the graph information as an instrumental variable to handle hidden confounders, but its performance is impeded due to the lack of proper techniques to handle graph data.

4.6 RQ2: Performance under Different Levels of Treatment Entanglement and Confounders

To evaluate our method more comprehensively, we test it under different levels of treatment entanglement. In the simulation, we control the treatment entanglement with parameter λ : the larger

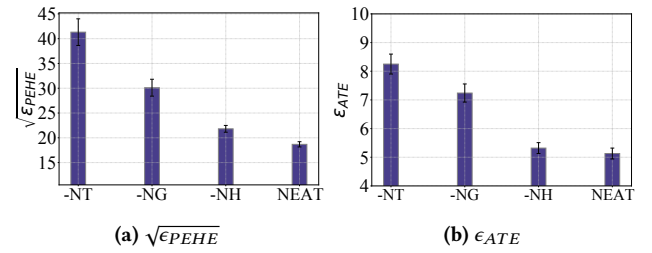


Figure 5: Ablation study for different variants of NEAT on Random dataset.

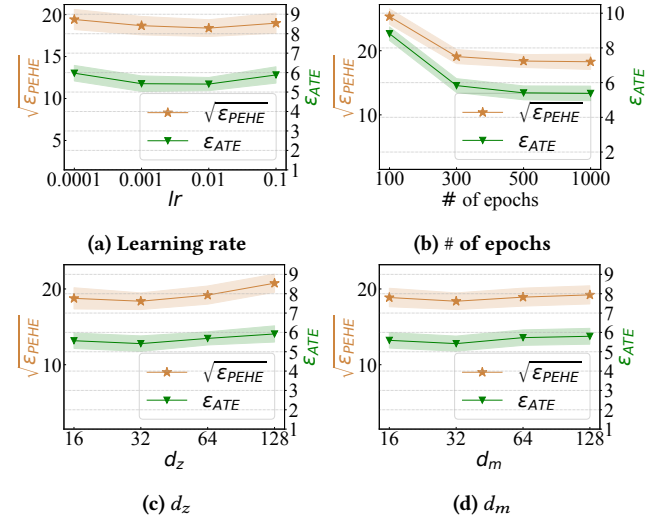


Figure 6: Parameter study for NEAT on Random dataset.

λ is set, the stronger the treatment assignment of each node is entangled with neighbors. Fig. 3 shows the causal effect estimation performance when we set λ as different values. Generally, we observe more obvious performance gain when λ is larger. This observation indicates that our method can well handle the entangled treatments by leveraging the graph structure. We only show the

Table 3: Estimated treatment effect of roommate number on MRSA infection in different populations of patients.

Population	T=0	T=1	T=2
All	0	0.025 ± 0.002	0.082 ± 0.004
General Surgery	0	0.016 ± 0.002	0.058 ± 0.003
Intensive Care	0	0.033 ± 0.003	0.119 ± 0.005
Gerontology	0	0.024 ± 0.002	0.082 ± 0.004

results on the Random dataset, but similar observations can also be found on other datasets.

We also evaluate our method under different levels of hidden confounders. In Fig. 4, we show the results when we change the strength of hidden confounders. Specifically, we change the strength by multiplying the hidden confounders in simulation with the parameter $\beta \geq 0$. From Fig. 4, it can be observed that compared with baselines, our method is more robust with hidden confounders. This is because we effectively utilize the graph as an instrumental variable to mitigate confounding biases.

4.7 RQ3: Ablation Study

To verify the effectiveness of each component in our method, we conduct an ablation study including the following variants: (1) **NEAT-NT**: In this variant, we replace the treatment modeling module with a random sampling over the space of treatment assignment; (2) **NEAT-NG**: In this variant, we do not use the graph in treatment modeling, and replace the input adjacency matrix with an identity matrix. (3) **NEAT-NH**: In this variant, we remove the RNN in our method and do not use historical information. Fig. 5 reports the performance of our method and these variants. The results show that all the different components contribute to the final superior performance of our method.

