Causal Inference with Latent Variables: Recent Advances and Future Prospectives

Yaochen Zhu
University of Virginia
Charlottesville, VA, USA
uqp4qh@virginia.edu

Yinhan He
University of Virginia
Charlottesville, VA, USA
nee7ne@virginia.edu

Jing Ma
Case Western Reserve University
Cleveland, OH, USA
jing.ma@case.edu

Mengxuan Hu
University of Virginia
Charlottesville, VA, USA
qtq7su@virginia.edu

Sheng Li
University of Virginia
Charlottesville, VA, USA
shengli@virginia.edu

Jundong Li
University of Virginia
Charlottesville, VA, USA
jundong@virginia.edu

Abstract

Causality lays the foundation for the trajectory of our world. Causal inference (CI), which aims to infer intrinsic causal relations among variables of interest, has emerged as a crucial research topic. Nevertheless, the lack of observation of important variables (e.g., confounders, mediators, exogenous variables, etc.) severely compromises the reliability of CI methods. The issue may arise from the inherent difficulty in measuring the variables. Additionally, in observational studies where variables are passively recorded, certain covariates might be inadvertently omitted by the experimenter. Depending on the type of unobserved variables and the specific CI task, various consequences can be incurred if these latent variables are carelessly handled, such as biased estimation of causal effects, incomplete understanding of causal mechanisms, lack of individual-level causal consideration, etc. In this survey, we provide a comprehensive review of recent developments in CI with latent variables. We start by discussing traditional CI techniques when variables of interest are assumed to be fully observed. Afterward, under the taxonomy of circumvention and inference-based methods, we provide an in-depth discussion of various CI strategies to handle latent variables, covering the tasks of causal effect estimation, mediation analysis, counterfactual reasoning, and causal discovery. Furthermore, we generalize the discussion to graph data where interference among units may exist. Finally, we offer fresh aspects for further advancement of CI with latent variables, especially new opportunities in the era of large language models (LLMs).

CCS Concepts

• **Mathematics of computing → Causal networks**.

Keywords

Causal inference; latent variable models; confounding analysis

ACM Reference Format:
Yaochen Zhu, Yinhan He, Jing Ma, Mengxuan Hu, Sheng Li, and Jundong Li. 2024. Causal Inference with Latent Variables: Recent Advances and Future Prospectives. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3671450

KDD '24, August 25–29, 2024, Barcelona, Spain
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671450

1 Introduction

Our world is a woven web of causes and effects, where everything that occurs is the consequence of some prior actions [6, 159]. For example, my headache disappeared because of the aspirin I took this afternoon, and I gained muscle because I worked out regularly every day. From the levity of language, we may be under the illusion that reasoning with causality from experiences can be simple and straightforward. However, formal causal inference did not emerge until decades ago, which enabled rigorously derivation of causal relationships of interest from the observational data [109].

In hindsight, what prevented the emergence of formal causal inference (CI) is the lack of mathematical language to describe causality [100]. One tempting choice is to use conditional distributions from probability theory for causal reasoning [15]. For example, if an event T causes another event Y (where $T, Y = 1$ means that the event *happened* and 0 otherwise), we usually have $p(Y = 1|T = 1) > p(Y = 1|T = 0)$. However, if we use the converse, i.e., the increase of probability, to denote causality, *correlation* can be easily mistaken for *causation*. For example, we can observe that people eat more ice cream when they wear fewer clothes. However, the former is clearly not a cause for the latter, as both are caused by a third variable: hot weather. Here, the issue lies in the fact that $T = 1$ in the conditional distribution means that T is passively observed, but what makes the relation between T and Y causal is that Y will happen if we *make* T happen. That is why Rubin claimed that "there is no causality without intervention" and introduced the potential outcome $Y(T = 1)$ to describe the event Y if $T = 1$ is *made* to happen for all population [55]. Similarly, Pearl introduced the *do*-operator, where $p(Y|do(T))$ denotes the distribution of Y if we make the event T happen instead of observing it passively [96].

With new symbols defined to facilitate causal reasoning, various causal questions can be formed in a rigorous manner. One common CI task is average treatment effect (ATE) estimation [4], which aims to estimate the expected influence of an event (T) on another (Y), e.g., the change of recovery rate Y if drug T is prescribed to all patients. Since ATE compares the outcomes of two interventions, i.e., treatment/no treatment, it can be directly formulated as

$\mathbb{E}[Y(T=1)] - \mathbb{E}[Y(T=0)]$ or $\mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)]$. In addition, causal mediation analysis [53] is also feasible via the new symbols, which aims to determine the fine-grained causal effect of T on Y mediated by other factors. For example, if we know that drug T cures the disease by reducing the blood pressure Z but it also thickens the blood vessel wall, we can define the causal effects of T on Y mediated by Z as the effect as if the drug has no side effect. Furthermore, the individual level causal effect also becomes tractable [67]. For example, for Alice, who has received the treatment and survived, we can formulate the question "would she also survive if no treatment had been provided?". Finally, we can even formulate causal discovery with the new symbols [119], where causal relations among variables of interest (e.g., treatment, mediators) can be automatically discovered from data.

Nevertheless, representing causal questions with new symbols is not enough. After all, directly obtaining the causal estimand $Y(T)$ or $p(Y|do(T))$ requires intervention upon T, which is not always feasible. One strategy is to simulate interventions with randomized experiment (RE) [39], where the randomization ensures that treatment T is the only contributor to the variation of Y. However, RE can be expensive and unethical (e.g., we cannot randomly decide whether or not to give drugs to patients). Therefore, CI with observational studies gains more attention, where the experimenter has no manipulation over the treatment assignment. The aim is to show that if certain assumptions hold for the data (i.e., identification criteria), causal estimand with causal symbols can still be calculated with conditional relations measurable in the collected data. For example, if all confounders (i.e., factors that simultaneously affect cause T and effect Y) are observed and recorded, backdoor adjustment [127] and propensity score weighting [108] can be used to estimate ATE. In addition, if the mediator of interest M is measured, the path-specific effect of T on Y mediated by M can be obtained under certain identification criteria [54]. If exogenous variables (e.g., individual factors not considered as the main variables of interest) are known, individual-level counterfactuals can be calculated [96]. Finally, if all variables of interest are known, mature algorithms such as the PC algorithm [118] are off-the-shelf for causal discovery.

However, important variables for CI can be latent, which hinders the reliability of existing CI techniques [120]. The issue lies primarily in two folds: ***(i)*** First, certain variables can be intrinsically difficult to measure, e.g., the socioeconomic status of a patient, which is a crucial confounder for drug effect evaluation [72]. ***(ii)*** In addition, in the observational study, important covariates for CI may not be recorded in the collected data [91]. The consequences of carelessly handling latent variables for CI can be multi-faceted. First, unobserved confounders can lead to bias in ATE estimation [21]; for example, if the severity of disease is not considered, we may erroneously conclude that an effective drug lowers the recovery rate, as more severe patients tend to be treated with the drug. In addition, missing important mediators could result in an incomplete understanding of the causal mechanism [87]. For example, the debate over the causal relation between tobacco smoking and lung cancer was not resolved until the mediator Tar deposit was determined to cause lung cancer for smokers [111]. Exogenous variables are usually considered as noise and are not explicitly included in the observational data [67]. However, without them, individual differences in treatment effects cannot be estimated, which hinders personalized counterfactual analysis. Finally, if not all variables of interest are available, causal discovery would be impossible [16].

Recent years have witnessed a plethora of works on causal inference with latent variables [72]. Generally, the methods can be categorized into two classes: ***(i)*** *Circumvention-based Methods* and ***(ii)*** *Inference-based Methods.* Circumvention-based strategies eschew direct modeling of latent variables; instead, they show that under certain stringent assumptions/conditions, latent variables can be avoided while the causal estimand can still be identified with observational data. However, there is no free lunch, and the price being paid could be the requirement to measure more variables (where errors could be introduced) [36] and an increase in estimation variance [9]. Inference-based methods, in contrast, explicitly model the latent variables based on the observations. This usually includes proxy of the latent variables (e.g., their noisy observations). However, latent variables may not be identifiable given the observed data, where bias can still remain in the causal estimations [61]. In addition, the proxy of latent variables may contain undesirable components, and carelessly ignoring them can ruin the estimation results [84]. Both strategies on the main CI tasks, as well as their generalization to graph data where interference exists, will be thoroughly discussed. Our contribution can be summarized as:

- ***Timely Topic.*** CI with latent variable is an important topic while scattered in different CI areas. This survey provides a timely and comprehensive review of the state-of-the-art.
- ***Novel Taxonomy.*** We provide novel taxonomy on existing CI methods to address latent variables, where two main categories of methods on four CI tasks are thoroughly discussed.
- ***New Hope.*** Based on existing techniques, we provide insights into the future advancement of CI with latent variables, especially the new opportunities with large language models (LLM).

