Estimate Causal Effects of Entangled Treatment on Graphs using Disentangled Instrumental Variables

Jingyuan Chou
Department of Computer Science
Biocompleixty Institute
University of Virginia
Charlottesville, USA
jc2wv@virginia.edu

Jiangzhuo Chen
Biocompleixty Institute
University of Virginia
Charlottesville, USA
chenj@virginia.edu

Madhav Marathe
Department of Computer Science
Biocompleixty Institute
University of Virginia
Charlottesville, USA
mvm7hz@virginia.edu

2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825713

***Abstract*—Causal effect estimation on a graph of connected units is often complicated by entangled treatments, where the treatment assignment is not independent for each individual. This presents multiple challenges: accurately modeling treatment assignment mechanisms, adjusting for both observed and unobserved confounders to mitigate confounding bias, and constructing instrumental variables to adjust unobserved confounders within a graph structure. Prior research on estimating the causal effects of entangled treatments either assumed no unobserved confounders or relied on the manual selection of IVs, leading to gaps in the methodology. To bridge these gaps and build upon previous work, we introduce the Graph-Disentanglement Instrumental Variable (GDIV) model, a novel approach employing both Graph Neural Networks (GNNs) and Adversarial Networks to assess the causal effects on nodes in a graph, considering observed/unobserved confounders and the intricacies of treatment entanglement. Our GDIV estimator is validated through extensive experiments across synthetic and semi-synthetic datasets, demonstrating its better performance over state-of-the-art methods. The ablation studies and robustness experiments verify the benefits of leveraging adversarial networks to generate IVs that satisfy the required assumptions.**

I. Introduction

Causal effect estimation helps us understand how treatments impact outcomes, which is essential across many fields. Unlike randomized controlled trials, where subjects are independent, observational studies are more complex. This makes it harder to reliably infer causal effects since assumptions used in randomized settings may not hold. For example, the *strong ignorability assumption*, crucial for adjusting for confounders influencing both treatment and outcome, is difficult to ensure when unobserved factors may be present. Additionally, observational studies typically provide only the observed (factual) outcomes, leaving the counterfactual outcomes—central to causal inference—unavailable. This challenge intensifies in cases where treatments depend on connections between units, as seen in graph-structured data.

Consider a social network of students connected by friendships. We want to determine how participating in a new educational program (the treatment) affects a student's academic performance, measured by grades (the outcome). A student's decision to join the program often depends on whether their friends also enroll, creating treatment interdependencies across the network. However, *the Stable Unit Treatment Value Assumption (SUTVA)* can still hold here if we assume each student's academic outcome depends only on their own enrollment and the characteristics of their friends, not on whether those friends are also enrolled. This setup, where treatments are interconnected yet outcomes depend solely on individual treatments and neighboring covariates, introduces unique challenges for estimating causal effects in graph-based settings.

Our research addresses causal effect estimation in networked environments with entangled treatments, as discussed by Toulis et al. [1], [2] and Ma et al. [3]. Prior studies show that ignoring unit interconnectedness can lead to inaccuracies by attributing treatment effects to individual characteristics instead of network context. Toulis et al. [1], [2] used network features, such as node degrees, for modeling treatment assignments but relied on the *unconfoundedness* assumption, which may overlook certain biases. Building on this, Ma et al. [3] introduced an instrumental variable (IV) approach tailored for networks to account for unobserved confounders, addressing gaps in Toulis's method.

While addressing a similar setting [3], our study diverges by focusing on constructing an IV model directly from the graph structure. We aim to estimate causal effects on graphs by not only addressing biases from unobserved confounders but also explicitly modeling the complex dynamics of entangled treatments. Two key challenges arise: i) **Adjusting for Unobserved Confounders**: Approaches typically use either proxy variables to approximate hidden confounders [4]–[6] or construct Instrumental Variables (IVs) [7]–[12]. IVs allow causal estimation by being associated with the treatment but not directly with the outcome or confounders, mimicking random treatment assignment. We adopt the IV approach. ii) **Modeling Treatment Assignment under Entanglement**: Prior work [1], [2] applies specific functions over the graph to model this mechanism, though these functions are typically unknown in practice.

To address these issues and complement existing IV methods [3], we propose the data-driven **Graph-Disentanglement Instrumental Variable (GDIV)** model. GDIV addresses both challenges: i) **Unobserved Confounders**: Unlike manual IV

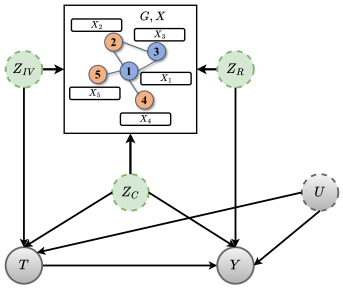

Fig. 1. Causal DAG: dashed circles represent unobserved factors (Z_{IV}, Z_C, Z_R, U) and solid circles for available variables (G, X, T, Y)

selection, which requires domain knowledge, our data-driven approach, inspired by [9], [11], [13], [14], disentangles graph information into three components—IV (affecting treatment only), latent confounder (affecting both treatment and outcome), and risk factor (affecting outcome only), as shown in Fig. 1. This disentangled IV corrects biases from hidden confounders in causal estimation. To ensure IV validity, we incorporate a dual adversarial framework that enforces IV assumptions. ii) **Unknown Treatment Assignment Mechanism**: Rather than assuming a functional form, we leverage the graph structure and introduce a learnable graph neural network model to capture dependencies in treatment assignments.

Our model differs from existing causal estimation methods on graphs in several ways: i) We do not use the graph as a proxy for hidden confounders. ii) We assume no interference between units (i.e., each unit's outcome is unaffected by other units' treatments). iii) While traditional causal assumptions such as *positivity* and *SUTVA* hold in our work, *strong ignorability* does not. Our main contributions are:

- We propose a novel data-driven IV model to adjust unobserved confounders with entangled treatment on graphs. This method constructs IVs automatically from graph information, which avoids manual IV selection that requires domain knowledge and external resources.
- We design a dual adversarial learning framework to guarantee the correctness of the disentangled IV concerning the three assumptions required for IV models.
- Extensive experimental results on synthetic and semi-synthetic datasets demonstrate that our model outperforms state-of-the-art baselines. We show that our model consistently outperforms the baselines under various levels of treatment entanglement and unobserved confounding, signifying the robustness of the proposed model. We also run ablation studies to verify the benefits of leveraging adversarial networks to satisfy the IV assumptions.