4.8 RQ4: Parameter Study

To investigate the performance of our proposed method under different parameter settings, we vary the parameters including: learning rate in range of $\{1e-4, 1e-3, 1e-2, 1e-1\}$, number of epochs in the range of $\{100, 300, 500, 1000\}$, node representation dimension $d_z \in \{16, 32, 64, 128\}$, and historical embedding dimension $d_m \in \{16, 32, 64, 128\}$. From the results shown in Fig. 6, we observe that our method is generally not sensitive to parameter setting, but proper choices of parameters still benefit the performance.

4.9 Case Study on Real-world Hospital Data

Methicillin-resistant *Staphylococcus aureus* (MRSA) is a difficult-to-treat pathogen (owing to multi-drug resistance) that is known to spread efficiently within hospitals via contact. One important avenue of hospitalized patient-to-patient MRSA transmission is thought to be through contamination of hospital room surfaces and equipment [29]. In addition, patients may be more or less susceptible to acquiring MRSA given individual factors [40], and MRSA transmission rates may vary according to particular hospital wards (or hospital units) [32].

Table 4: Estimated treatment effect of hospital unit type on MRSA infection.

Hospital Unit Type	Estimated ATE
General Surgery	0 (baseline)
Intensive Care	0.135 ± 0.002
Gerontology	0.138 ± 0.004
Transitional Care	-0.042 ± 0.006
Internal Medicine	0.000 ± 0.002
Cardiology	0.072 ± 0.004
Orthopedic Surgery	0.000 ± 0.001
Gastroenterology	0.000 ± 0.001
Hematology and Oncology	-0.083 ± 0.005

The MRSA dataset contains observational data including patient covariates, room-sharing information, and MRSA test record from a real-world hospital. We construct a room-sharing network, in which an edge connects two patients (nodes) if and only if they have appeared in at least one same room simultaneously. We use our method to investigate the following causal questions: (1) How does the number of in-room contacts causally influence the MRSA infection risk? (2) How do other treatments, such as the type of hospital unit (e.g. Cardiology, Internal Medicine, etc.) causally influence the MRSA infection risk? As the ground-truth causal model is unknown, it is infeasible to evaluate our method on this dataset with the aforementioned metrics. Instead, we show some case studies and verify our key findings with domain knowledge.

For the first question, we map the number of in-room contacts into three levels of treatment. Here, treatments 0, 1, 2 represent the roommate number from low to high. We take $T = 0$ as the control group, and calculate the treatment effect for $T = 1$ and $T = 2$ by comparing the estimated potential outcomes of them with the case of $T = 0$, respectively. Table 3 shows the estimated averaged treatment effect (ATE) of roommate number on MRSA infection over all the patients, and also shows the estimated conditional averaged treatment effect (CATE) conditioned on each subpopulation of patients in a specific group of rooms. From the results, we observe that: 1) In general, the increase in roommate number could result in an increase in MRSA infection risk. This observation holds in the whole population and different subpopulations. As MRSA is contagious through physical contact, this observation is consistent with domain knowledge. 2) The CATE of roommate number on MRSA infection is the strongest in Intensive Care and Gerontology. In Intensive Care, it is frequent for patients to share devices such as ventilators, which leads to a more severe risk of infection when the number of in-room contacts increases. Besides, most patients in Gerontology rooms are older adults with comorbidities associated with MRSA susceptibility (i.e., age >79 , prior nursing home residence, antibiotic exposure, dementia, stroke, or diabetes), which brings a higher risk for acquiring MRSA from the environment with more physical contact [41].

For the second question, we take the hospital unit type as treatment, and show the estimated ATE of each hospital unit type on MRSA infection in Table 4. Here, we take General Surgery as the

baseline treatment (control group). From Table 4, we observe that staying in Intensive Care and Gerontology rooms increases the MRSA infection risk most obviously. The reason might lie in the properties of these units (equipment sharing in the intensive care units, and more MRSA carriers in Gerontology). We also observe a relatively low treatment effect among beds in Transitional Care and Hematology/Oncology units. Most of these rooms are private (as opposed to other semi-private or 2-patient shared rooms), and may lead to less infection risk.