2 Preliminaries

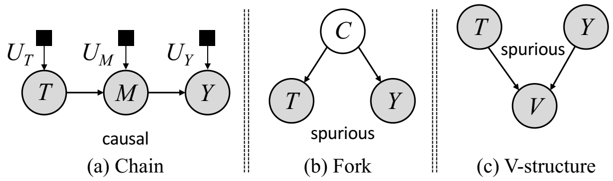
2.1 Symbol System

For most CI tasks, there are two main variables of interest, i.e., treatment T and outcome Y, on which the causal relation is scrutinized. We consider T as a binary variable by default, but the cases of continuous/multiple/high-dimensional treatments will also be covered in detail. The outcome Y can be arbitrary results of interest under the potential causal influence of T. In addition, we use X to denote other observed covariates in the system, which may have certain causal relations with T and Y depending on the context.

2.2 Rubin's Causal Model

To study the causal relation between treatment T and outcome Y, Rubin's causal model (SCM) starts by comparing individual-level counterfactuals, i.e., for unit i, what the outcome Y is if the unit is treated ($T=1$) or is not treated ($T=0$). Although the two results cannot be observed for the same unit i simultaneously, we can still hypothetically define them as potential outcomes as follows:

Definition 2.1. ***(Potential Outcome).*** We use the notations $\{Y_i(T=1), Y_i(T=0)\}$ (which are shortened as $Y_i(1), Y_i(0)$ if the treatment is clear from the context) to denote the potential outcomes (PO) of Y for unit i if the treatment $T=1$ or 0 is imposed on the unit.



(a) Chain (b) Fork (c) V-structure

Figure 1: Atomic structures of SCM, where mutually independent exogenous variables $\mathcal{U}$ are omitted for (b) and (c).

Accordingly, **R.V.** $Y(T=1)$, $Y(T=0)$ reason with the distribution of the POs if all units are uniformly treated or non-treated (i.e., interventions). However, $Y(T=1)$, $Y(T=0)$ cannot be obtained due to lack of individual counterfactuals. To estimate $Y(T=1)$, $Y(T=0)$, most strategies need to collect the outcomes of two groups of treated and non-treated units. Here, we use conditional R.V. $Y|T=1$ and $Y|T=0$ to denote the distribution of Y for the two groups. Only in rare cases, e.g., randomized experiments, can $Y|T=t$ provide an unbiased estimate for $Y(T=t)$. In other cases, the purpose of RCM is to show that under certain assumptions, causal estimand with PO can be reduced to conditional relations measurable in the data (usually involving other covariates X).

2.3 Structural Causal Model

Pearl's structural causal model (SCM), in contrast, reasons with causality via a pre-defined direct acyclic causal graph **G** that encodes the belief of causal relations among variables of interest [96]. Based on the causal graph, SCM can be formally defined as follows:

Definition 2.2. ***(SCM).*** Structural causal model (SCM) can be defined as a triplet of sets $(\mathcal{U}, \mathcal{V}, \mathcal{F})$, where $\mathcal{U}$ is the set of latent exogenous variables, $\mathcal{V}$ is a set of observed endogenous variables, and $\mathcal{F}$ is a set of structural equations. For an endogenous variable $V \in \mathcal{V}$, we have $V = f_V(\mathcal{A}_U(V), \mathcal{A}_V(V))$, where $\mathcal{A}_U(V), \mathcal{A}_V(V)$ are the exogenous, endogenous parents of V in **G**, respectively.

In SCM, each unit i is associated with a set of exogenous variables $\mathcal{U} = \mathcal{U}_i$ that causally determines the endogenous variables, e.g., T, Y, X. The prior for $\mathcal{U}$ is $p(\mathcal{U})$. Mutually-independent exogenous variables are usually ignored when average causal effects are considered, but they are vital for counterfactual reasoning since they represent unit variations. Three atomic structures exist in a causal graph (Fig. 1): ***(i)*** chains $T \longrightarrow M \longrightarrow Y$, ***(ii)*** forks $T \longleftarrow C \longrightarrow Y$, and ***(iii)*** V-structure $T \longrightarrow V \longleftarrow Y$. T and Y are correlated if *mediator* M is not unobserved for chains (causal), *confounder* C is not observed for forks (not causal), and *collider* V is observed for V-structures (not causal). Therefore, to distinguish causation from correlation, Pearl introduces the *do*-operator, where $p(Y|do(T=t))$ means that we set $T=t$ as an intervention and calculate Y via $f_Y(T=t, \mathcal{A}_U(Y), \mathcal{A}_V(Y)/T)$, regardless of observed parents of T.

2.4 Connections between SCM and RCM

If an SCM is correctly specified, potential outcome $Y_i(T=t)$ can be derived by ***(i)*** replacing the structural equation f_T in $\mathcal{F}$ with $f_T^{do} = t$ (i.e., intervention), which results in a new set of structural equations $\mathcal{F}^{do}$, ***(ii)*** setting the exogenous variables $\mathcal{U} = \mathcal{U}_i$ (i.e., the individual factors for unit i), and ***(iii)*** calculating the outcome Y based on $\mathcal{U}_i$ and the new structural equations $\mathcal{F}^{do}$. R.V. $Y(T=t)$ can be similarly derived by using the prior of $\mathcal{U}$, i.e., $p(\mathcal{U})$, instead of $\mathcal{U}_i$. Therefore, the two frameworks are fundamentally equivalent.

2.5 Overview of Causal Inference Tasks

2.5.1 *Treatment Effect Estimation.*

Treatment effect estimation aims to quantitatively measure the causal influence of treatment T (e.g., drug) on outcome Y (e.g., survival rate). The most commonly used metric is the **average treatment effect (ATE)** [4], which is the expected causal effect of T on Y for the entire population. ATE can be formulated via the two frameworks as follows:

$$ATE = \mathbb{E}[Y|do(T=1)] - \mathbb{E}[Y|do(T=0)] = \mathbb{E}[Y(T=1) - Y(T=0)]. \tag{1}$$

For a pretreatment variable X (e.g., age), we can also define the **conditional average treatment effect (CATE)** [2] as follows:

$$CATE(X) = \mathbb{E}[Y|do(T=1), X] - \mathbb{E}[Y|do(T=0), X], \tag{2}$$

which is important when the treatment effect is heterogeneous, i.e., different sub-populations X have different responses to treatment.

2.5.2 *Causal Mediation Analysis.*

Causal mediation analysis aims to quantitatively study the fine-grained causal relationship between treatment T (e.g., drug) and outcome Y (e.g., survival rate) mediated by certain factors M (e.g., blood pressure) [53]. When there is only one mediator, the most common metric is the **natural indirect effect (NIE)**, which can be formulated as follows [99]:

$$NIE = \mathbb{E}[Y(M(T=1), T=0))] - \mathbb{E}[Y(M(T=0), T=0)]. \tag{3}$$

Here, $Y(M(T=t), T=0))$ is a nested potential outcome (NPO) denoting three interventions: ***(i)*** $T \longleftarrow t$ along the path $M \longleftarrow T$, ***(ii)*** $T \longleftarrow 0$ along the path $Y \longleftarrow T$, and ***(iii)*** $M \longleftarrow M(T=t)$ along path $Y \longleftarrow M$. NIE excludes the direct effect of T along path $T \longrightarrow Y$, while enabling the indirect effect of T mediated by M. Furthermore, NIE can be generalized to an arbitrary causal path, i.e., path-specific causal effect, which can be defined in a similar way via NPO [53].

2.5.3 *Counterfactual Reasoning.*

Counterfactuals can be broadly defined as causal estimands (represented by *do*-operator or potential outcomes) that *contradict* the factual observations (represented by conditional distributions). For example, the **average treatment effect on the treated (ATT)** is defined as follows:

$$ATT = \mathbb{E}[Y(1)|T=1] - \mathbb{E}[Y(0)|T=1]^{1}. \tag{4}$$

Here, $\mathbb{E}[Y(0)|T=1]$ in Eq. (4) denotes the expected outcome Y in a counterfactual world where the treated units (denoted by the condition $T=1$) had not been treated (denoted by $Y(T=0)$).

2.5.4 *Causal Discovery.*

Given variables of interest, causal discovery aims to recover the causal graph **G** given the observed data, such that the parent nodes are the direct cause of the child [119].

3 Treatment Effect Estimation

In this section, we discuss the treatment effect estimation with latent variables. Specifically, we mainly focus on the latent confounders, which can systematically bias the estimation if handled carelessly. We first introduce traditional methods where confounders are assumed to be fully observed. We then discuss the *circumvention*-based and *inference*-based strategies to handle latent confounders.

[1] Here, please note ***consistency*** is always assumed, i.e., $\mathbb{E}[Y(1)|T=1] = \mathbb{E}[Y|T=1]$.

