Received: 22 August 2022 | Accepted: 21 September 2022

DOI: 10.1002/aaai.12070

ARTICLE

Learning causality with graphs

Jing Ma | **Jundong Li**

University of Virginia, Charlottesville, Virginia, USA

Correspondence
Jing Ma, University of Virginia, Charlottesville, VA 22904, USA.
Email: jm3mr@virginia.edu

Funding information
JP Morgan Chase Faculty Research Award; 3Cavaliers; National Science Foundation, Grant/Award Numbers: CNS-2154962, IIS-2006844, IIS-2144209

Abstract

Recent years have witnessed a rocketing growth of machine learning methods on graph data, especially those powered by effective neural networks. Despite their success in different real-world scenarios, the majority of these methods on graphs only focus on predictive or descriptive tasks, but lack consideration of causality. Causal inference can reveal the causality inside data, promote human understanding of the learning process and model prediction, and serve as a significant component of artificial intelligence (AI). An important problem in causal inference is causal effect estimation, which aims to estimate the causal effects of a certain treatment (e.g., prescription of medicine) on an outcome (e.g., cure of disease) at an individual level (e.g., each patient) or a population level (e.g., a group of patients). In this paper, we introduce the background of causal effect estimation from observational data, envision the challenges of causal effect estimation with graphs, and then summarize representative approaches of causal effect estimation with graphs in recent years. Furthermore, we provide some insights for future research directions in related area. Link to video abstract: https://youtu.be/BpDPOOqw-ns

INTRODUCTION OF CAUSAL EFFECT ESTIMATION WITH GRAPHS

Graphs have been extensively used for modeling a plethora of real-world systems, including social media platforms (Bazarova and Choi 2014), collaboration networks (Newman 2001), biological networks (Junker and Schreiber 2011), and critical infrastructure systems (Ouyang 2014), to name a few. Currently, the mainstream learning tasks on graphs are either predictive (e.g., node classification) or descriptive (e.g., measuring centrality) in nature. Most of studies (Wu et al. 2020; Zhou et al. 2020) on graphs address these tasks only from a statistical perspective, for example, utilizing the statistical correlations between node features, graph structure, and labels for node classification. But beyond the statistical level, we may also want to understand the causality of the learning process, which is often considered a significant component of human-level intelligence and can serve as the foundation of artificial intelligence (AI). Causal inference (Pearl 2009; Imbens and Rubin 2015) is the process of investigating causality. In causal inference, an important problem is to estimate the causal effects of a certain treatment (e.g., prescription of medicine) on an important outcome (e.g., cure of disease) at the individual/instance level (e.g., each patient) or a population level (e.g., a group of patients). The problem is often known as treatment effect estimation or causal effect estimation, and has a wide range of applications such as economics, public health, education, and environmental science.

The gold standard of causal effect estimation is conducting randomized controlled trials (RCTs), which randomly allocate treatment assignment to participants and compare the difference between the outcomes of participated individuals with different treatment assignments. However, RCTs are often expensive, impractical, or even unethical to conduct in many real-world scenarios (Goldstein et al. 2018). Thus, a large body of research in the past couple of decades has been dedicated to causal effect estimation from observational data. However, most of these research works assume the observational data are independent and identically distributed (i.i.d.), while this traditional setting does not fit well in many scenarios, for example, there may exist additional relational information such as social networks among individuals. With the rocketing availability of graph data across a myriad of influential areas, we are interested in developing novel frameworks to enable causal effect estimation with graphs to facilitate many downstream applications, such as policy evaluation and decision making. For example, given a social network of users, service providers need to decide whether the advertisement of a product (treatment) will help an individual user make a purchase (outcome) to provide better personalized recommendations; given a contact network among individuals for an infectious disease, government agencies, and healthcare providers need to quantify how different intervention strategies (e.g., self-quarantine, school closure) will impact the infections of each individual or a certain population group.

In this paper, we first introduce some most representative methods of causal effect estimation from observational data in Section “Existing works of causal effect estimation from observational i.i.d. data,” including traditional methods and state-of-the-art representation learning-based approaches. These methods, albeit perform effective in i.i.d. data, cannot be directly applied to graph data due to a couple of fundamental challenges as analyzed in Section “Challenges of causal effect estimation with graphs.” We then summarize current works of causal effect estimation with graphs in Section “Methods of causal effect estimation with graphs.” Specifically, these works can be mainly divided into three categories: (1) **Causal effect estimation with hidden confounders on graphs**. These works utilize the graph structure among individuals to mitigate the confounding biases in causal effect estimation caused by hidden confounders (confounders are variables which influence both treatment and outcome). (2) **Causal effect estimation with interference**. These works estimate causal effect under the assumption that there exists interference among different individuals (i.e., the outcome of each individual may be causally influenced by the treatment assignment of other individuals). (3) **Causal effect estimation with graph-structured treatments**. These works differ from the traditional setting in which the treatment is a scalar (even a binary value in most cases), and model the treatment with graph structure. Beyond these works of causal effect estimation with graphs, we further introduce more works in causality learning with graphs in a bigger picture, such as policy evaluation, counterfactual fairness on graphs, and causality in graph neural networks (GNNs) in Section “Beyond causal effect estimation.” Furthermore, we discuss several potential directions in this area in Section “Discussions and future work” and provide insights for future research exploration.

