Learning Causal Effects on Hypergraphs

Jing Ma
University of Virginia
Charlottesville, VA, USA
jm3mr@virginia.edu

Mengting Wan
Microsoft
Redmond, WA, USA
mengting.wan@microsoft.com

Longqi Yang
Microsoft
Redmond, WA, USA
longqi.yang@microsoft.com

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

Brent Hecht
Microsoft
Redmond, WA, USA
brent.hecht@microsoft.com

Jaime Teevan
Microsoft
Redmond, WA, USA
teevan@microsoft.com

ABSTRACT

Hypergraphs provide an effective abstraction for modeling multi-way group interactions among nodes, where each hyperedge can connect any number of nodes. Different from most existing studies which leverage *statistical dependencies*, we study hypergraphs from the perspective of *causality*. Specifically, in this paper, we focus on the problem of individual treatment effect (ITE) estimation on hypergraphs, aiming to estimate how much an intervention (e.g., wearing face covering) would causally affect an outcome (e.g., COVID-19 infection) of each individual node. Existing works on ITE estimation either assume that the outcome on one individual should not be influenced by the treatment assignments on other individuals (i.e., no *interference*), or assume the interference only exists between pairs of connected individuals in an ordinary graph. We argue that these assumptions can be unrealistic on real-world hypergraphs, where higher-order interference can affect the ultimate ITE estimations due to the presence of group interactions. In this work, we investigate high-order interference modeling, and propose a new causality learning framework powered by hypergraph neural networks. Extensive experiments on real-world hypergraphs verify the superiority of our framework over existing baselines.

CCS CONCEPTS

• **Mathematics of computing** → **Causal networks**; **Hypergraphs**; • **Information systems** → **Social networks**.

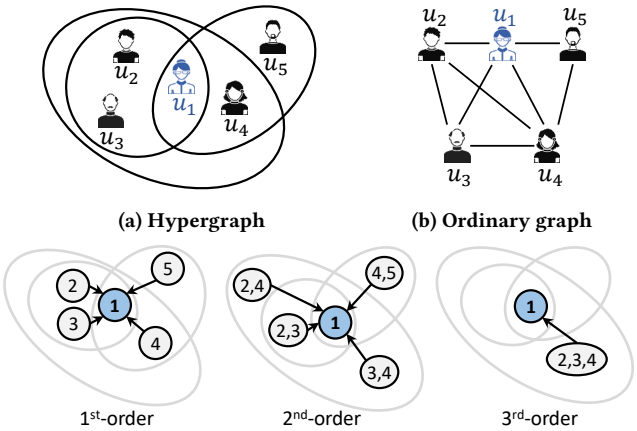
KEYWORDS

Causal Inference; Graph Mining; Hypergraph; Interference

ACM Reference Format:
Jing Ma, Mengting Wan, Longqi Yang, Jundong Li, Brent Hecht, and Jaime Teevan. 2022. Learning Causal Effects on Hypergraphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539299

1 INTRODUCTION

Group interactions among individuals exist in a wide range of scenarios, e.g., massive gathering events, day-to-day group chats on WhatsApp or WeChat, and workplace interactions on Microsoft Teams or Slack channels. Although the conventional pairwise graph definition covers a vast number of applications (e.g., person-to-person physical contact networks or social networks [10]), it fails to capture the complete information of these group interactions (where each interaction may involve more than two individuals) [5, 13, 44]. The notion of the *hypergraph* can thus be introduced to address this limitation. Consider a hypergraph example that individuals are connected via in-person social events, each gathering event can be represented as a *hyperedge* (Fig. 1a). Each hyperedge can connect an arbitrary number of individuals, in contrast to an ordinary edge which connects exactly two nodes (Fig. 1b).



(a) Hypergraph (b) Ordinary graph

(c) First, second, and third-order interferences with u_1.

Figure 1: (a) An illustrative example of group interactions on a hypergraph, where each circle represents a hyperedge (group); (b) An ordinary graph projected from this hypergraph; (c) Illustration of interferences with u_1 from its neighbors on the hypergraph. Note interference on (b) is pairwise (first-order only) while higher-order interference exists on the original hypergraph (a).

While many studies have been devoted to utilizing such a generalized hypergraph structure to facilitate machine learning tasks [5, 13, 44, 52], the majority were still executed at the statistical correlation level, e.g., predicting the COVID-19 infection risk on each individual (node) by capturing the *correlations* between one's demographic information (node features), in-person group gathering history (hypergraph structure) and the infection outcomes (node labels). A critical limitation here is the lack of *causality*, which

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539299

is particularly important for understanding the impact of a policy intervention (e.g., wearing face covering) on an outcome of interest (e.g., COVID-19 infection). For individuals connected as in Fig. 1a, one may ask "how would each individual's face covering practice (treatment) *causally* influence their infection risk (outcome)?" Such a causal inference task requires constructing the counterfactual state of the same individual by holding all other possible factors constant except the treatment variable of interest. This is a particularly hard problem on hypergraph data, since the outcome of each individual is not only affected by their own confounding factors (e.g., one's health conditions and vaccine status) but also interfered by other individuals on the hypergraph (e.g., face covering practice of other individuals who may physically contact the target individual through a gathering event).