5 RELATED WORK

In this section, we introduce some representative studies related to this work, including causal inference on graph data and instrumental variable analysis.

Causal inference on graph data. Causal inference on graph data has recently attracted arising attention [12, 26, 31, 48, 51]. Under this broad area, the topics which are most related to this work include: 1) *Entangled treatment*: a few initial explorations [44, 45] have been made for entangled treatment. These works discuss the challenges of entangled treatment modeling, and extend the traditional propensity score method for this problem. In our work, we do not limit the method to be propensity score-based, and consider a more general setting of entangled treatment with unknown treatment function, hidden confounders, and dynamic data. 2) *Network deconfounding*: A line of works [12, 26] leverage the graph structure among units to capture the hidden confounders. Netdeconf [12] develops a GCN-based framework to learn the representations of hidden confounders, and adjusts for the confounders on top of the learned representations. DNDC [26] further proposes to learn time-varying confounder representations from observational dynamic graphs. Although we also allow the existence of hidden confounders, our work differs from their application scenarios, as we focus on the setting in which the graph structure is an IV rather than a proxy for confounders. 3) *Network interference*: Traditional causal effect estimation studies are based on the Stable Unit Treatment Value (SUTVA) assumption [34, 35] that the treatment of each unit does not causally affect the outcome of other units (i.e., interference does not exist). However, interference often exists between connected units in graph data [2, 5, 52]. There have been many works [2, 19, 27, 28, 43, 50] addressing the problem of causal inference under interference. Our work differs from them as we do not assume the existence of interference in graphs. Instead, we focus on the case when the graph influences the treatment assignment.

Instrumental variable. Hidden confounders can bring biases in causal effect estimation. Different from most causal inference methods which assume that all the confounders are observed, instrumental variable (IV) based methods provide an alternative approach to identifying causal effects even with the existence of hidden confounders. One of the most well-known lines of IV studies is two-stage methods [1, 8, 16, 30]. The two-stage least squares method (2SLS) [1] is the most representative work in this line, which first fits a linear model to predict treatment with features and IVs, and then fits another linear model to predict the outcome with the features and the predicted treatment. 2SLS is based on two strong assumptions: homogeneity (treatment effect is the same for different units) and linearity (the linear models are correctly specified). There have been many follow-up works to relax these assumptions.

DeepIV [16] is a neural network-based two-stage method that allows nonlinearity and heterogeneity. Another line of IV studies is based on the generalized method of moments (GMM) [15, 24]. Among them, DeepGMM [4] leverages the moment conditions to identify the counterfactual generation function and estimate causal effects. But most of the existing IV studies focus on instrument variables in simple structures, such as scalars and vectors.

6 CONCLUSION

In this paper, motivated from the task of investigating the impact of close contact on MRSA infection in a room-sharing network, we studied the problem of causal effect estimation under entangled treatment. We discussed the related challenges and applications of this problem. To address this problem, we proposed a novel method NEAT, which leverages the graph structure to better model the treatment assignments, and mitigates the confounding biases by using the graph structure as an instrumental variable. Considering the fact that the observational data is often time-varying in the real world, we further generalize the problem to a dynamic setting. Extensive experiments on synthetic, semi-synthetic, and real-world graph data validate the effectiveness of the proposed method. Especially, the validation of our method on real-world data provides insights for its future applications in real-world clinical studies. In the future, interesting directions of entangled treatment modeling on graphs include incorporating different levels of graph information (e.g., local-level and global-level) in treatment modeling, and considering entanglements in different types of graph data such as heterogeneous graphs and knowledge graphs.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under grants (IIS-1955797, IIS-2006844, IIS-2144209, IIS-2223769, CNS-2154962, BCS-2228534, CCF-1918656), the Commonwealth Cyber Initiative awards (VV-1Q23-007 and HV-2Q23-003), the JP Morgan Chase Faculty Research Award, the Cisco Faculty Research Award, the Jefferson Lab subcontract 23-D0163, the UVA 3 Cavaliers seed grant, the 4-VA collaborative research grant, the National Institutes of Health (NIH) grant 2R01GM109718-07, and CDC MIND cooperative agreement U01CK000589. One of the authors, Gregory Madden, is an iTHRIV Scholar. The iTHRIV Scholars Program is partly supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Numbers UL1TR003015 and KL2TR003016. This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates (“JP Morgan”), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