EXISTING WORKS OF CAUSAL EFFECT ESTIMATION FROM OBSERVATIONAL I.I.D. DATA

In this section, we introduce some previous works of causal effect estimation from observational i.i.d. data. First, as various traditional studies (Austin 2011; Funk et al. 2011; Hill 2011; Wager and Athey 2018) have been dedicated to this task, we briefly review several well-known traditional methods. Second, with the rapid development of machine learning and neural network in recent years, the representation learning-based methods (Louizos et al. 2017; Shalit, Johansson, and Sontag 2017; Yao et al. 2018) for causal effect estimation have attracted significant attention. Thus, we also summarize the progress of recent representation learning-based methods (Shalit, Johansson, and Sontag 2017; Yao et al. 2018; Shi, Blei, and Veitch 2019; Louizos et al. 2017) for causal effect estimation from observational i.i.d. data.

Traditional methods

One of the key challenges of causal effect estimation from observational data is the existence of *confounders*. Confounders are variables that influence both treatment assignment and outcome. For example, when estimating the causal effect of advertising (treatment) on the purchase pattern (outcome) for each user (individual), the user’s preferences can be considered as confounders, which influence both the advertisement the user receives as well as his/her purchase patterns. Without an effective controlling of the influence of confounders, the causal effect estimation methods would utilize the statistical dependency rather than the causal relation between the treatment and outcome, and thus suffer from confounding biases.

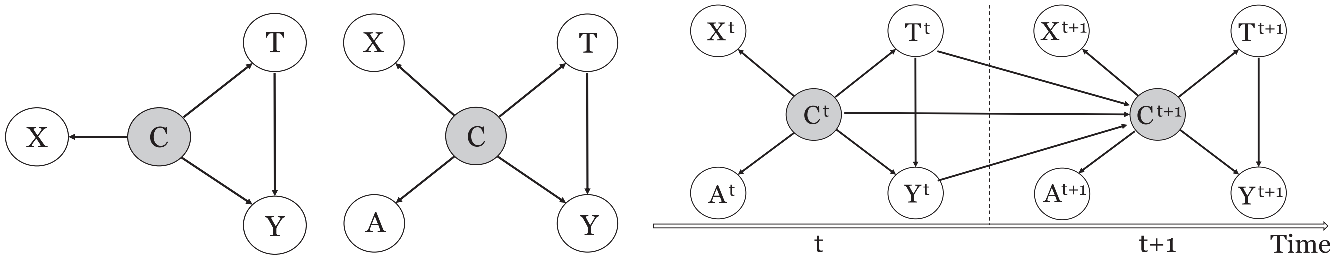
To mitigate the confounding bias, most of the existing works of causal effect estimation are based on the strong ignorability assumption (Hill 2011; Shalit, Johansson, and Sontag 2017; Wager and Athey 2018), which assumes that all the confounders are in the observed features, that is, *hidden confounders* (confounders, which are not directly observed) do not exist. Among these methods, regression methods build predictors for the *potential outcomes* (the outcomes which would be realized if a certain treatment had been assigned for an individual, including the *factual/observed outcome* corresponding to the observed treatment assignment, as well as the *counterfactual outcome* corresponding to a treatment assignment different from the observed one). That is, regression methods model the distribution $P(\mathrm{Y}|\mathrm{X}, \mathrm{T})$, where X, T, Y denote features, treatment assignment, and outcome for each individual. In this way, the causal effect estimation task is transformed into a supervised learning problem with partially labeled data (observed outcomes), and the causal effect can be estimated based on the predicted potential outcomes. Another classical methods are based on propensity scores (Austin 2011). Propensity score methods use a function $f(\mathrm{X})$ to model the propensity score $P(\mathrm{T}|\mathrm{X})$, and assume the treatment assignments are sampled from the true propensity scores. Matching or covariate balancing can be conducted based on these propensity scores for unbiased causal effect estimation. The propensity score-based methods can be mainly grouped into four categories (Austin 2011): propensity score matching (PSM), propensity score stratification, inverse probability of treatment weighting (IPTW), and adjustment based on propensity score. Doubly robust estimation (DRE) (Funk et al. 2011) uses an estimator, which combines both regression-based model and propensity score model to achieve better robustness. In DRE, the estimation of averaged treatment effect can be unbiased as long as at least one of the two models is correctly specified. Furthermore, many other causal effect estimation methods have also been developed to handle the scenarios that hidden confounders exist. These methods mainly include instrumental variable methods (Angrist, Imbens, and Rubin 1996; Angrist and Imbens 1995), front-door criterion (Pearl 1995), and regression discontinuity design (Angrist and Lavy 1999).

Representation learning-based methods

Recently, deep learning are progressing at an astounding rate, powered by effective neural network models. The progress in deep learning also stimulates the studies on representation learning-based causal effect estimation methods. Among them, Shalit, Johansson, and Sontag (2017) propose treatment-agnostic representation network (TARNET), which learns the representations of confounders with input as instance features, and predicts the potential outcomes based on the learned representations. To further mitigate the biases in estimation, another framework counterfactual regression (CFR) is also proposed (Shalit, Johansson, and Sontag 2017), which additionally uses representation balancing techniques to minimize the distribution distance between confounders' representations of the treatment group (the group of individuals which get treated) and the control group (the group of individuals which get controlled, i.e., not treated). Typical representation balancing techniques can be based on Wasserstein-1 distance (Villani 2009) (CFR-Wass) or maximum mean discrepancy (CFR-MMD) (Gretton et al. 2012). A local similarity preserved individual treatment effect (SITE) estimation method based on representation learning (Yao et al. 2018) follows a similar design, but preserves the local similarity information and balances data distributions by focusing on several hard samples in each mini-batch. Most of these methods are also based on the strong ignorability assumption (Hill 2011; Shalit, Johansson, and Sontag 2017; Wager and Athey 2018). Different from them, causal effect variational autoencoder (CEVAE) (Louizos et al. 2017) assumes that it can infer the hidden confounders based on a deep latent-variable model. CEVAE is based on the causal graph shown in Figure 1A, where C denotes the hidden confounders, and CEVAE assumes that the confounders can be inferred from observed features X. Most of these above representation learning-based methods have achieved outstanding performance in the problem of causal effect estimation from observational data.