3.1 Brief Review of Traditional Methods

Traditional treatment effect estimation methods assume away the latent confounders via the following ignorability assumption:

Assumption 1. ***(Ignorability).*** $Y(T=0,1) \perp\!\!\!\perp T|X$.

Combined with other common assumptions for CI (e.g., positivity, non-interference, etc. see [55]), ATE and CATE can be identified from observational data by controlling X as follows:

$$CATE(X) = \mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X], ATE = \mathbb{E}_{p(X)}[CATE(X)] \quad (5)$$

From the SCM's perspective, X blocks all backdoor paths that lead to spurious correlations between T and Y (see Section 2.3), such that in each stratum of $X = x$, the correlation between T and Y is causal. Based on Eq. (5), adjustment-based methods use non-parametric methods [8] or fit parametric models $f(t, X)$ (which will be denoted as $f_t(X)$ if different models are used for different T) to estimate $\mathbb{E}[Y|T=t,X]$, including linear models [55], tree-based methods [133], and deep neural networks (DNN) [114]. Another line of methods reweights samples via inverse propensity score $\mathbb{E}[T = t|X]$, such that they can be viewed as pseudo-random samples.

3.2 Circumvention-based Methods

However, if important confounders C are missing from the observed covariates X, Assumption 1 failed, and Eq. (5) is a biased estimation for C/ATE. To address the latent confounding bias, circumvention-based methods show that, under certain stringent conditions, causal effects can still be unbiasedly estimated without the direct or indirect measurement of latent confounders or their proxies.

3.2.1 *Small Randomized Data.*

If a small amount of randomized data is available (which cannot be directly used to estimate CATE due to high variance), we can use them to correct the bias in large-scale observational data with latent confounders [24, 101, 124, 129]. One exemplar work is [60], which first fits a biased CATE estimator $f_t^{obs}(X)$ on the observational data as Eq. (5) and correct the bias with another estimator $e_t^{exp}(X)$ fitting on the error of $f_t^{obs}(X)$ evaluated on the randomized data. Since the value of the bias is usually smaller than the CATE, the estimation variance can be reduced compared to directly fitting the CATE estimator on the small-scale randomized data. In contrast, Yang et al. [149] directly tackles confounding bias in the observational data. They define the latent confounding bias with the confounding function as follows:

$$\lambda(X) = \mathbb{E}^{obs}[Y(0)|T=1,X] - \mathbb{E}^{obs}[Y(0)|T=0,X], \quad (6)$$

which measures the systematic difference of the expected baseline PO $Y(0)$ between the treatment/non-treatment group in the observational data. They further show that under the ***transportability*** assumption, i.e., the treatment effect is the same between the randomized samples and the treated samples in the observational data, the confounding bias $\lambda(X)$ can be estimated as follows[2]:

$$\begin{aligned}\lambda(X) =& (\mathbb{E}^{obs}[Y|T=1,X] - \mathbb{E}^{obs}[Y|T=0,X]) - \\ & (\mathbb{E}^{exp}[Y|T=1,X] - \mathbb{E}^{exp}[Y|T=0,X]).\end{aligned} \quad (7)$$

Correction of the above bias upon the biased CATE estimator fit on the observational data leads to an unbiased CATE estimator with variance lower than direct estimation with randomized data. Recently, Wu and Yang [143] improved over [149] by adopting the R-learner [92] to model the confounding function, which allows flexible ML models such as trees and DNNs as the estimator.

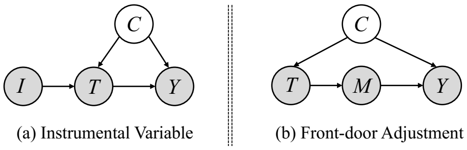


Figure 2: SCM for IV methods and front-door adjustment

The advantage of using randomized data to tackle latent confounding is that the randomized data are guaranteed to be unbiased (but with high variance due to small scale). However, these methods fail in the case where even a small number of randomized samples cannot be obtained, e.g., when the dataset was collected in the past.

3.2.2 *Instrumental Variable.*

If randomized data are not possible, we can use instrumental variables (IV) to "extract" pseudo-randomized data embedded inside the observational dataset to unbiasedly estimate ATE/CATE. Formally, IV is defined as follows:

Definition 3.1. ***(Instrumental Variable, IV)*** A variable that ***(i)*** has no confounding with the outcome Y, ***(ii)*** affects the treatment T (relevance), ***(iii)*** affects the outcome Y only through T (restriction).

For a binary IV, ATE can be unbiasedly estimated via [40]:

$$\hat{ATE} = (\mathbb{E}[Y|I=1] - \mathbb{E}[Y|I=0])/(\mathbb{E}[T|I=1] - \mathbb{E}[T|I=0]), \quad (8)$$

if we view I as the assigned treatment and T as the treatment received, the numerator can be viewed as the intention-to-treat effect of the treatment assignment (I) on outcome (Y), and the denominator as the compliance with the assigned treatment. General IV-based methods follow a similar two-stage procedure. Assuming linear causal relations, the two-stage least squares algorithm (2SLS) ***(i)*** first calculates the conditional mean of the treatment T given the IV I, i.e., $\hat{T} = \mathbb{E}[T|I]$, and ***(ii)*** regresses Y on $\hat{T}$, where the coefficient gives the causal relation between Y and T [5]. Afterward, efforts have been devoted to generalizing 2SLS to nonlinear cases [30, 85, 116]. For example, Deep IV [48] estimates the conditional density in $\hat{T} = \mathbb{E}[T|I,X]$ from stage ***(i)*** with categorical distribution (for discrete T) or mixture of Gaussian distribution (for continuous T) parameterized by DNN, and predicts the outcome Y in stage ***(ii)*** via another DNN $\hat{Y} = f_{nn}(\hat{T}, X)$. However, to make the objective optimizable, Deep IV assumes simple distributions, which fail when the treatment T is high dimensional. To address this issue, Bennett et al. [13] proposed to use a generalized method of moments to allow more flexible DNNs as treatment/outcome networks [47].

However, finding suitable IVs is still difficult. Recently, Yuan et al. [151] proposed the Auto IV, which finds IVs $\hat{I}$ from candidates X that satisfy Definition 3.1 by maximizing the mutual information (MI) between $\hat{I}$ and T to ensure *relevance*, and minimizing the conditional MI between $\hat{I}$, Y given T to ensure the *restriction* criteria.

The advantages of IV-based methods are that ***(i)*** no randomized data are required to address latent confounding, and ***(ii)*** mature methods exist with good theoretical properties. However, it is difficult to find IV that satisfies the Definition 3.1. In addition, if the IV is weak, i.e., has mild influences on the received treatments, the estimation will have a high variance even with large data.

[2] Proof is straightforward with consistency and transportability assumptions.

3.2.3 ***Front-door Adjustment***. In addition, if the causal mechanism between the treatment T and outcome Y is known, i.e., all mediators M are observable and unconfounded with T and Y, front-door adjustment can be used to address latent confounders [95]. Specifically, based on the probability theory, we have

$$\mathbb{E}[Y|do(T)] = \mathbb{E}_{p(M|do(T))}[Y|do(M)]. \tag{9}$$

Since no backdoor path exists between T and M, $p(M|do(T)) = p(M|T)$. In addition, since T blocks the backdoor path between M and Y, $P(Y|do(M)) = \mathbb{E}_{P(T)}[P(Y|M,T)]$, where Eq. (9) is reduced to conditional relations measurable from the data. However, similar to IV-based methods, mediators that satisfy the front-door criterion are difficult to find. Therefore, Xu et al. [147] proposed to infer latent mediators that satisfy the front criterion from the covariate X with the identifiable variational auto-encoder (iVAE) [61].

3.2.4 ***Multiple Causes***. Finally, we consider the case of multiple treatments, where we are interested in estimating the combined causal effects of all the treatments in T (e.g., prescribing bundled drugs) on Y. If we can determine that the latent confounders are shared among different treatments (i.e., single-cause ignorability [139]), various methods can be used to address the confounding bias. The deconfounder-based methods prove that if latent variables Z can be found that render different treatments conditional independent, controlling Z adjusts for the confounding bias due to multi-cause confounders C^m [139]. The proof is simple and elegant: if C^m are still active after conditioning on Z, they will render the treatments dependent (see Section 2.3), which results in contradiction. Linear models [140] and DNNs [110, 163, 164] are used to estimate Z from T. Recently, Ma et al. [78] proposed to learn latent Z with latent clustering, which can well accommodate new treatments. Observing that under single-cause ignorability assumption, the data is unconfounded for every single cause, Qian et al. [104] proposed to learn a single cause interventional model for each cause, and perturb the cause to generate counterfactually-augmented datasets, which they show are beneficial to learn multi-cause models.