REFERENCES

- [1] Joshua D Angrist and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- [2] Peter M Aronow and Cyrus Samii. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11, 4 (2017), 1912–1947.
- [3] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.
- [4] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. 2019. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems* 32 (2019).
- [5] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. 2020. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*.
- [6] Daniel Borrajo, Manuela Veloso, and Sameena Shah. 2020. Simulating and classifying behavior in adversarial environments based on action-state traces: An application to money laundering. In *Proceedings of the First ACM International Conference on AI in Finance*. 1–8.
- [7] 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).
- [8] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. 2011. Non-parametric instrumental regression. *Econometrica* 79, 5 (2011), 1541–1565.
- [9] Manlio De Domenico, Antonio Lima, Paul Mougél, and Mirco Musolesi. 2013. The anatomy of a scientific rumor. *Scientific reports* 3, 1 (2013), 1–9.
- [10] Abdulrahman M El-Sayed, Lars Seemann, Peter Scarborough, and Sandro Galea. 2013. Are network-based interventions a useful antiobesity strategy? An application of simulation models for causal inference in epidemiology. *American Journal of Epidemiology* 178, 2 (2013), 287–295.
- [11] Paul Erdős, Alfred Rényi, et al. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* (1960).
- [12] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *International Conference on Web Search and Data Mining*.
- [13] Kurinchi Selvan Gurusamy, Rahul Koti, Clare D Toon, Peter Wilson, and Brian R Davidson. 2013. Antibiotic therapy for the treatment of methicillin-resistant *Staphylococcus aureus* (MRSA) infections in surgical wounds. *Cochrane Database of Systematic Reviews* 8 (2013).
- [14] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).
- [15] Lars Peter Hansen, John Heaton, and Amir Yaron. 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics* 14, 3 (1996), 262–280.
- [16] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*. PMLR, 1414–1423.
- [17] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011).
- [18] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [19] Michael G Hudgens and M Elizabeth Halloran. 2008. Toward causal inference with interference. *J. Amer. Statist. Assoc.* 103, 482 (2008), 832–842.
- [20] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*.
- [21] Alexander J Kallen, Yi Mu, Sandra Bulens, Arthur Reingold, Susan Petit, KEN Gershman, Susan M Ray, Lee H Harrison, Ruth Lynfield, Ghinwa Dumyati, et al. 2010. Health care-associated invasive MRSA infections, 2005–2008. *Jama* (2010).
- [22] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [23] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences* 116, 10 (2019), 4156–4165.
- [24] Greg Lewis and Vasilis Syrgkanis. 2018. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164* (2018).
- [25] Jing Ma, Yushun Dong, Zheng Huang, Daniel Mietchen, and Jundong Li. 2022. Assessing the Causal Impact of COVID-19 Related Policies on Outbreak Dynamics: A Case Study in the US. In *Proceedings of the ACM Web Conference 2022*. 2678–2686.
- [26] Jing Ma, Ruocheng Guo, Chen Chen, Aidong Zhang, and Jundong Li. 2021. Deconfounding with Networked Observational Data in a Dynamic Environment. In *ACM International Conference on Web Search and Data Mining*.
- [27] Jing Ma, Mengting Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. 2022. Learning Causal Effects on Hypergraphs. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [28] Yunpu Ma and Volker Tresp. 2021. Causal Inference under Networked Interference and Intervention Policy Enhancement. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3700–3708.