In this paper, we focus on learning causal effects on hypergraphs. We are specifically interested in estimating the individual treatment effect (ITE) under hypergraph interference from observational data. Our study is motivated by the following gaps: (i) *Empirical constraints of randomized experiments.* One of the most reliable approaches for treatment effect estimation is randomized controlled trials (RCTs). Nevertheless, running RCTs is often expensive, impractical, even unethical [15], and they are especially difficult on graphs due to the dependencies among connected nodes [39]. (ii) *High-order interference on hypergraphs.* Our work focuses on the problem of ITE estimation, which aims to estimate the causal effect of a certain treatment (e.g., face covering practice) on an outcome (e.g., COVID-19 infection) for each individual. The classic ITE estimation is based on the Stable Unit Treatment Value (SUTVA) assumption [14, 36] that there is no interference [20, 37] (i.e., spillover effect) among instances (also referred to as *units* in causal inference literature). That means the outcomes for any instance are not influenced by the treatment assignment of other instances. This assumption can be impractical in the real-world, thus resulting in flawed causal effect estimations, especially on graphs where the interference among instances are ubiquitous [1, 47, 50]. There have been many efforts addressing this problem [3, 6, 21, 25, 28, 37, 39, 49], but most assume the interference only exists in a pairwise way on ordinary graphs (as shown in Fig. 1b). This pairwise interference notion is insufficient to characterize the high-order interference that exists on hypergraphs. As shown in Fig. 1c, within a gathering event (hyperedge) between u_1, u_2 and u_3, an individual's (u_1) infection outcome can be affected by the *first-order* interference from other individuals ($u_2 \rightarrow u_1$ and $u_3 \rightarrow u_1$) as well as the *high-order* interference from the interactions among other individuals (the interaction between u_2 and u_3 may also act on influencing the exposure of the virus to u_1; consequently, u_1's infection risk can be affected by this second-order interaction effect, i.e., $u_2 \times u_3 \rightarrow u_1$). Notice that the number of such high-order interference items grows combinatorially as the size of a hyperedge increases, leading to a significant information gap between the original hypergraph and the projected pairwise ordinary graph (which accounts for the first-order interference only). This demands techniques capable of modeling high-order interference, but to the best of our knowledge, very little work has been done in this area.

In this paper, we propose a novel framework— **C**ausal **I**nference under **S**pillover Effects in **Hyper**graphs (***HyperSCI***)—to model high-order interference. At a high-level, this framework controls for the confounders and models high-order interference based on representation learning, then estimates the outcomes based on the learned representations. More specifically: (i) *Controlling for Confounders.* Our framework is based on the widely accepted unconfoundedness assumption [33], i.e., the confounders are contained in the observed features. With this assumption, we leverage representation learning techniques to capture and control for confounders from the features of each individual. Note as shown in previous works [34], the discrepancy between confounder distributions in the treatment group and the control group can lead to biases in causal effect estimations. Therefore, we also propose to use a representation balancing technique to mitigate the discrepancy between these two distributions. (ii) *Modeling High-order Interference.* Modeling high-order relationships can be challenging due to the complexity of enumerating multi-way interactions among nodes within each hyperedge. Historically, one may need to simplify the original hypergraph and approximate it through a series of projected ordinary graphs [48]. This obstacle is fortunately unblocked by the recent advances of hypergraph neural networks [5, 44]. We extend this line of techniques to model interference by learning interference representations for each node. To learn the interference representations, the learned confounder representations and the treatment assignment are propagated via hypergraph convolution and attention operations. (iii) *Outcome Prediction.* Based on the learned representations of confounders and interference, we predict the potential outcomes corresponding to different treatment assignments for each individual. Overall, the main contributions of this work can be summarized as follows:

- We formalize the problem of ITE estimation under high-order interference on hypergraphs. To the best of our knowledge, it is the first work for this problem.
- We propose a novel framework **HyperSCI** for the studied problem. **HyperSCI** models confounders and high-order interference via representation learning and hypergraph neural networks.
- We validate the effectiveness of the proposed framework through extensive experiments and provide in-depth analysis on how it acts on different nodes and hyperedges.

2 PROBLEM DEFINITION AND ANALYSIS

We provide the formal problem definition and a brief theoretical analysis of our studied problem in this section. A notation table is provided in Appendix A.