3.3 Inference-based Methods

In this subsection, we introduce inference-based methods, which assume that even if confounders C cannot be directly observed, we can observe their proxies W, which could be conducive to the inference of latent confounders to address confounding bias [126].

3.3.1 ***Proxy-based Methods***. Statistical methods generally assume simple forms of C and its causal relations with observed proxies W [98]. However, even for the simplest relation, i.e., $C \rightarrow W$, directly controlling W leaves the backdoor path $T \leftarrow C \rightarrow Y$ open, which cannot adjust for all the confounding bias. To address this, Kuroki and Pearl [66] assumed that W contains two independent views of C to recover $p(W|C)$, such that $p(Y|do(T))$ can be identified from $p(W|C)$ and other observable relations. Miao et al. [83] relax the assumption, allowing for an unbiased estimation without recovering the confounder measurement error mechanism $p(W|C)$.

However, CI usually faces high-dimensional confounders and proxies W, where the statistical methods may fail to scale up to. Observing that even if C is complex, only a small part of the information in C is necessary to adjust for confounding (e.g., if $f(C)$ preserves the propensity score, i.e., $\mathbb{E}[T|C] = \mathbb{E}[T|f(C)]$, adjusting for $f(C)$ in Eq. (5) still gives an unbiased estimate of ATE/CATE [108]), Kallus et al. [59] proposed to use matrix factorization (MF) to obtain low-rank components of W, which they show are better approximations of true confounders C. Louizos et al. [72] proposed the causal effect variational auto-encoder (CEVAE), which uses the VAE [64] to recover the joint distribution $p(C,T,W,Y)$ and infer latent confounders C from the observations $\{T,W,Y\}$.

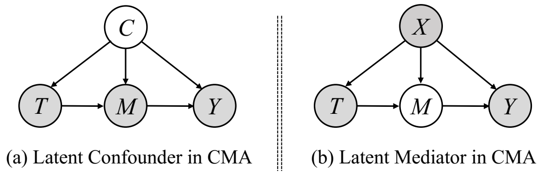
In addition to unbiasedness, other aspects need to be carefully considered as well. To address high variance due to non-overlapping covariate W (i.e., certain values of W appear only when $T = 0$ or $T = 1$, which is common if W is high-dimensional), Wu and Fukumizu [144] proposed to map W to low dimensional space with better overlapping w.r.t. the prognostic score [46] based on the iVAE [61], which is sufficient for the identification of ATE/CATE. Furthermore, in certain cases, we cannot identify latent variables Z that lead to an unbiased estimation of ATE/CATE. Therefore, Hu et al. [51] proposed an adversarial learning [38]-based method to bound the error by finding the max/min values of the possible ATE via a generator and ensures that the distribution parameterized by the generator is faithful to the observations via another discriminator.

3.3.2 ***Covariate Disentanglement***. If the proxy W scrambles variables other than latent confounders C, various issues could be incurred if naively using the above-introduced proxy-based methods. To address this issue, covariate disentanglement (CD) methods are proposed to further scrutinize the latent variables Z that generate the proxy W. Most methods assume that W is generated from three types of latent variables: IVs I (see Definition 3.1), confounders C, and adjusters A, i.e., variables that causally influence only the outcome Y. Previous work has proven that controlling C eliminates confounding bias, controlling A could reduce estimation variance, while controlling IVs I could *increase* the variance [50, 88].

To address the issue, most methods rely on the statistical property between $\{I,C,A\}$ and $\{T,Y\}$: IVs I are correlated with only the treatment T, adjusters A are correlated only with the outcome Y, while confounders C are correlated with **both** T and Y. To leverage this property, DR-CFR [49] designs three encoders to infer three sets of latent variables $\hat{I}, \hat{C}, \hat{A}$ from W, which are learned by making $\hat{I}, \hat{C}$ predictive for T (i.e., maximizing $p(T|\hat{I},\hat{C})$) and $\hat{C}, \hat{A}$ predictive for Y (i.e., maximizing $p(Y|\hat{A},\hat{C})$). Similarly, TEDVAE [155] splits the encoder in CEVAE into three parts for $\hat{I}, \hat{C}, \hat{A}$, respectively, and maximizes $p(T|\hat{I},\hat{C})$, $p(Y|T,\hat{A},\hat{C})$ to achieve disentanglement.

The disentanglement can also be achieved by relying on other properties. For example, assuming I, C, A are separable in the observed proxy W, AFS [136] shows that $Z = \{C, A\}$, provided that the efficient influence curve of Z, $D^{eff}(Z)$ [130] is minimized. The empirical estimate of $D^{eff}(Z)$ is then used as the reward to learn a mask to select C, A from W. In addition, NICE [115] uses invariant risk minimization (IRM) [7] to find all causal parents of Y (including C, A), which can effectively exclude IVs from the control set.

However, the above methods fail when latent post-treatment variables M' are scrambled in the proxy W, as similar to C, post-treatment variables M' can be correlated with **both** the treatment T and the outcome Y, and can be the causal parents of Y. Recently, CiVAE [161] was proposed to disentangle C from M'. Specifically, after individually identifying latent variables $\hat{Z}$ that generate W, independence tests are conducted for each pair of $\hat{Z}_i, \hat{Z}_j$, and the pairs



(a) Latent Confounder in CMA

(b) Latent Mediator in CMA

Figure 3: SCM for latent variables in CMA

with increased correlation after conditioning on T are selected as confounders. In contrast, if both $\hat{Z}_i, \hat{Z}_j$ are post-treatment variables or one is a post-treatment variable and another is a confounder, the correlation will decrease after conditioning on T (see [96]).

4 Causal Mediation Analysis

Causal mediation analysis (CMA) [103] aims to understand the fine-grained causal mechanism between treatment T and outcome Y by identifying the effect mediated by another factor M, which mediates the causal effect from the treatment to the outcome.

4.1 Brief Review of Traditional Methods

Traditional CMA identifies the causal mediation effect based on the assumptions of *measurable* confounders and mediators. Specifically, unobserved confounders are assumed away via the Sequential Ignorability assumption defined as follows:

Assumption 2. ***(Sequential Ignorability, SI). (i)*** *$M(t), Y(t,m) \perp\!\!\!\perp T|X$; **(ii)** $Y(t,m) \perp\!\!\!\perp M(t)|T, X$ for all t, m.*

Intuitively, SI assumption states no unobserved confounding between ***(i)*** the treatment T and the mediator M, ***(ii)*** the mediator M and the outcome Y, and ***(iii)*** the treatment T and the outcome Y. In addition, since M is the factor of which the mediated effect is interested in, it should be observed and measurable.

Assumption 3. ***(Measurable Mediator).*** *M is observed.*

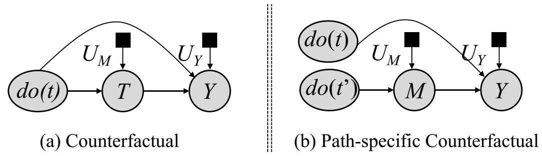
With Assumptions 2, 3, the causal effect mediated by M in the form of natural indirect effect (NIE, see Eq. (3)) can be calculated as

$$\begin{aligned} NIE = \mathbb{E}_{p(X)} \big[& \mathbb{E}_{p(M|T=1,X)} \left[Y|T=0,X,M \right] - \\ & \mathbb{E}_{p(M|T=0,X)} \left[Y|T=0,X,M \right] \big], \end{aligned} \tag{10}$$

which holds $T = 0$ fixed on the direct path $T \longrightarrow Y$, and change T from 0 to 1 on the indirect path $T \longrightarrow M \longrightarrow Y$. The natural direct effect (NDE) of T on Y can be calculated as $ATE - NIE$. In practice, the conditional distributions required by NIE and NDE can be estimated using various methods, such as linear regression [12], logistic regression [82], or machine learning techniques [53], such as decision trees and deep neural networks [34, 53].

4.2 Latent Confounders in CMA

If latent confounders C exist and are not included in X, the sequential ignorability assumption breaks and traditional methods that rely on Eq. (10) to estimate NIE will give biased results. To address the issue, various proxy-based methods are proposed. Here, an exemplar work is causal mediation analysis variational auto-encoder (CMAVAE) [17], which assumes that the latent confounder C confounds all pair-wise relations among T, Y, M, of which a noisy proxy W can be observed (definition of proxy see Section 3.3.1). Inspired by CEVAE [72], they prove that the NIE can be identified by estimating the joint distribution $p(C, W, M, T, Y)$, which are parameterized with DNNs, where the posterior distribution $q(C|W, M, T, Y)$ is obtained via variational inference [64]. Finally, they sample latent confounders C from $q(C|W, M, T, Y)$ to unbiasedly estimate NIE.



(a) Counterfactual

(b) Path-specific Counterfactual

Figure 4: SCM for (path-specific) counterfactuals

4.3 Latent Mediation Analysis

In this part, we discuss CMA with latent mediators, where Assumption 3 fails, and Eq. (10) cannot be used directly for estimation.