- [29] Carlene A Muto, John A Jernigan, Belinda E Ostrowsky, Hervé M Richet, William R Jarvis, John M Boyce, and Barry M Farr. 2003. SHEA guideline for preventing nosocomial transmission of multidrug-resistant strains of *Staphylococcus aureus* and enterococcus. *Infection Control & Hospital Epidemiology* 24, 5 (2003), 362–386.
- [30] Whitney K Newey and James L Powell. 2003. Instrumental variable estimation of nonparametric models. *Econometrica* 71, 5 (2003), 1565–1578.
- [31] Elizabeth L Ogburn, Ilya Shpitser, and Youjin Lee. 2020. Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)* (2020).
- [32] Jan Ohst, Fredrik Liljeros, Mikael Stenhem, and Petter Holme. 2014. The network positions of methicillin resistant *Staphylococcus aureus* affected units in a regional healthcare system. *EPJ Data Science* 3 (2014), 1–15.
- [33] Judea Pearl. 2009. *Causality*.
- [34] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association* 75, 371 (1980), 591–593.
- [35] Donald B Rubin. 1986. Statistics and causal inference: Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* 81, 396 (1986), 961–962.
- [36] Donald B Rubin. 2005. Bayesian inference for causal effects. *Handbook of Statistics* (2005).
- [37] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
- [38] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*. PMLR, 3076–3085.
- [39] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*.
- [40] Erica S Shenoy, Molly L Paras, Farzad Noubary, Rochelle P Walensky, and David C Hooper. 2014. Natural history of colonization with methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus* (VRE): a systematic review. *BMC infectious diseases* 14, 1 (2014), 1–13.
- [41] Andrew F Shorr, Daniela E Myers, David B Huang, Brian H Nathanson, Matthew F Emons, and Marin H Kollef. 2013. A risk score for identifying methicillin-resistant *Staphylococcus aureus* in patients presenting to the hospital with pneumonia. *BMC infectious diseases* 13, 1 (2013), 1–7.
- [42] Stefania Stefani, Doo Ryeon Chung, Jodi A Lindsay, Alex W Friedrich, Angela M Kearns, Henrik Westh, and Fiona M MacKenzie. 2012. Methicillin-resistant *Staphylococcus aureus* (MRSA): global epidemiology and harmonisation of typing methods. *International journal of antimicrobial agents* (2012).
- [43] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. 2012. On causal inference in the presence of interference. *Statistical methods in medical research* 21, 1 (2012), 55–75.
- [44] Panos Toulis, Alexander Volfovsky, and Edoardo M Airoidi. 2018. Propensity score methodology in the presence of network entanglement between treatments. *arXiv preprint arXiv:1801.07310* (2018).
- [45] Panos Toulis, Alexander Volfovsky, and Edoardo M Airoidi. 2021. Estimating causal effects when treatments are entangled by network dynamics. (2021).
- [46] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *ICLR* (2018).
- [47] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018).
- [48] Yinwei Wei, Xiang Wang, Liqiang Nie, Shaoyu Li, Dingxian Wang, and Tat-Seng Chua. 2022. Causal Inference for Knowledge Graph based Recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).
- [49] 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).
- [50] Yuan Yuan, Kristen Altenburger, and Farshad Kooti. 2021. Causal Network Motifs: Identifying Heterogeneous Spillover Effects in A/B Tests. In *Proceedings of the Web Conference 2021*. 3359–3370.
- [51] Elena Zheleva and David Arbour. 2021. Causal inference from network data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4096–4097.
- [52] Corwin M Zigler and Georgia Papadogeorgou. 2021. Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics* 36, 1 (2021), 109.