II. Related Work

Our work lies in the intersection of the IV model, causal estimation on graphs, and entangled treatment. We describe how our work relates to these three topics, respectively.

- **IV Modeling**: Instrumental Variable (IV) models [12], [15] address unobserved confounders, with two-stage least squares (2SLS) [16] being a common linear method. For greater flexibility, newer models use neural networks, such as DeepIV [10] for non-linear estimation, and DeepGMM [8], [17] leveraging moment conditions for causal effects. With the development of Variational Autoencoders (VAE) [18], VAE-based models emerged: TEDVAE [11], [19] disentangles covariates into IV, confounding, and risk factors, and DVAE.CIV [9] introduces conditional IVs for causal estimation. IV generation methods [20]–[25] learn IVs from features or group IVs for various treatment settings, though some may struggle without specified valid IVs or in graph contexts. Unlike these approaches, our method leverages the graph structure to learn disentangled IVs automatically, reducing bias and using adversarial networks to satisfy IV assumptions.
- **Causal Estimation on Graphs**: Numerous methods for causal estimation on graphs exist [3]–[5], [7], [26]–[37]. Ma et al. [4] and Guo et al. [5] use graphs to adjust for unobserved confounders, while Veitch et al. [35] employ network embeddings with observed proxies. Shi et al. [38] leverage propensity scores in Dragonnet, and Ma et al. [7] propose HyperSci for high-order interference in networks. These approaches inform our method. In addition to causal effect estimation, prior research has addressed spatial confounders and spillover effects [6], [30], [36], [39]–[41]. Cristali et al. [6] formalize methods to address homophily and apply embeddings for peer effect estimation. Fatemi et al. [30] and Cai et al. [36] use independent sets to disentangle peer effects, with Cai et al. guaranteeing estimator performance under their design. For network confounding, [41] formalizes non-local confounding (NLC) within the potential outcomes framework, distinguishing it from causal interference. Papadogeorgou et al. [40] propose spatial causal graphs to address both interference and bias from unmeasured spatial confounding. In our work, we address NLC by designing an IV model for bias mitigation.
- **Entangled Treatment**: Prior work [1], [2] models treatment entanglement as predefined functions on the graph, adapting traditional propensity scores to incorporate graph structure but constrained by the unconfoundedness assumption. Ma et al. [3] address unobserved confounders by using node structures as instrumental variables, enhancing causal effect estimation. However, their approach requires manual IV selection, relying on domain knowledge and limiting practicality. Our work shares this problem setting but advances IV methodology by providing an automated alternative to their IV approach.

In summary, existing models face challenges in estimating causal effects on graphs due to entangled treatments and unobserved confounders, often lacking valid IVs or support for graph structures. Our model addresses this by automatically constructing and validating IVs using the graph structure, reducing bias and eliminating manual IV selection.

III. Problem Formulation

Let $G = (\mathcal{V}, \mathcal{E})$ be an undirected graph with $N = |\mathcal{V}|$ units and a set of edges $\mathcal{E}$. Let $A \in \{0,1\}^{N \times N}$ be the adjacency matrix of G. Each unit $i \in [N]$ is associated with three variables $\{X_i, T_i, Y_i\}$, where $X_i \in \mathbb{R}^{d_x}$ is d_x-dimensional pre-treatment covariates; T_i, a categorical variable, is the treatment of i; and $Y_i \in \mathbb{R}$ is the outcome of i. We describe the causal DAG of our work in Fig. 1. We denote U as an unobserved confounder that impacts both treatment T and outcome Y. We adopt the potential outcome framework [42], and denote the potential outcome $Y(t) = \{Y_i(t)\}_{i \in [N]}$ under treatment $T = t$. If we consider $T = t_0$ as the baseline treatment, we define the Average Treatment Effect (ATE) over N nodes on G as τ:

$$\tau = \mathbb{E}[Y_i(t) - Y_i(t_0) | X, G] \tag{1}$$

To account for the entangled treatments on the graph, similar to previous work [3], we attribute the entangled treatments to the interaction between nodes, thus we define treatment to be a function Φ of G, U, and X as:

$$T = \Phi(G, X, U) \tag{2}$$

Then, we formally define the problem we study in this work: *Given the observational data $\{X, G, T, Y\}$, we aim to estimate the treatment effect τ_i for different units with entangled treatments in the graph in the presence of U.*

Similar to [10], we assume the outcome Y is structurally determined by treatment T, features X, graph G, and unobserved confounder U as:

$$Y = f(G, X, T, U) \tag{3}$$

We assume f is some unknown and potentially non-linear function, and U cannot be directly observed. To mitigate the bias brought by U, we need to identify the source of variation in the treatment that is not confounded by U, which motivates an IV model on the graph. As noted, we do not require the *strong ignorability* assumption, but we do assume *consistency* and *SUTVA*, and consider U independent of X and G. Given these assumptions, our goal is to answer: *How can we build an IV model based on (G, X, Y) to adjust for U and accurately estimate causal effects? Does the constructed IV satisfy IV assumptions?* To address this, we integrate GraphITE [43], an encoder-decoder framework, into a dual adversarial learning framework for IV construction.

IV. Model Framework

In this section, we describe our IV model, which follows a two-stage least squares (2SLS) approach: we estimate treatment $\hat{T}$ in the first stage, then use $\hat{T}$ to predict potential outcomes in the second stage.

Our model adjusts for unobserved confounders by automatically constructing a valid IV, avoiding manual selection and domain knowledge required in prior work [3]. Inspired by [9], [11], we use a data-driven disentanglement module to build an IV that meets the necessary assumptions.