4.3.1 ***Circumvention-based Methods.*** Circumvention-based methods are difficult in CMA with latent mediators. Nevertheless, Derkach at. al. [29] proposed a method without any utilization of observable proxies of latent mediators. They made a strong assumption that the distribution of Y belongs to an exponential family $f(Y; \xi, \phi_Y) = exp[Y\xi_Y - b(\xi_Y)]/a(\phi_Y) + c(Y, \phi_Y)$, where ξ_Y and ϕ_Y are modeled as functions of the latent mediator. They observe that NIE can be represented by the parameters of the distribution. Therefore, to estimate the NIE, it is sufficient to estimate the parameters of $f(Y; \xi, \phi_Y)$, which is solved via an expectation-maximization (EM) algorithm: In each iteration of the algorithm, the expected latent mediators are first calculated, then the distribution parameters are updated via likelihood maximization.

4.3.2 ***Proxy-based Methods.*** If the mediator of interest M is not directly observed, utilizing its observed proxies W is an effective method for NIE estimation. Kuroki et al. [65, 66] showed that the NIE of treatment T on an outcome Y could be identified in linear models given two independent proxies of an unobserved mediator. In addition, Albert et al. [3] proposed a maximum likelihood-based approach to estimate causal mediation effects with a continuous latent mediator measured by multiple observed proxies. Their method is based on fitting a generalized structural equation model (GSEM) [86] using an approximate Monte Carlo EM algorithm. The fitted GSEM is then used to estimate natural direct and indirect effects [97]. In addition to the latent mediator, this approach also accommodates mediator-outcome confounding and mixed continuous and categorical outcomes. However, it relies on parametric modeling assumptions and may be computationally intensive. Recently, Sun et al. [123] proposed a joint modeling approach that incorporates multiple latent mediators and a survival outcome. Specifically, a Bayesian approach with a Markov chain Monte Carlo algorithm is developed to perform an efficient estimation of NIE.

5 Counterfactual Analysis

In this section, we focus on counterfactual reasoning, which aims to explain the outcomes of a specific individual if a different treatment was taken in the past. According to SCM, exogenous variables U contain individual varieties (see Section 2.3). Therefore, counterfactual reasoning under SCM is naturally a latent variable problem

[96]. Please note that throughout this section, we assume Sequential Ignorability (see Assumption 2) holds, so that we can devote the main discussions to the latent exogenous variables.

5.1 Overview

With a pre-defined SCM $\mathbf{G}$, counterfactuals generally have the following form: $\mathbb{E}[Y_U(T = t')|W, T = t]$, where W is the evidence (observed values for variables in $\mathbf{G}$) and $T = t$ is the observed treatment. Here, we use the subscript U in $Y_U(T = t')$ to denote the dependence of the PO $Y(T)$ on exogenous variables U. The counterfactual inference involves three steps as follows:

- ***(abduction)*** The prior of U, i.e., $p(U)$, is updated into posterior $p(U|W, T)$ based on the observed W and $T = t$.
- ***(action)*** Structural equation f_T is substituted with $f_{do}(T) = t'$.
- ***(prediction)*** The outcome Y is computed with $p(U|W, T = t')$, $f_{do}(T) = t'$, and other structural equations in $\mathcal{F}$.

The key to counterfactual inference lies in the abduction of latent exogenous variables U, as the other two steps are straightforward.

5.2 Circumvention-based Methods

We can circumvent latent exogenous variables U if the counterfactuals of interest are not required to be qualitatively determined. For example, when studying counterfactual fairness of ML models, we only need to judge whether two counterfactuals are the same:

$$\mathbb{E}[\hat{Y}_U(T = t')|W = w, T = t] \overset{?}{=} \mathbb{E}[\hat{Y}_U(T = t)|W = w, T = t]. \quad (11)$$

Intuitively, Eq. (11) asks that given evidence $W = w$ and $T = t$ (where T could be the sensitive features such as race, gender, etc.), whether the prediction $\hat{Y}$ would be the same for a unit U if T is set to another value t'. Kusner et al. [67] showed that Eq. (11) holds when predictor $\hat{Y}$ does not use any descendant of T, which precludes the dependence of $\hat{Y}$ on T. Afterward, Chiappa [18] proposed path-specific counterfactual fairness (PSCF), which allows the causal influence of T on $\hat{Y}$ along certain causal paths. For example, in the single mediator case, we may allow $T \longrightarrow M \longrightarrow \hat{Y}$ while forbid $T \longrightarrow \hat{Y}$, where M is called a resolving variable [63]. In this case, the counterfactual question can be formulated via NPOs as follows:

$$\mathbb{E}[\hat{Y}_U(M(t), T = t')|W = w, T = t] \overset{?}{=} \mathbb{E}[\hat{Y}_U(M(t), T = t)|W = w, T = t]. \quad (12)$$

If Eq. (12) holds, the predictor $\hat{Y}$ is precluded from using the descendants of T along the unfair paths (T itself included) [89].

5.3 Inference-based Methods

In other cases, when counterfactuals need to be calculated or bounded, inference-based methods become more useful [160]. Observing that exogenous variables U satisfy exactly the non-descendant requirement of T while containing all individual information, Kusner et al. [67] assumed linear structural equations and fitted linear additive models on the observed data, where the error terms are viewed as the estimand of U and used for fair predictions. Zuo et al. [165] further generalized [67] to the case of partially observed SCMs, under the assumption that T has no endogenous ancestor. Wu et al. [146] proposed to bound Eq. (12) by dividing the U space into equivalent regions via response functions [10], and search the upper and lower limit of Eq. (12) while making the response functions compatible with the observed W and T. To achieve PSCF, Chiappa [18] proposed to use VAE to infer U and use it to correct the dataset by setting T of all samples to the baseline value along the unfair paths.

6 Causal Discovery

Previous sections primarily focus on CI with a *pre-defined* causal graph. However, when accurate causal relations cannot be obtained (e.g., lack of domain knowledge), it becomes imperative to automatically discover the causal relations from data via causal discovery (CD). In this section, we first introduce traditional CD methods, including constraint-based and score-based methods. We then discuss CD strategies when unobserved confounders exist.

6.1 Brief Review of Traditional Methods

Causal discovery (CD) aims to infer causal relations among variables of interest $\mathcal{V}$ from the observational dataset, with the goal of constructing a causal graph $\mathbf{G} = (\mathcal{V}, \mathcal{E})$. Most traditional CD methods rely on the assumptions of faithfulness and causal sufficiency (which assume away unobserved confounders) as follows:

Assumption 4. ***(Faithfulness).*** *If two disjoint sets of variables $\mathcal{M}$ and $\mathcal{N}$ are independent in the distribution P when conditioning on $\mathcal{Z}$, then it implies that $\mathcal{M}$ and $\mathcal{N}$ are d-separated [118] in the graph $\mathbf{G}$ conditioning on $\mathcal{Z}$, denoted as: $\mathcal{M} \perp\!\!\!\perp_P \mathcal{N}|\mathcal{Z} \Longrightarrow \mathcal{M} \perp\!\!\!\perp_{\mathbf{G}} \mathcal{N}|\mathcal{Z}$.*

Assumption 5. ***(Causal Sufficiency).*** *For any two observed variables V_i and V_j in the data, all common causes must also be observed.*

Generally, CD methods can be categorized into two classes: ***(i)*** constraint-based and ***(ii)*** score-based methods [93]. Constraint-based methods use conditional independence tests to identify edges in the graph based on the faithfulness assumption. For example, the Peter-Clark (PC) algorithm [118] and its variants [14, 27, 68, 137] first identify an undirected causal graph (i.e., *skeleton*) by removing edges from a complete causal graph with conditional independence tests, and then determine the edge direction by a set of orientation propagation rules with V-structures and acyclicity property [118]. For example, consider a path in the skeleton $A - B - C$, where A and C are not adjacent. If A and C became dependent conditioning on B, then the PC algorithm orients the edges as $A \longrightarrow B \longleftarrow C$ based on the property of V-structures (see Section 2.3). In contrast, score-based algorithms [19, 106, 141] aim to identify the best candidate graph by maximizing a fitness score, such as the Bayesian Information Criterion (BIC), to discover the causal graph from the data.

6.2 Proxy-based Methods

There are a few proxy-based methods for CD with unobserved confounders. Liu et al. [71] studied causal discovery between two variables T, Y with a latent confounder U. Assuming a proxy W, i.e., a causal descendant of U, can be observed, they discretize W and use [83] introduced in Section 3.3.1 to estimate the causal effect and judge whether an edge exists between T and Y. However, such proxies may not exist in reality. Recently, [70] introduced time series data to address the issue, where each variable is assumed to be its causal parent in the next time step, serving as the proxy W.