A ANALYSIS

In this section, we provide a more detailed analysis of the proposed method. Again, the outcome generation function defined in Eq. (5) is:

$$Y^P = \mathcal{Y}(T^P, X^P, M^P) + g(U^P). \quad (23)$$

Inspired by [16], a counterfactual prediction function is defined as:

$$\mathcal{H}(T^P, X^P, M^P) = \mathcal{Y}(T^P, X^P, M^P) + \mathbb{E}[g(U^P)|X^P, M^P]. \quad (24)$$

Here, $\mathcal{H}(T^P, X^P, M^P)$ is what we aim to estimate. As the hidden confounders U^P cannot be observed, it is difficult for classical methods to directly fit this function from observational data. Fortunately, based on the assumptions mentioned in Section 2.2, we have:

$$\begin{aligned} \mathbb{E}[Y^P|X^P, M^P, A^P] &= \mathbb{E}[\mathcal{Y}(T^P, X^P, M^P) + g(U^P)|X^P, M^P, A^P] \\ &= \mathbb{E}[\mathcal{Y}(T^P, X^P, M^P)|X^P, M^P, A^P] \\ &\quad + \mathbb{E}[g(U^P)|X^P, M^P, A^P] \\ &= \mathbb{E}[\mathcal{Y}(T^P, X^P, M^P)|X^P, M^P, A^P] \\ &\quad + \mathbb{E}[g(U^P)|X^P, M^P] \\ &= \int \mathcal{Y}(T^P, X^P, M^P) d\mathcal{F}(T^P|X^P, M^P, A^P) \\ &\quad + \int \mathbb{E}[g(U^P)|X^P, M^P] d\mathcal{F}(T^P|X^P, M^P, A^P) \\ &= \int \mathcal{H}(T^P, X^P, M^P) d\mathcal{F}(T^P|X^P, M^P, A^P), \end{aligned} \quad (25)$$

where $\mathcal{F}(T^P|X^P, M^P, A^P)$ is the conditional distribution of treatment. Here, $\mathcal{H}$ can be estimated with an inverse problem based on observable functions $\mathbb{E}[Y^P|X^P, M^P, A^P]$ and $\mathcal{F}(T^P|X^P, M^P, A^P)$. In our two-stage IV analysis, the first stage can model $\mathcal{F}(T^P|X^P, M^P, A^P)$, and the second stage can model $\mathcal{H}(T^P, X^P, M^P)$.

B DETAILS OF EXPERIMENTS

In this section, we introduce more details of the experimental setup for the reproducibility of the experimental results.

B.1 Baseline Settings

Here are more details for the settings of each baseline:

- **S-Learner:** We use linear regression as the estimator in S-Learner.
- **Causal forest:** We set the number of trees as 100, the minimum number of samples required to be at a leaf node as 10, and the maximum depth of the tree as 10.
- **Counterfactual regression:** The number of epochs is set as 500, the learning rate is 0.001, the batch size is 4000, the

representation dimension is 25. We choose Wasserstein-1 distance [39] for representation balancing.

- **NetDeconf:** We set the number of epochs as 500, the learning rate as 0.005, the representation dimension as 32, and the representation balancing weight as 0.5.
- **DNDC:** We set the number of epochs as 800, the learning rate as 0.001, and the representation dimension as 32.
- **DeepIV:** We use the default parameter setting in the EconML package.

B.2 Experiment Settings

All the experiments in this work are conducted in the following environment:

- Ubuntu 18.04
- Python 3.6
- Scikit-learn 1.0.1
- Scipy 1.6.2
- Pytorch 1.10.1
- Pytorch-geometric 1.7.0
- Networkx 2.5.1
- Numpy 1.19.2
- Cuda 10.1

B.3 Dataset Details

Transaction. This dataset is collected from the anti-money laundering (AML) financial system [3, 6] which provides transaction records between users over time. At each timestamp, we construct a transaction network to represent the transactions occurring inside this timestamp. In the transaction network, each user is represented by a node, and a transaction is an edge between users. We use the user profiles such as location as their covariates.

Social. This dataset contains a real-world social network of people at different timestamps based on tracking from smart devices [9]. Each node represents a user, and each edge represents a friendship between two users.

B.4 Simulation Details

By default, in our potential outcome simulation, we generate the elements in Θ_y from a random Gaussian distribution $\mathcal{N}(0, 0.1^2)$, and Θ_0 from $\mathcal{N}(0, 0.05^2)$. In the dynamic setting, we generate the weights W_*^r with random sampling from $\mathcal{N}(1 - (r/R), (1/R)^2)$. In this way, when r is larger, the weights become smaller. This is consistent with the general observation that earlier information has weaker influence on the future data.