CHALLENGES OF CAUSAL EFFECT ESTIMATION WITH GRAPHS

Despite the success of the above methods in causal effect estimation from observational data, most of them are still limited in i.i.d. data. Recent studies (Guo, Li, and Liu 2020b) have revealed the great importance and benefit to conduct causal effect estimation with graphs. However, performing causal effect estimation with graphs remains a daunting task and is a rather underexplored area due to the following challenges: (1) **Different modalities in graph data**. Most of the existing causal inference methods focus on observational i.i.d. data. However, the graph data contain different modalities including instance features as well as graph structure. Jointly utilizing these different modalities for causal effect estimation is the first challenge to address. (2) **Existence of hidden confounders**. Existing efforts often ignore the influence of hidden confounders, and are mostly based on the strong ignorability assumption (Hill 2011; Shalit, Johansson, and Sontag 2017; Wager and Athey 2018), which assumes that all the confounders are in the observed features. However, such assumption is difficult to be satisfied in real-world observational data. As in the aforementioned example,



(A) Causal graph on i.i.d. data. (B) Causal graph on graph data. (C) Causal graph on dynamic graph data.

FIGURE 1 Causal graphs commonly used under different assumptions of observational data, including i.i.d. data (Louizos et al. 2017), graph data (Guo, Li, and Liu 2020b), and dynamic graph data (Ma et al. 2021). Each circle denotes a variable, and each arrow stands for a causal relation. The gray circles are unobserved variables, while other white ones are observed. X, C, T, Y, A here denotes instance features, confounders, treatment assignment, outcome, and graph structure, respectively. The superscript $(\cdot)^t$ denotes a variable at time stamp t.

confounders such as user's preferences are often not explicitly measured. Without controlling the influence of hidden confounders, existing methods based on such assumption may result in biased estimation of causal effect. (3) **Complicated forms of graphs**. Many real-world graphs are complicated, for example, dynamic graphs. Typical examples include interactions among individuals in an epidemic and connections among users in a social network during different time periods. Performing causal effect estimation on these graphs is often difficult as it requires us to control for the hidden confounders in a complicated evolving environment. (4) **Network interference**. Most of the existing works of causal effect estimation are based on the stable unit treatment value (SUTVA) assumption (Fisher 1936; Splawa-Neyman, Dabrowska, and Speed 1990). SUTVA assumption requires that the interference (i.e., spillover effect) among individuals does not exist, that is, the outcome of any individual is not influenced by the treatment assignment of other individuals. However, interference among individuals is ubiquitous in the real world, especially in networked data (Rakesh et al. 2018; Ma and Tresp 2021). (5) **Graph-structured treatments**. In traditional settings of causal effect estimation, treatment assignment of each individual is often a scalar (even a binary value in most cases), but in many real-world scenarios, treatments can be naturally modeled as graph structures, for example, when the treatments are molecular structures of chemotherapy drugs. In these cases, traditional methods cannot be directly adopted. Handling graph-structured treatment is a new challenge in this area.

METHODS OF CAUSAL EFFECT ESTIMATION WITH GRAPHS

Despite the aforementioned challenges, opportunities are also unequivocally present with the graph data—although the hidden confounders are notoriously hard to measure, we can capture their patterns and control their influence by incorporating the underlying graph structure. For example, the purchasing preferences (hidden confounders) of an individual can affect the recommended items to him/her and his/her purchasing behaviors. However, although the purchasing preferences of an individual are difficult to be directly measured from observational data, they are often encoded implicitly in the social network, such as which community he/she belongs to. We introduce several works, which bridge the knowledge gap by developing novel causal inference frameworks for causal effect estimation with graphs. More specifically, we mainly focus on the following categories of works: (1) Causal effect estimation with hidden confounders with graphs. These works utilize the graph structure among individuals to better infer the hidden confounders and achieve unbiased causal estimation; (2) causal effect estimation under interference. Most of these works relax the SUTVA assumption (Fisher 1936; Splawa-Neyman, Dabrowska, and Speed 1990) and assume that the outcome of each individual can be influenced by the treatment assignments of other individuals (e.g., neighbor nodes in the graph). (3) Causal effect estimation with graph-structured treatments. Unlike traditional works, which take treatment as a scalar, in these works, treatments are naturally modeled as graph structures, such as molecular structure of chemotherapy drugs.

Causal effect estimation with hidden confounders on graphs

Network deconfounder. To mitigate the confounding biases for individual-level treatment effect (ITE) estimation, Guo, Li, and Liu (2020b) use a weaker version of strong ignorability assumption, and assume that the hidden confounders can be captured from the proxy variables for them, that is, the variables which have dependencies with the hidden confounders. In the aforementioned example, the graph structure of social network among individuals in observational data can often reflect the hidden confounders. As the causal graph shown in Figure 1B, both instance features X and graph structure A can serve as proxy variables to infer the hidden confounders C. Based on this assumption, Guo, Li, and Liu (2020b) propose a method—network deconfounder (Guo, Li, and Liu 2020b), which utilizes the graph structure as well as the instance features to infer hidden confounders for ITE estimation on graphs. More specifically, network deconfounder maps the features and the network structure simultaneously into a latent space with graph convolutional networks (GCNs) (Kipf and Welling 2017) to learn the representations of hidden confounders. Based on the learned representations of confounders and treatment assignment, network deconfounder makes predictions for potential outcomes of each individual. Representation learning-based methods have been leveraged for ITE estimation in previous studies (Johansson, Shalit, and Sontag 2016; Shalit, Johansson, and Sontag 2017; Louizos et al. 2017). Different from them, network deconfounder is the first work, which utilizes auxiliary network information to learn the confounder representation for ITE estimation.