6.3 Circumvention-based Methods

6.3.1 ***Constraint-based Methods.*** If unobserved confounders exist, the causal sufficiency assumption will not hold, and naive independence tests (i.e., correlation) cannot indicate causal relations among variables of interest. To address this issue, Spirtes et al. [120] proposed the FCI algorithm, which extends the PC algorithm by introducing three more relations (in addition to $X \rightarrow Y$) to model the uncertainty regarding confounders: ***(i)*** $X \leftrightarrow Y$ indicates the presence of unmeasured confounders; ***(ii)*** $X\circ \;\rightarrow\; Y$ represents that either X causes Y or there are unmeasured confounders; ***(iii)*** $X \circ - \circ Y$ can represent any of the following scenarios: (1) X causes Y, (2) Y causes X, or (3) there are unmeasured confounders, where a new orientation rule [118, 152] is used to orient edges.

Subsequent research has been proposed to extend FCI from various perspectives, such as with enhanced efficiency [25, 26, 117], tailored for sparse causal graphs [121], improved scalability [105], and incorporating different conditional independence tests [56].

6.3.2 ***Score-based Methods.*** Score-based algorithms, e.g., GES and Fast GES (FGES) [26], find optimal causal graph by greedily adding and deleting edges based on predefined scores measuring the fitness of a graph on observational data. However, they face challenges when latent confounders exist (i.e., causal insufficiency). To address the issue, the recent trend is to use confounders-robust constraint-based methods such as FCI to correct the bias. However, it underperforms when the sample sizes are small due to an inaccurate estimation of the independence relations [120]. The Greedy FCI (GFCI) algorithm [94] combines the strengths of both approaches. It uses GES to identify a supergraph of the skeleton, then employs FCI to prune the supergraph and determine the orientations to handle unmeasured confounders. This integration enhances performance while maintaining asymptotic correctness under causal insufficiency. However, GFCI's scoring function cannot be applied to mixed variables, which is addressed by the Bayesian Constraint-Based Causal Discovery (BCCD) algorithm [22] via utilizing a hybrid constraint and score-based approach for causal search.

7 Future Directions

In this section, we discuss several promising directions to further advance causal inference studies with latent variables and discuss new opportunities in the era of large language models (LLM).

7.1 On Theories and Model Design

Firstly, there has been a growing interest in causal representation learning [112], which aims to develop models capable of automatically extracting and representing causal concepts and relations from data. An in-depth study of causal representation learning with latent variables would be interesting in real-world applications.

Additionally, integrating multi-modal information of the unit, such as textual [132, 158], visual [148], and sensor data [128], offers opportunities to compensate for the absence of observation in a single modality, which also increases the chance of finding applicable circumvention or inference methods to address the latent variable.

Furthermore, improving the interpretation [20, 33, 42] (especially towards latent variables) is essential to foster trust and transparency in causal learning systems with latent variables, allowing users to comprehend and validate causal conclusions effectively.

Finally, exploring uncertainty quantification techniques [1], such as conformal prediction [113], can provide valuable information on the reliability and robustness of CI under latent variables, facilitate more informed decision-making, and provide pessimistic/optimistic bounds when exact causal effects cannot be identified.

7.2 Opportunities in the LLM Era

Recently, large language models (LLM) exhibit remarkable in-context learning and reasoning capabilities [107, 134, 138, 156, 162]. Although LLM itself is nowhere causal [153] (after all, it still fits conditional distributions parameterized by transformer networks on corpora [131]), recent research has shown some promising results of LLMs to facilitate CI, e.g., causal reasoning [58], counterfactual analysis [154], and causal discovery [11, 23, 62]. For example, Jin et al. [58] showed that when provided with few-shot examples with chain-of-thought (CoT) [142] causal reasoning steps in the prompts, LLMs can construct causal graphs, formulate causal questions with the two frameworks and manage to solve it with observational data.

Based on the above examples, we speculate that LLMs can also provide opportunities to advance CI with latent variables. Here, we provide the following interesting future perspectives. ***(i)*** First, it is promising to see LLM facilitate the automatic identification of important latent variables that could be neglected by human beings. As such, issues of neglecting important variables can be prevented in advance. ***(ii)*** In addition, if the absence of important variables is inevitable, LLM may have the potential to reason with new strategies to circumvent or infer the variables from proxy based on the reasoning ability to the causal relation of latent variables and observed variables at hand (i.e., automatic causal discovery). ***(iii)*** Furthermore, LLM may provide a usable and user-friendly interpretation of latent variable models for CI [145], as well as how biases are generated and eliminated. ***(iv)*** Finally, recent advances in multi-modal LLM [32] are also promising to systematically consider multi-modal features of a unit, where the more comprehensive causal graph can be established by the LLM to increase the chance of finding good solutions to address the latent variables.

8 Conclusions

In this survey, we review recent advances in causal inference (CI) with latent variables, covering four main CI tasks, i.e., causal effect estimation, causal mediation analysis, counterfactual reasoning, and causal discovery. We start by briefly reviewing CI methods where important variables are assumed to be observed. Then, under the new taxonomy of inference-based and circumvention-based methods, we introduce methods that account for the absence of crucial variables. Furthermore, we generalize the above method to graphs, an important area for machine learning. Finally, we discuss future perspectives, especially the new opportunity in the LLM era.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) under grants IIS-2006844, IIS-2144209, IIS-2223769, IIS-2316306, CNS-2154962, and BCS-2228534; the Commonwealth Cyber Initiative Awards under grants VV-1Q23-007, HV-2Q23-003, and VV-1Q24-011; the JP Morgan Chase Faculty Research Award; the Cisco Faculty Research Award; and Snap gift funding.

Appendix

A Generalization to Graph Data

Causal inference on graph data (e.g., social networks) naturally faces unique challenges compared with traditional tabular data due to the intrinsic interconnection and interactions among units under study. In the last few decades, there have been substantial efforts in marrying causal inference with graph mining [28, 75, 79], where latent variables still severely impede the robustness and trustworthiness of causal conclusions. Here, we extend the methodology introduced in the main paper to graph data.

Treatment Effect Estimation. Estimating treatment effect on graphs inevitably requires particular method to handle the challenges brought by the graph structure. Studies in this area mainly include the following branches: ***(i)*** *Proxies including graph structure*: although latent confounders on graphs are easily neglected by regular methods, fortunately, the graph structure itself can serve as proxies for the latent confounders in many cases [43, 44, 74]. ***(ii)*** *Circumvention-Based Methods*: Under certain circumstances, graph structure affects the treatment assignments and plays the role of an instrumental variable [73]. Therefore, IV-based causal effect estimation approaches can be applied. ***(iii)*** *Interference*: one major issue of treatment effect estimation on graphs is that there often exists interference between connected units (graph nodes), i.e., the treatment of one unit may causally influence the outcome of other units. This, however, violates the SUTVA assumption [96] in traditional causal inference. There have been numerous explorations [35, 52, 80, 81] in this problem, covering different types of graphs.

Counterfactual Analysis. On graphs, counterfactual reasoning targets on generating a different graph under certain circumstances different from the factual one. As the graph structure is involved, counterfactual analysis on graphs often involves additional considerations regarding the causal relations among nodes, as well as the discrete and unorganized structural space. Various investigations have been conducted for this problem, including different goals such as generalization [69, 122], explanation [41, 57, 76, 102, 125], and fairness [31, 45, 77] in many important applications.

Causal Discovery. The nature of graphs makes them closely associated with causal relations. Related causal discovery work in this area mainly includes ***(i)*** methods based on classical graphical models [37], which rely on causal graphical models and have been the mainstream of causal discovery; ***(ii)*** methods based on learnable graph adjacency matrices in neural networks [148, 157], which discover the causal relations inside data by learning an $N \times N$ adjacency matrix for a causal graph with N variables; ***(iii)*** methods based on graph neural networks (GNNs) [90, 135, 150], which explicitly leverage GNN techniques to facilitate causal discovery.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297.

[2] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. 2015. Estimating conditional average treatment effects. *JBES* 33, 4 (2015), 485–505.

[3] Jeffrey M Albert, Cuiyu Geng, and Suchitra Nelson. 2016. Causal mediation analysis with a latent mediator. *Biom. J.* (2016).

[4] Joshua Angrist and Guido Imbens. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 2 (1994), 467–475.

[5] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *JASA* (1996).

[6] Elizabeth Anscombe. 2018. Causality and determination. In *Agency and Responsiblity*.

[7] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv* (2019).

[8] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U.S.A* (2016).

[9] Michael Baiocchi, Jing Cheng, and Dylan S Small. 2014. Instrumental variable methods for causal inference. *Stat. Med.* (2014).

[10] Alexander Balke and Judea Pearl. 1994. Counterfactual probabilities: Computational methods, bounds and applications. In *UAI*. 46–54.

[11] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv* (2023).

[12] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* (1986).