In the first stage, we extend the GraphITE framework [43] to disentangle (X, G) into the IV component Z_{IV}, latent confounder Z_C, and risk factor Z_R, obtaining estimated treatments $\hat{T}$ for all nodes. The adversarial strategy ensures the three IV assumptions are met. In the second stage, an MLP predicts potential outcomes based on (Z_{IV}, Z_C, Z_R) and $\hat{T}$. The workflow is shown in Fig. 2.

A. First Stage: IV Assumptions

To address confounding by U, we construct an IV to reduce bias and improve causal effect estimation. In 2SLS methods, the first stage estimates $\hat{T}$ as the projection of T onto the IV, effectively modeling entangled treatments. We use GraphITE [43] to disentangle and construct $\{Z_{IV}, Z_C, Z_R\}$. To ensure Z_{IV} is a valid IV, the following assumptions must be satisfied:

- *Assumption 1:* ***Relevance****: Given X_i of a random unit i and G, the treatment T_i is relevant to $Z_{i,IV}$, $Z_{i,IV} \not\perp\!\!\!\perp T_i \mid X_i, G$*
- *Assumption 2:* ***Exclusion****: Shown in Fig. 1, the impact of Z_{IV} on Y is fully mediated by T. Furthermore, the disentanglement process is built upon the latent representation of every single node, which is an individual-level operation, thus there does not exist another path from Z_{IV} to Y.*
- *Assumption 3:* ***Instrumental Unconfoundedness****: Shown in Fig. 1, there does not exist a unblocked backdoor path between Z_{IV} and Y.*

To satisfy three assumptions, we build a dual adversarial learning framework.

B. First Stage: Graph Disentanglement

As shown in Fig. 2, we first disentangle (G, X) into $\{Z_{IV}, Z_C, Z_R\}$ and validate IV assumptions using adversarial learning. Specifically, we use GraphITE [43] as an encoder-decoder VAE framework for embedding generation. Following GraphITE, our model's encoder applies forward message passing, using a mean-field approximation to define the variational:

$$q_\phi(Z|A, X) \approx \prod_{i=1}^{n} q_{\phi_i}(Z_i|A, X) \tag{4}$$

Where Z_i is the embedding for node i, ϕ is the variational parameter, and $q(\cdot)$ is the distribution. We assume each variational marginal $q_{\phi_i}(Z_i|A, X)$ is re-parameterizable and easy to sample, ensuring low-variance gradients for ϕ_i [44]. We further assume isotropic Gaussian marginals with diagonal covariance, implying uncorrelated latent dimensions with equal variance in all directions, and use a GNN to specify the parameters of $q_{\phi_i}(Z_i|A, X)$.

$$\mu, \sigma = GNN_\phi(A, X) \tag{5}$$

Where μ and σ denote the vector of means and standard deviations for the variational marginals. Using the reparameterization technique, we obtain (Z_{IV}, Z_C, Z_R) and apply an adversarial learning strategy to validate assumptions.

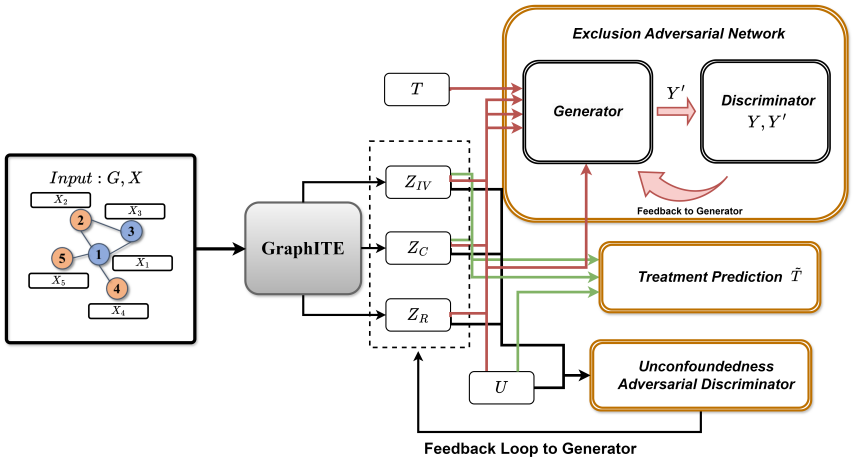

Fig. 2. In the first stage, we disentangle (G, X) into three sets of node embeddings (Z_{IV}, Z_C, Z_R) and use a dual adversarial network to ensure IV assumptions are met. In the second stage (not shown in the figure), an MLP estimates potential outcomes based on $Z_{IV}, Z_R, \hat{T}, U$.

C. Adversarial Learning for Assumptions Validation

To ensure the constructed IV satisfies the three assumptions, we design a dual adversarial learning module to reflect *exclusion* and *unconfoundedness* assumptions, and a treatment prediction module for the *relevance* assumption. The assumption validation workflow is illustrated in Fig. 2, where the three yellow boxes represent the modules. Specifically, the *exclusion* adversarial network includes a generator and discriminator, while the *unconfoundedness* adversarial network uses GraphITE as the generator, differing from the *exclusion* module. We describe each module in detail below.

1) ***Unconfoundedness Adversarial****:* To satisfy this assumption, we enforce mutual independence among (Z_{IV}, Z_C, Z_R, U),

$$q(Z_{IV}, Z_C, Z_R, U) = q(Z_{IV})q(Z_C)q(Z_R)q(U) \quad (6)$$

Where we model the joint distribution $q(Z_{IV}, Z_C, Z_R, U)$ as:

$$q(Z_{IV}, Z_C, Z_R, U) = q(Z_{IV}, Z_C, Z_R|G, X)q(U|T, Y) \quad (7)$$

The aim of Eq. 6 is to promote independence that the marginal distribution should be the same as the joint distribution. Mathematically, we formulate to minimize the *Total Correlation* (TC), which is the KL divergence between $q(Z_{IV}, Z_C, Z_R, U)$ and $q(Z_{IV})q(Z_C)q(Z_R)q(U)$, and we denote them as Q_{ind} and $\bar{Q}_{ind}$ for simplicity. To calculate the KL divergence between them, inspired by [45], we adopt the permutation trick to approximate this KL divergence. Specifically, through Eq. 7, we get the real samples as the view of Q_{ind}, then we randomly permute across batches for each latent factor to obtain the permuted samples, which could approximate $\bar{Q}_{ind}$ if the batch size is sufficiently large.