Minimax game between representation balancing and treatment prediction. To further enhance the performance of ITE estimation on graph data, Guo et al. (2020) consider two desiderata for ITE estimation: (1) On the group level, existing works (Shalit, Johansson, and Sontag 2017; Yao et al. 2018) have proved that minimizing the discrepancy between the representation distributions of treatment group and control group can help mitigate the biases in ITE estimation. (2) On the individual level, the learned confounder representations are desired to capture patterns of hidden confounders, which can predict treatment assignments. As these two desiderata often contradict each other, this work proposes a minimax game-based network ITE estimator (IGNITE) (Guo et al. 2020) to achieve these two desiderata.

Graph infomax adversarial learning for treatment effect estimation on graphs. Chu, Rathbun, and Li (2021) realize that the confounding bias in causal inference problem can cause the data imbalance not only between distributions of features in treatment group and control group, but also between their network structures. The imbalance in network structure between treatment and control groups can aggravate the imbalance of the representations learned by GNNs. To address this problem, (Chu, Rathbun, and Li 2021) proposes a graph infomax adversarial learning (GIAL) model for treatment effect estimation, which captures more information by recognizing the imbalance in network structure. Specifically, GIAL maximizes the structure mutual information (Velickovic et al. 2019) between the learned representation vector and the structure summary vector to help GNNs to learn representations of confounders from the imbalanced networked data, and uses adversarial learning for representation balancing.

Deconfounding in dynamic networks. Despite the empirical success of works, which utilize network structure to infer hidden confounders, these works overwhelmingly assume that the observational data and the relations among them are static. However, these data and their relations are naturally dynamic in many real-world scenarios (Li et al. 2017). For example, the purchasing preferences of users and their social connections are both evolving over time. Such data are referred as time-evolving networked observational data. The prevalence of such data in a wide spectrum of domains brings about new opportunities to unravel the patterns of hidden confounders towards unbiased ITE estimation. Considering this, Ma et al. (2021) propose a framework—dynamic network deconfounder (DNDC) for ITE estimation in dynamic networked observational data. This framework is based on the causal graph shown in Figure 1C, where the hidden confounders at current time stamp can be influenced by historical confounders, treatment assignments, and outcomes. Generally, DNDC learns representations of hidden confounders over time by mapping the current networked observational data and historical information into the representation space. More specifically, recurrent neural network (RNNs) (Medsker and Jain 2001) are utilized to capture the historical information, and GCNs (Kipf and Welling 2017) are used to encode the network structure in each time stamp. This work has been applied to assess the causal impact of different policies on the COVID-19 outbreak dynamics (e.g., the number of confirmed cases) at different time stamps (Ma et al. 2021).

Causal effect estimation with network interference

There have been lots of works to handle interference in causal inference (Aronow and Samii 2017; Basse and Feller 2018; Imai, Jiang, and Malani 2020; Kohavi et al. 2013; Tchetgen and VanderWeele 2012; Ugander et al. 2013; Yuan, Altenburger, and Kooti 2021; Ma and Tresp 2021; Bhattacharya, Malinsky, and Shpitser 2020). These works generally contain the following categories: (1) Studies on improving the assignment strategy. Most of these works (Ugander et al. 2013; Fatemi and Zheleva 2020) are based on cluster random assignment, and treat observations at a group level. Strong interference is assumed to exist within the groups while independence maintains across different groups. (2) Studies on causal effect estimation from observational data under interference, where the outcome of each individual is assumed to be causally influenced by the instance features and treatment assignment of both itself and other individuals. Here, we mainly introduce several recent studies on individual-level causal effect estimation under interference in pairs of individuals or graph data.

Identifying paired spillover effects. In causal inference, interference (i.e., spillover effect) occurs when the outcome of an individual is influenced by treatment of other individuals (Sobel 2006). Rakesh et al. (2018) propose a variational auto encoder (VAE) based framework—linked causal variational autoencoder (LCVA), to estimate the causal effect of a treatment on an outcome with the existence of spillover effects between pairs of individuals (i.e., paired spillover). More specifically, similar as traditional VAEs, the proposed framework LCVA reconstructs the inputs. But different from traditional VAEs, LCVA treats the latent embeddings as confounders. For each pair of individuals (i, j) and their treatment assignment and outcomes, the encoder of LCVA samples the confounders by conditioning on the observed features of i, the treatments of both i and j, as well as the outcome of i. The latent embeddings in LCVA are expected to capture the spillover effect.

Causal inference under networked interference. Recently, an ever-increasing number of efforts have been made to handle interference in networks (Ma and Tresp 2021; Aronow and Samii 2017; Basse and Feller 2018; Imai, Jiang, and Malani 2020). Among them, a line of works address this problem by relaxing SUTVA assumption and define the potential outcome as a function of the treatment assignment of each instance and a summary of its neighbors (e.g., k-hop neighbors on graph). Ma and Tresp (2021) study on causal effect estimation under the existence of network interference by proposing a novel GNN (Wu et al. 2020) based framework. GNNs are used as effective tools to capture the dependencies between nodes and links in the given graph. In this methods, GNNs are used to model the interference for each node by aggregating the information of this node's neighbors in the graph. Apart from such pairwise interference, a recent work (Ma et al. 2022) considers high-order interference in hypergraphs. Different from pairwise interference, high-order interference can influence each node through complicated interactions among multiple nodes on the same hyperedge. Correspondingly, a hypergraph neural network-based method HyperSCI (Ma et al. 2022) is proposed for causal effect estimation under interference in hypergraphs.