[13] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. 2019. Deep generalized method of moments for instrumental variable analysis. In *NeurIPS*.

[14] Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. 2020. Tuning causal discovery algorithms. In *ICPGM*. 17–28.

[15] Leo Breiman. 1992. *Probability*. SIAM.

[16] Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. 2019. Causal discovery with cascade nonlinear additive noise models. *arXiv* (2019).

[17] Lu Cheng, Ruocheng Guo, and Huan Liu. 2022. Causal mediation analysis with hidden confounders. In *WWW*.

[18] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *AAAI*, Vol. 33. 7801–7808.

[19] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3, Nov (2002), 507–554.

[20] Zhixuan Chu, Mengxuan Hu, Qing Cui, Longfei Li, and Sheng Li. 2024. Task-driven causal feature distillation: Towards trustworthy risk prediction. In *AAAI*, Vol. 38. 11642–11650.

[21] Zhixuan Chu and Sheng Li. 2023. Causal Effect Estimation: Recent Progress, Challenges, and Opportunities. *Machine Learning for Causal Inference* (2023).

[22] Tom Claassen and Tom Heskes. 2012. A Bayesian approach to constraint based causal inference. *arXiv* (2012).

[23] Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. 2023. Large Language Models for Constrained-Based Causal Discovery. In *AAAI Workshop*.

[24] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. 2024. Causal inference methods for combining randomized trials and observational studies: A review. *Stat. Sci.* (2024).

[25] Diego Colombo, Marloes H Maathuis, et al. 2014. Order-independent constraint-based causal structure learning. *JMLR* 15, 1 (2014), 3741–3782.

[26] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* (2012).

[27] Ruifei Cui, Perry Groot, and Tom Heskes. 2016. Copula PC algorithm for causal discovery from mixed data. In *ECML PKDD*. 377–392.

[28] Haixing Dai, Mengxuan Hu, Qing Li, Lu Zhang, Lin Zhao, Dajiang Zhu, Diez, et al. 2023. Graph-based counterfactual causal inference modeling for neuroimaging analysis. In *MICCAI*. 205–213.

[29] Andriy Derkach, Ruth M Pfeiffer, Ting-Huei Chen, and Joshua N Sampson. 2019. High dimensional mediation analysis with latent variables. *Biometrics* 75, 3 (2019), 745–756.

[30] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. 2020. Minimax estimation of conditional moment models. In *NeurIPS*.

[31] Yushun Dong, Jing Ma, Chen Chen, and Jundong Li. 2022. Fairness in Graph Mining: A Survey. *arXiv* (2022).

[32] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *ICLR*. 8469–8488.

[33] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[34] Helmut Farbmacher, Martin Huber, Lukáš Lafférs, Henrika Langen, and Martin Spindler. 2022. Causal mediation analysis with double machine learning. *The Econometrics Journal* (2022).

[35] Zahra Fatemi and Elena Zheleva. 2020. Minimizing interference and selection bias in network experiment design. In *AAAI*.

[36] Isabel R Fulcher, Ilya Shpitser, Stella Marealle, and Eric J Tchetgen Tchetgen. 2020. Robust inference on population indirect causal effects: The generalized front door criterion. *J. R. Stat.* (2020).

[37] Clark Glymour and Kun Zhang. 2019. Review of causal discovery methods based on graphical models. *Front. Genet.* (2019).

[38] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

[39] Sander Greenland. 1990. Randomization, statistics, and causal inference. *Epidemiology* (1990).

[40] Sander Greenland. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 29, 4 (2000), 722–729.

[41] Zihan Guan, Mengnan Du, and Ninghao Liu. 2023. XGBD: Explanation-Guided Graph Backdoor Detection. arXiv:2308.04406

[42] Zihan Guan, Mengxuan Hu, Sheng Li, and Anil Vullikanti. 2024. Ufid: A unified framework for input-level backdoor detection on diffusion models. *arXiv* (2024).

[43] Ruocheng Guo, Jundong Li, Yichuan Li, K Selçuk Candan, Adrienne Raglin, and Huan Liu. 2020. IGNITE: A minimax game toward learning individual treatment effects from networked observational data. In *IJCAI*.

[44] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *WSDM*.

[45] Zhimeng Guo, Jialiang Li, Teng Xiao, Yao Ma, and Suhang Wang. 2023. Towards fair graph neural networks via graph counterfactual. In *CIKM*. 669–678.

[46] Ben B Hansen. 2008. The prognostic analogue of the propensity score. *Biometrika* 95, 2 (2008), 481–488.

[47] Lars Peter Hansen. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* (1982).

[48] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *ICML*.

[49] Negar Hassanpour and Russell Greiner. 2020. Learning disentangled representations for counterfactual regression. In *ICLR*.

[50] Mengxuan Hu, Zhixuan Chu, and Sheng Li. 2023. DBRNet: Advancing Individual-Level Continuous Treatment Estimation through Disentangled and Balanced Representation. (2023).

[51] Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. 2021. A generative adversarial framework for bounding confounded causal effects. In *AAAI*. 12104–12112.

[52] Qiang Huang, Jing Ma, Jundong Li, Ruocheng Guo, Huiyan Sun, and Yi Chang. 2023. Modeling interference for individual treatment effect estimation from networked observational data. *TKDD* 18, 3 (2023), 1–21.

[53] Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychol. Methods* (2010).

[54] Kosuke Imai, Luke Keele, and Teppei Yamamoto. 2010. Identification, inference and sensitivity analysis for causal mediation effects. (2010).

[55] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*.

[56] Fattaneh Jabbari, Joseph Ramsey, Peter Spirtes, and Gregory Cooper. 2017. Discovery of causal models that contain latent variables through Bayesian scoring of independence constraints. In *ECML PKDD*.

[57] Utkarshani Jaimini and Amit Sheth. 2022. CausalKG: Causal Knowledge Graph Explainability using interventional and counterfactual reasoning. *IEEE Internet Computing* 26, 1 (2022), 43–50.

[58] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In *NeurIPS*.

[59] Nathan Kallus, Xiaojie Mao, and Madeleine Udell. 2018. Causal inference with noisy and missing covariates via matrix factorization. In *NeurIPS*.

[60] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. 2018. Removing hidden confounding by experimental grounding. In *NeurIPS*.

[61] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*.

[62] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv* (2023).

[63] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *NeurIPS*.

[64] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational Bayes. In *ICLR*.

[65] Manabu Kuroki. 2007. Graphical identifiability criteria for causal effects in studies with an unobserved treatment/response variable. *Biometrika* 94, 1 (2007), 37–47.

[66] Manabu Kuroki and Judea Pearl. 2014. Measurement bias and effect restoration in causal inference. *Biometrika* (2014).

[67] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *NeurIPS*.

[68] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. 2016. A fast PC algorithm for high dimensional causal discovery with multi-core PCs. *IEEE/ACM TCBB* 16, 5 (2016), 1483–1495.

[69] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2022. Out-of-distribution generalization on graphs: A survey. *arXiv* (2022).

[70] Mingzhou Liu, Xinwei Sun, Lingjing Hu, and Yizhou Wang. 2024. Causal discovery from subsampled time series with proxy variables. In *NeurIPS*.

[71] Mingzhou Liu, Xinwei Sun, Yu Qiao, and Yizhou Wang. 2023. Causal discovery with unobserved variables: A proxy variable approach. *arXiv* (2023).

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

[73] 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 *KDD*. 4584–4594.

[74] Jing Ma, Ruocheng Guo, Chen Chen, Aidong Zhang, and Jundong Li. 2021. Deconfounding with networked observational data in a dynamic environment. In *WWW*.

[75] Jing Ma, Ruocheng Guo, and Jundong Li. 2023. Causal Inference on Graphs. In *Machine Learning for Causal Inference*.

[76] Jing Ma, Ruocheng Guo, Saumitra Mishra, Aidong Zhang, and Jundong Li. 2022. Clear: Generative counterfactual explanations on graphs. In *NeurIPS*.

[77] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 2022. Learning fair node representations with graph counterfactual fairness. In *WSDM*.

[78] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. 2021. Multi-cause effect estimation with disentangled confounder representation. In *IJCAI*.

[79] Jing Ma and Jundong Li. 2022. Learning causality with graphs. *AI Magazine* 43, 4 (2022), 365–375.

[80] Jing Ma, Mengting Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. 2022. Learning causal effects on hypergraphs. In *KDD*. 1202–1212.

[81] Yunpu Ma and Volker Tresp. 2021. Causal inference under networked interference and intervention policy enhancement. In *AISTATS*. 3700–3708.

[82] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. 2007. Mediation analysis. *Annu. Rev. Psychol.* (2007).

[83] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* (2018).

[84] Jacob M Montgomery, Brendan Nyhan, and Michelle Torres. 2018. How conditioning on posttreatment variables can ruin your experiment and what to do about it. *AJPS* (2018).