Concretely, we train a discriminator $D_{ind,\psi}$ to output the probability of the sample coming from $\bar{Q}_{ind}$ instead of Q_{ind}. In the max-stage, we train $D_{ind,\psi}$ to be discriminative, while in the min-stage, we fix the parameters of $D_{ind,\psi}$ and train q_ϕ to generate latent factors with the distributions close to $\bar{Q}_{ind}$. Formally, the min-max learning objective O_{ind} is defined as:

$$\min_{q_\phi} \max_{D_{ind,\psi}} \mathbb{E}_{(Z_{IV}, Z_C, Z_R, U)\sim\bar{Q}_{ind}}[log(D_{ind,\psi}(Z_{IV}, Z_C, Z_R, U))] + \mathbb{E}_{(Z_{IV}, Z_C, Z_R, U)\sim Q_{ind}}[1 - log(D_{ind,\psi}(Z_{IV}, Z_C, Z_R, U))] \quad (8)$$

From Eq. 8, q_ϕ (GraphITE in Fig. 2) serves as the generator to obtain latent embeddings from the marginal distribution. Overall, this *unconfoundedness* adversarial network guarantees we have low TC so that (Z_{IV}, Z_C, Z_R, U) are mutually independent, and such mutual independence would also help validate the exclusion assumption in another parallel adversarial learning framework.

2) ***Exclusion Adversarial****:* In addition to the *unconfoundedness* module, we build an *exclusion* adversarial network to support the assumption that *the effect of* Z_{IV} *on* Y *is fully mediated by* T, meaning Z_{IV} should have no direct effect on Y. Thus, we design the generator and discriminator accordingly: The generator aims to produce data Y' that the discriminator cannot distinguish from real data. It is penalized if the discriminator detects a direct influence of Z_{IV} on Y'. The generator loss is given by:

$$L_{Gen} = -\mathbb{E}_{(T, Z_C, U, Z_R)\sim P, Y'\sim Gen(T, Z_C, U, Z_R)} [logD_{exl}(Z_{IV}, T, Z_C, U, Z_R, Y')] \quad (9)$$

Where Gen denotes the generator of this adversarial learning, D_{exl} denotes the discriminator of the *exclusion* adversarial module, and P denotes the true distribution. This loss encourages the generator to produce outcomes Y' that follow the exclusion assumption, indicating Z_{IV} only influences Y' through T. We formulate the loss for D_{exl} as:

$$L_{D_{exl}} = L_{fake} + L_{real} \quad (10)$$

$$L_{fake} = -\mathbb{E}_{(Z_{IV}, T, Z_C, U, Z_R)\sim P, Y'\sim Gen(T, Z_C, U, Z_R)} [log(1 - D_{exl}(Z_{IV}, Y', T, Z_C, U, Z_R))] \quad (11)$$

$$L_{real} = -\mathbb{E}_{(Z_{IV}, Y, T, Z_C, U, Z_R)\sim P}[logD_{exl}(Z_{IV}, Y, T, Z_C, U, Z_R)] \quad (12)$$

For real data in Eq. 12, the discriminator should output a high probability for Y conditioned on (T, Z_C, Z_R, U) with no direct influence from Z_{IV}, supporting the *exclusion* assumption. For generated data Y' in Eq. 11, Y' may violate this assumption, and the discriminator penalizes the generator if Z_{IV} directly affects Y'. Overall, the *exclusion* adversarial learning enforces the *exclusion* assumption by penalizing the generator whenever the discriminator detects Z_{IV}'s direct influence on Y'. Meanwhile, the *unconfoundedness* adversarial learning promotes the independence among (Z_{IV}, Z_R, Z_C, U), which explicitly suggests that there does not exist another causal path from Z_{IV} to Y, further strengthen the *exclusion* assumption.

D. First Stage: Entangled Treatment Prediction

We satisfy the *relevance* assumption using a treatment prediction module, reflecting the causal relationship $Z_{IV} \rightarrow T$, $Z_C \rightarrow T$, and $U \rightarrow T$. We frame a GNN model $\mathcal{F}_T$ to account for such causal relationships. Mathematically, given obtained Z_{IV}, Z_C, and G, we use a one-layer GCN to be $\mathcal{F}_T$:

$$\hat{T} = \eta(\tilde{A}(Z_{IV} \oplus Z_C)W_T \oplus UW_u) \tag{13}$$

Here, η is the sigmoid/softmax function, $\tilde{A}$ is the normalized adjacency matrix of G, $\oplus$ denotes concatenation, and W_T and W_U are learnable parameters.

1) Loss Objective in First Stage: Generally, the objective of the first stage of the IV model is to model the causal effect of IV on treatment. To validate the constructed IV, we design a dual adversarial learning framework to satisfy the IV assumptions. The overall loss for the first stage consists of different types of losses. Specifically, we construct it as:

$$\mathcal{L}_{First} = \mathcal{L}_T + \mathcal{L}_{exl} + \mathcal{L}_{ind} + \alpha\|\Theta\|^2 \tag{14}$$

Where $\mathcal{L}_T = NLL(T, \hat{T})$, and NLL denotes the negative log-likelihood loss, and the last term is the L2 regularization loss, α is the hyperparameters for regularization, and Θ denotes all the parameters involved in the first stage. $\mathcal{L}_{ind}$ corresponds to Eq. 8, and

$$\mathcal{L}_{exl} = \min_{Gen} \max_{D_{exl}} L_{Gen} L_{D_{exl}} \tag{15}$$

Where L_{Gen} and $L_{D_{exl}}$ correspond to Eq. 9 and Eq. 10.