Causal effect estimation with graph-structured treatments

Different from traditional scalar treatments, in many real-world scenarios, treatments are naturally modeled as graphs. For example, when the treatment is the nutritional content of meals or molecular structure of chemotherapy drugs. In these scenarios, the traditional methods cannot be applicable for graph-structured treatments. Here, we introduce several recent works in this direction.

Individual treatment effect estimation with graph-structured treatments. GraphITE (Harada and Kashima 2020) is a method, which addresses the problem of causal effect estimation with graph-structured treatments. GraphITE learns representations of graph-structured treatments with GNNs, and mitigates the estimation biases with a Hilbert–Schmidt independence criterion (HSIC) (Gretton et al. 2007) regularization, which increases the independence of the representations of the target individuals and the treatments. One appealing advantage of GraphITE is the learned representations of treatments can enable zero-shot learning of unseen treatments.

Graph intervention networks (GIN). Kaddour et al. (2021) propose a framework GIN for causal effect estimation with graph-structured treatments. This work generalizes a well-known Robinson decomposition (Robinson 1988) of causal effect estimation to graph-structured treatments, and proposes a plug-in estimator, which decomposes causal effect estimation into separate and simpler optimization problems. This estimator can support any supervised learning methods. The evaluation on small-world and molecular graphs show its effectiveness and robustness to varying selection biases.

BEYOND CAUSAL EFFECT ESTIMATION

Aside from the above recent studies of causal effect estimation with graphs, various other studies (Guo, Li, and Liu 2020a; Kusner et al. 2017; Zečević et al. 2021) on causality learning with graphs also have attracted lots of attention. Generally, although studies on causality learning with graphs are challenging due to various barriers, such as the complicated relational information and interference among individuals, many of these works have still made remarkable contributions in different causality-related research problems, and also revealed the great potential of these works in many high-impact domains. Here, we summarize several recent works of causality learning with graphs to provide introduction of a bigger picture in this area.

Policy evaluation

One natural application of causality studies is to apply them into policy evaluation with observational data to save the effort of traditional A/B testing. Here, a policy refers to the intervention rule for treatment assignment. Policy evaluation from a causal perspective has been widely studied and applied in recommender systems (Schnabel et al. 2016), economics (Heckman 2000), and epidemiology (Chernozhukov, Kasahara, and Schrimpf 2021), but these works mostly focus on i.i.d. data. Recently, some works (Guo, Li, and Liu 2020a; Ma and Tresp 2021) have investigated policy evaluation on graph data from a causal perspective.

Counterfactual evaluation on graph data. Guo, Li, and Liu (2020a) study on the problem of counterfactual evaluation on observational graph data. Counterfactual evaluation aims to estimate the utility of a treatment assignment function, that is, the average outcome over a certain population under the treatments assigned by this function, without performing A/B testing to save the cost of randomized experiments. This work proposes a framework counterfactual network evaluator (CONE), which addresses the problem by exploiting the network structure and the observed instance features to mitigate hidden confounding effects.

Policy enhancement under network interference. Considering the existence of network interference, Ma and Tresp (2021) develop an algorithm to learn intervention rule for assignments (i.e., policy) to maximize the utility on the entire graph. A new utility function is defined on interconnected individuals. Based on it, the policy improves its decision rules through the utility function with budget constraints. Policy regret bounds under network interference and treatment capacity constraint are provided.

Fairness

Another interesting application of causality learning is to incorporate causality to handle the fairness issue in machine learning methods. Generally, a fair machine learning-based predictor aims to mitigate the discrimination of model prediction against certain demographic subpopulations regarding sensitive attributes such as race, gender, and age. Recently, fairness on graph mining is also an emerging field (Dong, Ma, Chen, and Li 2022). Among existing notions of fairness, counterfactual fairness (Kusner et al. 2017) measures the fairness of predictors from a causal perspective by comparing the predictions of each individual from the original data and the counterfactuals in which the sensitive attributes of this individual had been modified to a different value. However, most of these works only focus on i.i.d. data, while few of them have been applied on graphs.

Counterfactual fairness on graph data. Agarwal, Lakkaraju, and Zitnik (2021) propose to extend the notion of counterfactual fairness (Kusner et al. 2017) on graphs by developing a graph representation learning framework NIFTY, which targets on unifying Fairness and stability. NIFTY (Agarwal, Lakkaraju, and Zitnik 2021) is a GNN-based framework, which learns node representations that are both counterfactually fair and stable. This work learns counterfactual fair node representations by maximizing the agreement between the node representations learned from original graph and its counterfactual with Siamese networks (Bromley et al. 1993). The counterfactuals are generated by flipping the sensitive attribute values of all nodes in the graph. Theoretical analysis shows that NIFTY can promote counterfactual fairness and stability in the learned representations. Ma et al. (2022) propose a more comprehensive notion of graph counterfactual fairness, which further considers the following two types of biases: (1) biases induced by neighbors' sensitive attributes; and (2) biases induced by the causal effect of sensitive attributes on the graph structure and other features. A framework GEAR (Ma et al. 2022) is developed to learn node representations towards this notion of graph counterfactual fairness.