[85] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. 2020. Dual instrumental variable regression. In *NeurIPS*.

[86] Bengt Muthén. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 1 (1984), 115–132.

[87] Bengt Muthén and Tihomir Asparouhov. 2015. Causal effects in mediation modeling: An introduction with applications to latent variables. *Struct. Equ. Modeling* (2015).

[88] Jessica A Myers, Jeremy A Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. 2011. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology* 174, 11 (2011), 1213–1222.

[89] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *AAAI*.

[90] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. 2019. A graph autoencoder approach to causal structure learning. *arXiv* (2019).

[91] Austin Nichols. 2007. Causal inference with observational data. *The Stata Journal* (2007).

[92] Xinkun Nie and Stefan Wager. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* (2021).

[93] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 2 (2022), e1449.

[94] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. 2016. A hybrid causal search algorithm for latent variable models. In *ICPGM*. 368–379.

[95] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[96] Judea Pearl. 2009. *Causality*.

[97] Judea Pearl. 2012. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. *Causality: Statistical Perspectives and Applications* (2012), 151–179.

[98] Judea Pearl. 2012. On measurement bias in causal inference. *arXiv* (2012).

[99] Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*.

[100] Judea Pearl and Dana Mackenzie. 2018. *The book of why: The new science of cause and effect*.

[101] Alexander Peysakhovich and Akos Lada. 2016. Combining observational and experimental data to find heterogeneous treatment effects. *arXiv* (2016).

[102] Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo, and Fosca Giannotti. 2023. A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. *Comput. Surveys* (2023).

[103] Kristopher J Preacher. 2015. Advances in mediation analysis: A survey and synthesis of new developments. *Annu. Rev. Psychol.* (2015).

[104] Zhaozhi Qian, Alicia Curth, and Mihaela van der Schaar. 2021. Estimating multi-cause treatment effects via single-cause perturbation. In *NeurIPS*. 23754–23767.

[105] Vineet K Raghu, Joseph D Ramsey, Alison Morris, Dimitrios V Manatakis, Peter Sprites, Panos K Chrysanthis, Clark Glymour, and Panayiotis V Benos. 2018. Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics* 6 (2018), 33–45.

[106] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 2017. A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics* 3 (2017), 121–129.

[107] Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. 2024. A Survey of Large Language Models for Graphs. *arXiv* (2024).

[108] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* (1983).

[109] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Am. Stat. Assoc* (2005).

[110] Shiv Kumar Saini, Sunny Dhamnani, Akil Arif Ibrahim, and Prithviraj Chavan. 2019. Multiple treatment effect estimation using deep generative model with task embedding. In *WWW*. 1601–1611.

[111] AJ Sasco, MB Secretan, and K Straif. 2004. Tobacco smoking and cancer: A brief review of recent epidemiological evidence. *Lung Cancer* (2004).

[112] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.

[113] Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *JMLR* 9, 3 (2008).

[114] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*.

[115] Claudia Shi, Victor Veitch, and David M Blei. 2021. Invariant representation learning for treatment effect estimation. In *UAI*.

[116] Rahul Singh, Maneesh Sahani, and Arthur Gretton. 2019. Kernel instrumental variable regression. In *NeurIPS*.

[117] Peter Spirtes. 2001. An anytime algorithm for causal inference. In *AISTATS*. 278–285.

[118] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search.*

[119] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: Concepts and recent methodological advances. In *Appl. Inform.*

[120] Peter L Spirtes, Christopher Meek, and Thomas S Richardson. 2013. Causal inference in the presence of latent variables and selection bias. *arXiv* (2013).

[121] Eric V Strobl, Shyam Visweswaran, and Peter L Spirtes. 2018. Fast causal inference with non-random missingness by test-wise deletion. *JDSA* (2018).

[122] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. 2022. Causal attention for interpretable and generalizable graph classification. In *KDD*.

[123] Rongqian Sun, Xiaoxiao Zhou, and Xinyuan Song. 2021. Bayesian causal mediation analysis with latent mediators and survival outcome. *Structural Equation Modeling: A Multidisciplinary Journal* 28, 5 (2021), 778–790.

[124] Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. 2016. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *JBES* (2016).

[125] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *WWW*. 1018–1027.

[126] Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. 2020. An introduction to proximal causal learning. *arXiv* (2020).

[127] Jin Tian and Judea Pearl. 2002. A general identification condition for causal effects. In *IAAI*.

[128] Fani Tsapeli and Mirco Musolesi. 2015. Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach. *EPJ Data Science* 4, 1 (2015), 24.

[129] Mark van der Laan, Sky Qiu, and Lars van der Laan. 2024. Adaptive-TMLE for the Average Treatment Effect based on Randomized Controlled Trial Augmented with Real-World Data. *arXiv* (2024).

[130] Mark J Van der Laan, Sherri Rose, et al. 2011. *Targeted learning: Causal inference for observational and experimental data.* Vol. 4.

[131] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[132] Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *UAI*.

[133] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *JASA* (2018).

[134] Guangya Wan, Yuqi Wu, Mengxuan Hu, Zhixuan Chu, and Sheng Li. 2024. Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions. *arXiv* (2024).

[135] Dongjie Wang, Zhengzhang Chen, Jingchao Ni, Liang Tong, Zheng Wang, Yanjie Fu, and Haifeng Chen. 2023. Hierarchical graph neural networks for causal discovery and root cause localization. *arXiv* (2023).

[136] Haotian Wang, Kun Kuang, Haoang Chi, Longqi Yang, Mingyang Geng, Wanrong Huang, and Wenjing Yang. 2023. Treatment effect estimation with adjustment feature selection. In *SIGKDD*.

[137] Lun Wang, Qi Pang, and Dawn Song. 2020. Towards practical differentially private causal graph discovery. In *NeurIPS*. 5516–5526.

[138] Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, et al. 2023. Knowledge editing for large language models: A survey. *arXiv* (2023).

[139] Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *JASA* (2019).

[140] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020. Causal inference for recommender systems. In *RecSys*.

[141] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. 2017. Permutation-based causal inference algorithms with interventions. In *NeurIPS*.

[142] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

[143] Lili Wu and Shu Yang. 2022. Integrative *R*-learner of heterogeneous treatment effects combining experimental and observational studies. In *CLeaR*.

[144] Pengzhou Wu and Kenji Fukumizu. 2022. β-Intact-VAE: Identifying and estimating causal effects under limited overlap. In *ICLR*.

[145] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv* (2024).

[146] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-fairness: A unified framework for measuring causality-based fairness. In *NeurIPS*.

[147] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. 2024. Causal inference with conditional front-door adjustment and identifiable variational autoencoder. In *ICLR*.

[148] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2021. CausalVAE: Disentangled representation learning via neural structural causal models. In *CVPR*.

[149] Shu Yang, Donglin Zeng, and Xiaofei Wang. 2020. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv* (2020).

[150] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *ICML*. 7154–7163.

[151] Junkun Yuan, Anpeng Wu, Kun Kuang, Bo Li, Runze Wu, Fei Wu, and Lanfen Lin. 2022. Auto IV: Counterfactual prediction via automatic instrumental variable decomposition. *TKDD* (2022).

[152] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning* 151 (2022), 101–129.

[153] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv* (2023).

[154] Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Xin Wen, and Bingchen Zhao. 2023. What if the TV was off? Examining counterfactual reasoning abilities of multi-modal language models. In *CVPR*. 4629–4633.

[155] Weijia Zhang, Lin Liu, and Jiuyong Li. 2021. Treatment effect estimation with disentangled latent factors. In *AAAI*.

[156] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv* (2023).

[157] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with no tears: Continuous optimization for structure learning. In *NeurIPS*.

[158] Yaochen Zhu and Zhenzhong Chen. 2022. Mutually-regularized dual collaborative variational auto-encoder for recommendation systems. In *WWW*. 2379–2387.

[159] Yaochen Zhu, Jing Ma, and Jundong Li. 2023. Causal Inference and Recommendations. In *Machine Learning for Causal Inference*. Springer, 207–245.

[160] Yaochen Zhu, Jing Ma, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2023. Path-Specific Counterfactual Fairness for Recommender Systems. In *SIGKDD*. 3638–3649.

[161] Yaochen Zhu, Jing Ma, Liang Wu, Guo Qi, Liangjie Hong, and Jundong Li. 2024. Treatment effect estimation with mixed latent post-treatment variables. (2024).

[162] Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2024. Collaborative large language model for recommender systems. In *WWW*.

[163] Yaochen Zhu, Jing Yi, Jiayi Xie, and Zhenzhong Chen. 2022. Deep causal reasoning for recommendations. *ACM TIST* (2022).

[164] Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. 2020. Counterfactual prediction for bundle treatment. In *NeurIPS*.

[165] Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. 2022. Counterfactual fairness with partially known causal graph. In *NeurIPS*. 1238–1252.