E. Second Stage: Potential Outcome Prediction

With predicted $\hat{T}$ from the first stage, we design an outcome prediction module $H(\cdot)$ in the second stage. Specifically, for each node, we predict $\hat{Y}_i$ based on $(Z_{i,R}, Z_{i,C}, \hat{T}_i, U_i)$:

$$\hat{Y}_i = H(Z_{i,R}, Z_{i,C}, \hat{T}_i, U_i) \tag{16}$$

1) Loss Objective in Second Stage: We denote the loss function of the second stage as:

$$\mathcal{L}_{Second} = \sum_{i}^{N} MSE(Y_i, \hat{Y}_i) \tag{17}$$

Where MSE denotes the Mean Squared Error (MSE), and we naturally formulate the individual treatment effect of node i:

$$\hat{\tau}_i = \hat{Y}_i(t) - \hat{Y}_i(t_0) \tag{18}$$

V. Experiment

In this section, we experimentally validate our model by addressing the following research questions (RQs): **RQ1**: How does our model compare with state-of-the-art methods? **RQ2**: How does our model perform with varying levels of treatment entanglement? **RQ3**: How does our model perform with different levels of unobserved confounding? **RQ4**: What are the contributions of the exclusion and independence adversarial modules to model performance?

A. Evaluation Metrics

We adopt two well-received metrics for causal effect estimation, Precision in Estimation of Heterogeneous Effect (ϵ_{PEHE}, $\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{N}\sum_{i\in N}(\tau_i - \hat{\tau}_i)^2}$) [46], and Mean Absolute Error of ATE (ϵ_{ATE}, $\epsilon_{ATE} = \left\|\frac{1}{N}\sum_{i\in N}\tau_i - \frac{1}{N}\sum_{i\in N}\hat{\tau}_i\right\|$). We run each experiment 10 times and report the corresponding average value and standard deviation.

B. Simulation

In this section, we create synthetic datasets by simulating all variables and semi-synthetic datasets using real covariates and graphs (e.g., BlogCatalog and Flickr [5]). For the semi-synthetic datasets, treatments and outcomes are simulated. Table I shows dataset characteristics, with the simulation following the causal DAG in Fig. 1.

TABLE I
STATISTICS OF DATASETS

Datasets	# nodes	# edges	# features
Synthetic	1000	99629	100
BlogCatalog	5196	171743	8189
Flickr	7575	239738	12047

We formulate the detailed simulation strategy as follows:

- Hidden Confounders: we simulate U_i:

$$U_i \sim \mathcal{N}(0, \gamma I) \tag{19}$$

  Where I denotes the identity matrix with size d_u, the dimension of hidden confounders, and γ is the scaling factor, which is set to 20 per [3].
- Features: To align causal DAG in Fig. 1, and also to make features comprehensive, we simulate both continuous and categorical features as:

$$X_{i,cont} \sim \mathcal{N}(0, \gamma\mathbf{I}) \quad , \quad X_{i,cate} \sim Categorical(p_{cat}) \tag{20}$$

  We assign a uniform probability distribution to simulated categorical features with four categories, $p_{cat} = \{0.25, 0.25, 0.25, 0.25\}$. The dataset includes 80 continuous features and 20 categorical features. For BlogCatalog and Flickr, we use the available covariates.
- Graph: Graph is a random graph generated using the Erdös-Rényi model [47]. We use NetworkX [48] to generate graphs for synthetic datasets, and the generated graph is employed for simulations of the treatment variable.
- Treatment: To account for the non-local confounding scenarios that T_i is causally influenced by X_i, U_i and

X_j, where j is the neighbors of node i, we simulate the treatment as:

$$T_i = BI((1-\lambda)\theta_{t,x}^T X_i + \lambda \frac{1}{\mathcal{N}_i} \sum_{j \in \mathcal{N}_i} (\theta_{t,x}^T X_j) + \theta_{t,u}^T U_i + \epsilon_t) \tag{21}$$

Here, $\theta_{t,x}$ and $\theta_{t,u}$ are parameters with sizes d_x and d_u, sampled from a Gaussian distribution $\mathcal{N}(0, 0.5^2)$. $\mathcal{N}_i$ is the set of immediate neighbors of node i. The parameter $\lambda \in [0, 1]$ controls the level of treatment entanglement: $\lambda = 0$ means T_i depends only on X_i, while $\lambda = 1$ means T_i depends entirely on the characteristics of $\mathcal{N}_i$. $BI(\cdot)$ is the sigmoid function, converting input to a probability and sampling output via the Bernoulli distribution. $\epsilon_t \in \mathcal{N}(0, 0.01^2)$ represents random Gaussian noise.

- Potential Outcome: After we simulate the treatment, we formulate outcome Y_i as:

$$Y_i(t) = T_i \cdot \theta_y^T X_i + \frac{1}{\mathcal{N}_i} \sum_{j \in \mathcal{N}_i} (\theta_0^T X_j) + \beta U_i \theta_u^T + \epsilon_y \tag{22}$$

Where θ_y and θ_0 are parameters of size d_x, and θ_u is of size d_u, and $\beta \geq 0$ controls the level of unobserved confounding, and ϵ_y represents random Gaussian noise.

C. Baselines

We categorize baselines into five types for comparison:

- Independent Units: Assuming unit independence. We use Causal Forest (CF), Counterfactual Regression (CFR), and Logistic Regression (LR).
- Distribution Modeling: Uses distribution modeling approaches. CEVAE, TEDVAE, and GANITE employ variational autoencoders or adversarial training to model joint outcomes or generate counterfactuals.
- IV Generation: Data-driven IV generation methods including VIV [22], AutoIV [20], and GIV [21], which use adversarial or VAE-based approaches to create valid or group IV candidates.
- Deep Learning-Based: Includes TarNet, for balanced representation learning, and DeepIV, an instrumental variable-based approach. For each node i in DeepIV, we use the i-th row of Z_{IV} from our method.
- Network-Based: Graph-structure-based methods DNDC and Netdeconf, which use network connections as proxies for hidden confounders.