Causality in graph neural networks

Recently, inspired by the remarkable success of deep neural networks and especially GNNs, which are arguably universal approximators, some works (Zečević et al. 2021; Yu et al. 2019) investigate the connection between causality and GNNs, take advantage of the power of neural networks as functional approximators to discover causal relations, identify and estimate the causal effects between variables in complex data distribution.

Relating GNNs to structural causal models. GNNs have achieved the state-of-the-art performance in various machine learning tasks on structured data. Zečević et al. (2021) consider GNNs as a viable candidate for causal learning, and analyze the relations between GNNs and structural causal model (SCM) (Pearl 2009). This work makes an exploration to show how to use GNNs for causal computations and embeds causality within neural models. A new model class for GNN-based causal inference that is necessary and sufficient for causal effect identification is developed, and theoretical analysis of this new model class is provided w.r.t. its feasibility, expressivity, and identifiability.

DAG structure learning with GNNs. Bayesian networks (BN) plays an important part in causal inference (Pearl 1988, 2009), however, learning a faithful directed acyclic graph (DAG) structure from samples of data distribution is a challenging task (NP-hard) (Chickering, Heckerman, and Meek 2004) due to the intractable search space superexponential in the number of graph nodes. Encouraged by the success of neural networks in approximation, Yu et al. (2019) develop a graph-based deep generative model to recover the underlying DAG. This method employs the machinery of variational inference, and parameterizes the encoder and decoder with specially designed GNNs.

DISCUSSIONS AND FUTURE WORK

There have been an increasing interest in causality learning with graphs. Despite the contributions of existing works, there still remain lots of potential challenges and tasks to address in the future: (1) **Causal effect estimation under complicated networked data**: Currently, most existing works of causality learning with graphs assume these graphs are simple and homogeneous graphs, while in many real-world scenarios, more complicated relational information can also be considered, for example, heterogeneous graphs, hypergraphs, and knowledge graphs. (2) **Network interference in complicated scenarios**: Although many efforts have been made to address the problem of causal effect estimation under the existence of interference, most of them are still based on some strong assumptions with respect to the network structure (e.g., simple graphs) and the existence of hidden confounders (e.g., strong ignorability assumption) (Ma and Tresp 2021). Future works, which can relax these assumptions, would be highly impactful in real-world applications of causal effect estimation with graphs. (3) **Interpretation in causality learning**: Aside from the performance of causal effect estimation, most of existing works, which control for confounders, especially those based on neural network, lack any interpretation of the causal effect estimation process. In these works, the captured representations are often a mixture of unknown factors without any human-understandable interpretation. A line of works (Hassanpour and Greiner 2019; Zhang, Liu, and Li 2020) identifies disentangled representations to distinguish the underlying factors which influence the treatment, the outcome, or both of them. These works promote the interpretation of the learned representations in a causal perspective, but still lack any semantic understanding. Differently, Ma et al. (2021) uses a disentangled representation learning mechanism to improve the semantic interpretation of learned confounder representations, but this work relies on multicause setting. Besides, existing works of interpretation in causality learning are mostly based on i.i.d. data, while it would be even more challenging for graph data due to the complex structure of relational information. Further works in interpretation of causality learning on graphs would be an interesting direction. (4) **Causal domain adaptation/generalization on graphs**: Recent studies have revealed the importance of incorporating causal perspective to remove the spurious correlations (Arjovsky et al. 2019; Mahajan et al. 2021) and enhance the performance of domain adaptation/generalization. Nevertheless, few of them have considered the graph structure among different instances, while it becomes a more challenging task on graphs as the biases brought from spurious correlations might be amplified through the graph structure. (5) **Online experiments and observational studies on graph data**: One of the key factors hindering the progress in causality learning is the lack of counterfactual data in an individual level. Even through there are some benchmark datasets such as Twins (Almond, Chay, and Lee 2005) and IHDP (Hill 2011) for causal effect estimation evaluation, most of these existing datasets are in i.i.d. scenarios. There have been little work in designing online experiments and finding out comparable counterparts with different treatment assignments. Future works in designing such experiments (e.g., A/B testing in graph data) would be influential in causal inference on graphs; besides, as such experiments are hard to conduct, combining the online experiments together with the offline causal effect estimation from the observational data is also a promising direction.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under Grant IIS-2006844, IIS-2144209, and CNS-2154962, the JP Morgan Chase Faculty Research Award, and the 3 Cavaliers seed grant.

CONFLICT OF INTEREST

The authors declare that there is no conflict.

ORCID

Jing Ma https://orcid.org/0000-0003-4237-6607

REFERENCES

Agarwal, C., H. Lakkaraju, and M. Zitnik. 2021. “Towards a unified framework for fair and stable graph representation learning.” In *Uncertainty in Artificial Intelligence*, pp. 2114–24.

Almond, D., K. Y. Chay, and D. S. Lee. 2005. “The costs of low birth weight.” *The Quarterly Journal of Economics* 120: 1031–83.

Angrist, J. D., and G. W. Imbens. 1995. “Two-stage least squares estimation of average causal effects in models with variable treatment intensity.” *Journal of the American statistical Association* 90: 431–42.

Angrist, J. D., G. W. Imbens, and D. B. Rubin. 1996. "Identification of causal effects using instrumental variables." *Journal of the American statistical Association* 91: 444–55.

Angrist, J. D., and V. Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics* 114: 533–75.

Arjovsky, M., L. Bottou, I. Gulrajani, and D. Lopez-Paz. 2019. "Invariant Risk Minimization." *arXiv preprint arXiv:1907.02893*.