D. Experimental Setting

We use one-hot encoding for categorical features during preprocessing and split the data into 60% training, 20% validation, and 20% testing sets. Each experiment runs for 10 iterations. Learning rates are searched in $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, with L2 regularization applied using a weight decay of 10^{-5}. The hidden dimension is set to 32, while d_u and d_x are set to 100 in simulations; λ and β are set to 0.5. During dual adversarial training, we alternate updates between the discriminators and the generator. First, the exclusion discriminator learns to detect if Z_{IV} directly influences Y, while the independence discriminator ensures (Z_{IV}, Z_C, Z_R, U) are independent. Then, the generator is updated to fool both discriminators, and this cycle repeats until convergence. For baseline implementations, we followed their default configurations. For CF and DeepIV methods, we used built-in methods from *EconML*. Specifically, for DeepIV, treatment and outcome models were created using two dense layers from *Keras*, with 10 components in the mixture density network. For CF, we set the number of trees to 100, keeping other parameters at default values. Note that training dual adversarial networks requires careful tuning for stable results. To make this process smoother, we pre-train the networks on smaller data subsets, providing a stable starting point and reducing the need for intensive tuning, and the learning rates of $1e^{-4}$ for the max stage and $1e^{-3}$ for the min stage work best for our adversarial setup through tuning.

E. Performance Comparison

We present the performance of all methods in Table II and observe that our proposed method consistently outperforms baselines on both synthetic and semi-synthetic datasets. This advantage is due to several factors: i). Data-driven IV methods (GIV, VIV, AutoIV) are the most competitive, but AutoIV struggles without a valid IV, and GIV is limited by its focus on group IVs. VIV achieves close performance to ours by using an adversarial network for the unconfoundedness assumption, though its lack of an exclusion mechanism impacts results slightly. ii). Methods assuming no unobserved confounders (CF, LR, CFR, TarNet) perform less effectively due to the confounding present in our setting. While TEDVAE (which inspired our work) disentangles covariates to allow IV creation, it lacks consideration of node connections, a limitation also present in CEVAE and GANITE. iii). DeepIV, which uses Z_{IV} from our model, performs competitively but is limited by its lack of support for graph data.

F. Robustness of GDIV

We address **RQ2** and **RQ3** using Table III. We choose baselines from the data-driven IV generation category in Table II and vary λ and β to test our model's robustness against different levels of treatment entanglement and unobserved confounding.

*1) **Varied Levels of Unobserved Confounding**:* To assess robustness to confounding, we set $\lambda = 0.5$ to fix treatment entanglement and vary β as $\{0, 0.5, 1\}$, generating synthetic datasets with varying confounding levels. Results in the upper part of Table III show that although error metrics $(\sqrt{\epsilon_{PEHE}}, \epsilon_{ATE})$ increase with higher confounding, our model consistently outperforms baselines.

*2) **Varied Levels of Treatment Entanglement**:* Similarly, we fix $\beta = 0.5$ and vary λ as $[0, 0.5, 1]$ to test robustness to treatment entanglement. Results in the lower part of Table III indicate that **GDIV** outperforms the baselines across all levels of treatment entanglement.

G. Ablation Study

To address **RQ4**, we perform ablation studies to evaluate the effectiveness of each module. We created two model vari-

TABLE II
PERFORMANCE TABLE FOR DIFFERENT FAMILY OF BASELINES

	Synthetic		BlogCatalog		Flickr	
Metric	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
LR	69.84 ± 9.84	11.08 ± 7.25	14.33 ± 4.80	7.47 ± 2.65	4.38 ± 2.72	1.74 ± 1.59
CF	35.46 ± 8.81	7.11 ± 5.92	18.16 ± 4.65	6.61 ± 2.18	4.11 ± 1.19	2.44 ± 1.98
CFR	48.45 ± 7.22	9.28 ± 3.15	72.2 ± 2.58	3.45 ± 2.91	48.81 ± 2.56	2.82 ± 1.91
DeepIV	29.16 ± 5.31	1.92 ± 0.90	12.19 ± 9.18	0.94 ± 0.49	5.10 ± 1.06	0.49 ± 0.40
TarNet	48.20 ± 8.49	10.90 ± 8.77	72.18 ± 3.51	5.06 ± 2.56	48.79 ± 2.07	1.70 ± 1.57
GANITE	29.73 ± 2.59	2.68 ± 1.18	10.62 ± 7.34	0.48 ± 0.35	6.44 ± 1.22	0.40 ± 0.24
CEVAE	30.20 ± 3.43	3.95 ± 2.01	16.15 ± 8.15	1.95 ± 0.97	7.15 ± 1.09	0.97 ± 0.74
TEDVAE	28.19 ± 3.11	2.34 ± 1.91	10.55 ± 3.54	0.49 ± 0.29	7.65 ± 1.25	0.56 ± 0.41
Net-Deconf	35.18 ± 3.35	2.67 ± 2.41	11.47 ± 3.52	1.52 ± 1.28	5.56 ± 1.43	0.86 ± 0.36
DNDC	60.68 ± 4.80	8.49 ± 7.66	71.46 ± 7.25	9.76 ± 2.58	49.79 ± 10.45	3.02 ± 1.22
VIV	24.11 ± 2.19	1.66 ± 0.82	$\mathbf{7.98 \pm 2.33}$	0.31 ± 0.09	2.68 ± 1.10	0.33 ± 0.12
AutoIV	31.22 ± 3.26	1.97 ± 0.87	11.22 ± 3.18	0.41 ± 0.15	3.92 ± 1.48	0.41 ± 0.17
GIV	27.87 ± 2.87	1.99 ± 0.94	9.18 ± 2.81	0.48 ± 0.16	3.59 ± 1.39	0.39 ± 0.17
GDIV∼(Ours)	$\mathbf{21.47 \pm 1.13}$	$\mathbf{1.32 \pm 0.85}$	8.04 ± 2.12	$\mathbf{0.28 \pm 0.16}$	$\mathbf{2.32 \pm 0.79}$	$\mathbf{0.18 \pm 0.10}$

TABLE III
ROBUSTNESS UNDER VARYING LEVELS OF UNOBSERVED CONFOUNDING AND TREATMENT ENTANGLEMENT

$\lambda = 0.5$	$\beta = 0$		$\beta = 0.5$		$\beta = 1.0$	
Metric	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
DeepIV	31.04 ± 3.99	2.16 ± 1.79	29.16 ± 5.31	1.92 ± 0.90	35.16 ± 4.11	4.19 ± 2.80
VIV	22.68 ± 1.98	1.37 ± 0.89	24.11 ± 2.19	1.66 ± 0.82	28.12 ± 3.35	2.65 ± 1.77
AutoIV	25.19 ± 2.11	1.40 ± 1.01	31.22 ± 3.26	1.97 ± 0.87	31.77 ± 3.69	3.01 ± 1.91
GIV	24.33 ± 1.63	1.39 ± 0.98	27.87 ± 2.87	1.99 ± 0.94	27.88 ± 3.19	2.64 ± 1.69
GDIV∼(Ours)	$\mathbf{20.65 \pm 1.03}$	$\mathbf{1.23 \pm 0.71}$	$\mathbf{21.47 \pm 1.13}$	$\mathbf{1.32 \pm 0.85}$	$\mathbf{23.22 \pm 1.44}$	$\mathbf{1.91 \pm 1.23}$
$\beta = 0.5$	$\lambda = 0$		$\lambda = 0.5$		$\lambda = 1$	
Metric	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
DeepIV	25.18 ± 3.97	2.09 ± 1.01	29.16 ± 5.31	1.92 ± 0.90	33.91 ± 4.70	2.13 ± 1.19
VIV	21.13 ± 1.16	1.33 ± 0.41	24.11 ± 2.19	1.66 ± 0.82	27.17 ± 3.11	2.01 ± 1.13
AutoIV	26.29 ± 3.91	1.89 ± 0.92	31.22 ± 3.26	1.97 ± 0.87	33.49 ± 3.98	2.33 ± 1.48
GIV	24.55 ± 3.31	1.81 ± 1.13	27.87 ± 2.87	1.99 ± 0.94	30.15 ± 2.88	2.21 ± 1.39
GDIV∼(Ours)	$\mathbf{20.51 \pm 0.98}$	$\mathbf{1.13 \pm 0.51}$	$\mathbf{21.47 \pm 1.13}$	$\mathbf{1.32 \pm 0.85}$	$\mathbf{23.86 \pm 1.93}$	$\mathbf{1.75 \pm 1.07}$

TABLE IV
PERFORMANCE UNDER ABLATION VARIANTS

Datasets	Synthetic		BlogCatalog		Flickr	
Variants	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
GDIV	$\mathbf{21.47 \pm 1.13}$	$\mathbf{1.32 \pm 0.85}$	$\mathbf{8.04 \pm 2.12}$	$\mathbf{0.28 \pm 0.16}$	$\mathbf{2.32 \pm 0.79}$	$\mathbf{0.18 \pm 0.10}$
GDIV-wo-exl	29.16 ± 3.89	2.00 ± 1.13	11.65 ± 2.88	0.46 ± 0.33	4.12 ± 1.17	0.31 ± 0.11
GDIV-wo-ind	33.51 ± 4.67	2.34 ± 1.46	12.41 ± 3.01	0.59 ± 0.29	3.98 ± 1.02	0.40 ± 0.20

ants: *GDIV-wo-exl*, which removes the *exclusion* adversarial network (allowing potential violation of the *exclusion* assumption), and *GDIV-wo-ind*, which removes the *independence* adversarial network (potentially violating the *instrumental unconfoundedness* assumption). Their performances, summarized in Table IV, were tested under the same settings as Table II. Results show that removing either adversarial module leads to significant drops in both metrics compared to the full model, confirming that violations of IV assumptions reduce causal effect accuracy. This analysis highlights the dual adversarial learning module's role in upholding IV assumptions and shows the model's practical, data-driven robustness.

VI. CONCLUSION

We address the problem of causal effect estimation on graph-structured data, focusing on the challenges of interconnected treatments and unobserved confounders. Our approach introduces a method for creating disentangled instrumental variables (IVs) from graph structures, supported by a dual adversarial network to reinforce IV model assumptions. Experiments on synthetic and semi-synthetic datasets show that our model consistently outperforms existing benchmarks across various levels of treatment entanglement and confounding, highlighting its robustness. Ablation studies confirm the advantage of using adversarial networks to meet IV assumptions.

REFERENCES

[1] P. Toulis, A. Volfovsky, and E. M. Airoldi, "Estimating causal effects when treatments are entangled by network dynamics," Tech. Rep., 2021.

[2] T. Panos, V. Alexander, and M. A. Edoardo, "Propensity score methodology in the presence of network entanglement between treatments," 2018.

[3] J. Ma, C. Chen, A. Vullikanti, R. Mishra, G. Madden, D. Borrajo, and J. Li, "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*, ser. KDD '23, 2023, p. 4584–4594.

[4] J. Ma, R. Guo, C. Chen, A. Zhang, and J. Li, "Deconfounding with networked observational data in a dynamic environment," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ser. WSDM '21, 2021, p. 166–174.

[5] R. Guo, J. Li, and H. Liu, "Learning individual causal effects from networked observational data," in *Proceedings of the 13th International Conference on Web Search and Data Mining*, ser. WSDM '20, 2020, p. 232–240.

[6] I. Cristali and V. Veitch, "Using embeddings for causal estimation of peer influence in social networks," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 15 616–15 628.

[7] J. Ma, M. Wan, L. Yang, J. Li, B. Hecht, and J. Teevan, "Learning causal effects on hypergraphs," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22, 2022, p. 1202–1212.