Aronow, P. M., and C. Samii. 2017. "Estimating average causal effects under general interference, with application to a social network experiment." *The Annals of Applied Statistics* 11: 1912–47.

Austin, P. C. 2011. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." *Multivariate Behavioral Research* 46: 399–424.

Basse, G., and A. Feller. 2018. "Analyzing two-stage experiments in the presence of interference." *Journal of the American Statistical Association* 113: 41–55.

Bazarova, N. N. and Y. H. Choi. 2014. "Self-disclosure in social media: extending the functional approach to disclosure motivations and characteristics on social network sites." *Journal of Communication* 64: 635–57.

Bhattacharya, R., D. Malinsky, and I. Shpitser. 2020. "Causal inference under interference and network uncertainty." In *Uncertainty in Artificial Intelligence*.

Bromley, J., J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah. 1993. "Signature verification using a "siamese" time delay neural network." *International Journal of Pattern Recognition and Artificial Intelligence* 7: 25.

Chernozhukov, V., H. Kasahara, and P. Schrimpf. 2021. "Causal impact of masks, policies, behavior on early covid-19 pandemic in the us." *Journal of Econometrics* 220: 23–62.

Chickering, M., D. Heckerman, and C. Meek. 2004. "Large-sample learning of Bayesian networks is NP-hard." *Journal of Machine Learning Research* 5: 1287–330.

Chu, Z., S. L. Rathbun, and S. Li. 2021. "Graph infomax adversarial learning for treatment effect estimation with networked observational data." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dong, Y., J. Ma, C. Chen, and J. Li. 2022. "Fairness in Graph Mining: A Survey." *arXiv preprint arXiv:2204.09888*.

Fatemi, Z., and E. Zheleva. 2020. "Minimizing interference and selection bias in network experiment design." In *Proceedings of the International AAAI Conference on Web and Social Media*.

Fisher, R. A. 1936. "Design of experiments." *British Medical Journal* 1: 554.

Funk, M. J., D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. 2011. "Doubly robust estimation of causal effects." *American Journal of Epidemiology* 173: 761–7.

Goldstein, C. E., C. Weijer, J. C. Brehaut, D. A. Fergusson, J. M. Grimshaw, A. R. Horn, and M. Taljaard. 2018. "Ethical issues in pragmatic randomized controlled trials: a review of the recent literature identifies gaps in ethical argumentation." *BMC Medical Ethics* 19: 14.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. 2012. "A kernel two-sample test." *The Journal of Machine Learning Research* 13: 723–73.

Gretton, A., K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola. 2007. "A kernel statistical test of independence." In *Proceedings of the Conference on Neural Information Processing Systems*.

Guo, R., J. Li, Y. Li, K. S. Candan, A. Raglin, and H. Liu. 2020. "Ignite: a minimax game toward learning individual treatment effects from networked observational data." In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Guo, R., J. Li, and H. Liu. 2020a. "Counterfactual evaluation of treatment assignment functions with networked observational data." In *Proceedings of the SIAM International Conference on Data Mining*.

Guo, R., J. Li, and H. Liu. 2020b. "Learning individual causal effects from networked observational data." In *Proceedings of the International Conference on Web Search and Data Mining*.

Harada, S., and H. Kashima. 2020. "Graphite: Estimating Individual Effects of Graph-structured Treatments." *arXiv preprint arXiv:2009.14061*.

Hassanpour, N., and R. Greiner. 2019. "Learning disentangled representations for counterfactual regression." In *Proceedings of the International Conference on Learning Representations*.

Heckman, J. J. 2000. "Causal parameters and policy analysis in economics: A twentieth century retrospective." *The Quarterly Journal of Economics* 115: 45–97.

Hill, J. L. 2011. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics* 20: 217–40.

Imai, K., Z. Jiang, and A. Malani. 2020. "Causal inference with interference and noncompliance in two-stage randomized experiments." *Journal of the American Statistical Association* 116: 1–39.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

Johansson, F., U. Shalit, and D. Sontag. 2016. "Learning representations for counterfactual inference." In *Proceedings of the International Conference on Machine Learning*.

Junker, B. H., and F. Schreiber. 2011. *Analysis of Biological Networks*. Hoboken, NJ: John Wiley & Sons.

Kaddour, J., Y. Zhu, Q. Liu, M. J. Kusner, and R. Silva. 2021. "Causal effect inference for structured treatments." *Advances in Neural Information Processing Systems*, 34: 24841–4.

Kipf, T. N., and M. Welling. 2017. "Semi-supervised classification with graph convolutional networks." In *Proceedings of the International Conference on Learning Representations*.

Kohavi, R., A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. 2013. "Online controlled experiments at large scale." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva. 2017. "Counterfactual fairness." In *Advances in Neural Information Processing Systems*.

Li, J., H. Dani, X. Hu, J. Tang, Y. Chang, and H. Liu. 2017. "Attributed network embedding for learning in a dynamic environment." In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. 2017. "Causal effect inference with deep latent-variable models." In *Advances in Neural Information Processing Systems*.

Ma, J., Y. Dong, Z. Huang, D. Mietchen, and J. Li. 2021. "Assessing the Causal Impact of Covid-19 Related Policies on Outbreak Dynamics: A Case Study in the US." *arXiv preprint arXiv:2106.01315*.

Ma, J., R. Guo, C. Chen, A. Zhang, and J. Li. 2021. “Deconfounding with networked observational data in a dynamic environment.” In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

Ma, J., R. Guo, M. Wan, L. Yang, A. Zhang, and J. Li. 2022. “Learning fair node representations with graph counterfactual fairness.” In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*.

Ma, J., R. Guo, A. Zhang, and J. Li. 2021. “Multi-cause effect estimation with disentangled confounder representation.” In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Ma, J., M. Wan, L. Yang, J. Li, B. Hecht, and J. Teevan. 2022. “Learning causal effects on hypergraphs.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1202–12).

Ma, Y., and V. Tresp. 2021. “Causal inference under networked interference and intervention policy enhancement.” In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Mahajan, D., S. Tople, and A. Sharma. 2021. “Domain generalization using causal matching.” In *Proceedings of the International Conference on Machine Learning*.

Medsker, L. R., and L. Jain. 2001. *Recurrent Neural Networks. Design and Applications*. Boca Raton: CRC Press.

Newman, M. E. 2001. “The structure of scientific collaboration networks.” *Proceedings of the National Academy of Sciences* 98: 404–9.

Ouyang, M. 2014. “Review on modeling and simulation of interdependent critical infrastructure systems.” *Reliability Engineering & System Safety* 121: 43–60.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California: Morgan Kaufmann Publishers.

Pearl, J. 1995. “Causal diagrams for empirical research.” *Biometrika* 82: 669–88.

Pearl, J. 2009. *Causality*. Cambridge, England: Cambridge University Press.

Rakesh, V., R. Guo, R. Moraffah, N. Agarwal, and H. Liu. 2018. “Linked causal variational autoencoder for inferring paired spillover effects.” In *Proceedings of the International Conference on Information and Knowledge Management*.

Robinson, P. M. 1988. “Root-n-consistent semiparametric regression.” *Econometrica: Journal of the Econometric Society* 56: 931–54.

Schnabel, T., A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. 2016. “Recommendations as treatments: debiasing learning and evaluation.” In *Proceedings of the International Conference on Machine Learning*.

Shalit, U., F. D. Johansson, and D. Sontag. 2017. “Estimating individual treatment effect: generalization bounds and algorithms.” In *Proceedings of the International Conference on Machine Learning*.

Shi, C., D. Blei, and V. Veitch. 2019. “Adapting neural networks for the estimation of treatment effects.” *Advances in Neural Information Processing Systems* 32.

Sobel, M. E. 2006. “What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference.” *Journal of the American Statistical Association* 101: 1398–407.

Splawa-Neyman, J., D. M. Dabrowska, and T. Speed. 1990. “On the application of probability theory to agricultural experiments. Essay on principles. Section 9.” *Statistical Science* 5: 465–72.

Tchetgen, E. J. T., and T. J. VanderWeele. 2012. “On causal inference in the presence of interference.” *Statistical Methods in Medical Research* 21: 55–75.

Ugander, J., B. Karrer, L. Backstrom, and J. Kleinberg. 2013. “Graph cluster randomization: network exposure to multiple universes.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Velickovic, P., W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. 2019. “Deep graph infomax.” In *Proceedings of the International Conference on Learning Representations (Poster)*.

Villani, C. 2009. *Optimal Transport: Old and New*. Berlin: Springer.

Wager, S., and S. Athey. 2018. “Estimation and inference of heterogeneous treatment effects using random forests.” *Journal of the American Statistical Association* 113: 1228–42.

Wu, Z., S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. 2020. “A comprehensive survey on graph neural networks.” *IEEE Transactions on Neural Networks and Learning Systems* 32: 4–24.

Yao, L., S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. 2018. “Representation learning for treatment effect estimation from observational data.” In *Advances in Neural Information Processing Systems*.

Yu, Y., J. Chen, T. Gao, and M. Yu. 2019. “DAG-GNN: DAG structure learning with graph neural networks.” In *Proceedings of the International Conference on Machine Learning*.

Yuan, Y., K. Altenburger, and F. Kooti. 2021. “Causal network motifs: identifying heterogeneous spillover effects in A/B tests.” In *Proceedings of the Web Conference*.

Zečević, M., D. S. Dhami, P. Veličković, and K. Kersting. 2021. “Relating Graph Neural Networks to Structural Causal Models.” *arXiv preprint arXiv:2109.04173*.

Zhang, W., L. Liu, and J. Li. 2020. “Treatment effect estimation with disentangled latent factors.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 10923–30.

Zhou, J., G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. 2020. “Graph neural networks: a review of methods and applications.” *AI Open* 1: 57–81.

AUTHOR BIOGRAPHIES

Jing Ma is a Ph.D. candidate in the Department of Computer Science at University of Virginia. Her research interests include causal inference, machine learning, data mining, and graph mining. Especially, her research targets on bridging the gap between causality and machine learning. Her works have been published in many top conferences and journals such as KDD, IJCAI, WWW, AAAI, TKDE, WSDM, SIGIR, and IPSN.

Jundong Li is an Assistant Professor in the Department of Electrical and Computer Engineering, with a joint appointment in the Department of Computer Science, and the School of Data Science at the University of Virginia. He received Ph.D. degree in Computer Science at Arizona State University in 2019. His research interests are in data mining and machine learning. He has published over 100 articles in high-impact venues

(e.g., KDD, WWW, AAAI, IJCAI, WSDM, EMNLP, CSUR, TPAMI, TKDE, TKDD, and TIST.). He has won several prestigious awards, including NSF CAREER Award, JP Morgan Chase Faculty Research Award, Cisco Faculty Research Award, and being selected for the AAAI 2021 New Faculty Highlights program.

How to cite this article: Ma, J., and J. Li. 2022. "Learning causality with graphs." *AI Magazine* 43: 365–375. https://doi.org/10.1002/aaai.12070