[8] A. Bennett, N. Kallus, and T. Schnabel, "Deep generalized method of moments for instrumental variable analysis," in *Advances in Neural Information Processing Systems*, 2019.

[9] D. Cheng, Z. Xu, J. Li, L. Liu, T. D. Le, and J. Liu, "Learning conditional instrumental variable representation for causal effect estimation," in *Machine Learning and Knowledge Discovery in Databases*, 2023.

[10] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 06–11 Aug 2017, pp. 1414–1423.

[11] W. Zhang, L. Liu, and J. Li, "Treatment effect estimation with disentangled latent factors," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10 923–10 930, May 2021.

[12] G. Imbens, "Instrumental variables: An econometrician's perspective," National Bureau of Economic Research, Tech. Rep. 19983, March 2014.

[13] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 6449–6459.

[14] D. Cheng, Z. Xu, J. Li, L. Liu, J. Liu, and T. le, "Causal inference with conditional instruments using deep generative models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 7122–7130, 06 2023.

[15] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, vol. 91, no. 434, pp. 444–455, 1996.

[16] J. D. Angrist and G. W. Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, vol. 90, no. 430, pp. 431–442, 1995.

[17] J. H. Lars Peter Hansen and A. Yaron, "Finite-sample properties of some alternative gmm estimators," *Journal of Business & Economic Statistics*, vol. 14, no. 3, pp. 262–280, 1996.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in *International Conference on Learning Representations*, 2020.

[20] J. Yuan, A. Wu, K. Kuang, B. Li, R. Wu, F. Wu, and L. Lin, "Auto iv: Counterfactual prediction via automatic instrumental variable decomposition," *ACM Transactions on Knowledge Discovery from Data*, vol. 16, no. 4, p. 1–20, Jan. 2022. [Online]. Available: http://dx.doi.org/10.1145/3494568

[21] A. Wu, K. Kuang, R. Xiong, M. Zhu, Y. Liu, B. Li, F. Liu, Z. Wang, and F. Wu, "Learning instrumental variable from data fusion for treatment effect estimation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 9, pp. 10 324–10 332, Jun. 2023. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/26229

[22] X. Li and L. Yao, "Distribution-conditioned adversarial variational autoencoder for valid instrumental variable generation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, pp. 13 664–13 672, Mar. 2024.

[23] V. Syrgkanis, V. Lei, M. Oprescu, M. Hei, K. Battocchi, and G. Lewis, "Machine learning estimation of heterogeneous treatment effects with instruments," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] D. Cheng, Z. Xu, J. Li, L. Liu, J. Liu, and T. D. Le, "Conditional instrumental variable regression with representation learning for causal inference," *arXiv preprint arXiv:2310.01865*, 2023.

[25] K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj, "Dual instrumental variable regression," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20, 2020.

[26] S. Harada and H. Kashima, "Graphite: Estimating individual effects of graph-structured treatments," ser. CIKM '21, 2021, p. 659–668.

[27] Q. Huang, J. Ma, J. Li, R. Guo, H. Sun, and Y. Chang, "Modeling interference for individual treatment effect estimation from networked observational data," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 3, dec 2023.

[28] Z. Fatemi and E. Zheleva, "Network experiment design for estimating direct treatment effects," in *KDD International Workshop on Mining and Learning with Graphs (MLG 2020)*, 08 2020.

[29] M. Laan, "Causal inference for a population of causally connected units," *Journal of Causal Inference*, vol. 2, 03 2014.

[30] Z. Fatemi and E. Zheleva, "Minimizing interference and selection bias in network experiment design," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 176–186, 05 2020.

[31] D. Eckles, R. Kizilcec, and E. Bakshy, "Estimating peer effects in networks with peer encouragement designs," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 7316–7322, 07 2016.

[32] P. Toulis and E. Kao, "Estimation of causal peer influence effects," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13, 2013, p. III–1489–III–1497.

[33] C. R. Shalizi and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies," *Sociological Methods & Research*, vol. 40, no. 2, pp. 211–239, 05 2011.

[34] F. Feng, W. Huang, X. He, X. Xin, Q. Wang, and T.-S. Chua, "Should graph convolution trust neighbors? a simple causal inference method," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '21, 2021, p. 1208–1218.

[35] V. Veitch, Y. Wang, and D. M. Blei, "Using embeddings to correct for unobserved confounding in networks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[36] C. Cai, X. Zhang, and E. M. Airoldi, "Independent-set design of experiments for estimating treatment and spillover effects under network interference," *arXiv preprint arXiv:2312.04026*, 2023.

[37] J. Kaddour, Y. Zhu, Q. Liu, M. J. Kusner, and R. Silva, "Causal effect inference for structured treatments," in *Neural Information Processing Systems*, 2021.

[38] C. Shi, D. Blei, and V. Veitch, "Adapting neural networks for the estimation of treatment effects," *Advances in neural information processing systems*, vol. 32, 2019.

[39] E. L. Ogburn and T. J. VanderWeele, "Causal Diagrams for Interference," *Statistical Science*, vol. 29, no. 4, pp. 559 – 578, 2014.

[40] G. Papadogeorgou and S. Samanta, "Spatial causal inference in the presence of unmeasured confounding and interference," 2024.

[41] M. Tec, J. Scott, and C. Zigler, "Weather2vec: Representation learning for causal inference with non-local confounding in air pollution and climate studies," 2022.

[42] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[43] A. Grover, A. Zweig, and S. Ermon, "Graphite: Iterative generative modeling of graphs," *arXiv preprint arXiv:1803.10459*, 2018.

[44] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *NIPS Workshop on Bayesian Deep Learning*, 2016.

[45] H. Kim and A. Mnih, "Disentangling by factorising," 2019. [Online]. Available: https://arxiv.org/abs/1802.05983

[46] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[47] P. Erdős, A. Rényi *et al.*, "On the evolution of random graphs," *Publ. math. inst. hung. acad. sci*, vol. 5, no. 1, pp. 17–60, 1960.

[48] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," Tech. Rep., 2008.