Definition 2.1. *Suppose a set of individuals $\mathcal{V} = \{v_i\}_{i=1}^n$ are connected via hyperedges $\mathcal{E} = \{\mathbf{e}_k\}_{k=1}^m$, together these form a **hypergraph** $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$ with n nodes and m hyperedges, where each hyperedge can connect an arbitrary number of nodes.*

The observational data on this hypergraph can be denoted as $\{\mathbf{X}, \mathcal{H}, \mathbf{T}, \mathbf{Y}\}$, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{T} = \{t_i\}_{i=1}^n$ and $\mathbf{Y} = \{y_i\}_{i=1}^n$ represent node features, treatment assignments, and observed outcomes, respectively. $\mathbf{H} = \{h_{i,e}\} \in \mathbb{R}^{n \times m}$ is an incidence matrix which describes the hypergraph structure of $\mathcal{H}$. $h_{i,e} = 1$ if node i is in hyperedge e, otherwise $h_{i,e} = 0$. For ease of discussion, we consider the treatment assignment for each node as a binary variable in this study (i.e., $t_i \in \{0, 1\}$), but our work can be extended to non-binary categorical variables and continuous variables.

Definition 2.2. *The **potential outcome** [33] of the instance i (denoted by y_i^1 or y_i^0) is defined as the realized value of outcome for instance i under the treatment value $t_i = 1$ or $t_i = 0$. These potential outcomes can be instantiated via a transformation $Y_i^{T_i} = \Phi_Y(T_i, X_i, T_{-i}, X_{-i}, H)$. Φ_Y can be regarded as a (non-deterministic) function to output potential outcomes, which takes each node's treatment assignment, node features, the information (treatment assignments and node features) of other nodes on the hypergraph, and the hypergraph structure as input[1], i.e., $y_i^{t_i} = \Phi_Y(t_i, \mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H})$, where the subscript $-i$ denotes all other nodes on $\mathcal{H}$ except i.*

Given the above preliminaries, we are ready to provide the formal definition of individual treatment effect on hypergraphs.

Definition 2.3. *For each node i on the hypergraph $\mathcal{H}$, the **individual treatment effect** (ITE) is defined by the difference between potential outcomes corresponding to $t_i = 1$ and $t_i = 0$:*

$$\begin{aligned}\tau(\mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H}) &= \mathbb{E}[Y_i^1 - Y_i^0 | X_i = \mathbf{x}_i, T_{-i} = \mathbf{T}_{-i}, X_{-i} = \mathbf{X}_{-i}, H = \mathbf{H}] \\ &= \mathbb{E}[\Phi_Y(1, \mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H}) - \Phi_Y(0, \mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H})].\end{aligned} \tag{1}$$

We clarify that the ITE in this paper is actually defined in the form of conditional average treatment effect (CATE), similar as [17, 28]. The expectation is taken over the potential outcome (output of Φ_Y) of the instances with same node features $\mathbf{x}_i$ and "environmental information" (hypergraph structure $\mathbf{H}$, other nodes' features $\mathbf{X}_{-i}$ and treatments $\mathbf{T}_{-i}$). The distribution of the output of Φ_Y is equivalent to the conditional distribution of the potential outcome conditioned on the parameters in Φ_Y with fixed values. For notation simplicity, we also denote $\tau_i = \tau(\mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H})$ in this paper. Meanwhile, we introduce the notion of spillover effect in this work to assess the level of interference on hypergraphs.

Definition 2.4. *The **spillover effect** of node i under its treatment t_i and other nodes' treatment assignment $\mathbf{T}_{-i}$ on the hypergraph $\mathcal{H}$ is defined as:*

$$\delta_i = \mathbb{E}[\Phi_Y(t_i, \mathbf{x}_i, \mathbf{T}_{-i}, \mathbf{X}_{-i}, \mathbf{H}) - \Phi_Y(t_i, \mathbf{x}_i, \mathbf{0}, \mathbf{X}_{-i}, \mathbf{H})]. \tag{2}$$

In this paper, given the observed data $\{\mathbf{X}, \mathcal{H}, \mathbf{T}, \mathbf{Y}\}$, we aim to estimate the ITE defined in Eq. 1 for each node in $\mathcal{H}$ with the existence of high-order interference defined in Eq. 2.

2.1 Theoretical Analysis

With the above definitions, we show that the ITE can be identifiable from the observational data under the following two assumptions.

Similar as the assumptions in other works of causal inference under network interference [28], for each individual, we assume there exists a summary function capable of characterizing all the "environmental" information related to this node on the hypergraph. Suppose there is a summary function $\mathtt{SMR}(\cdot)$: for each node i, $\mathtt{SMR}(\cdot)$ takes the hypergraph structure $\mathbf{H}$, the treatment assignment of other nodes $\mathbf{T}_{-i}$ and the features of these nodes $\mathbf{X}_{-i}$ as input, then maps them into a vector $\mathbf{o}_i$:

$$\mathbf{o}_i = \mathtt{SMR}(\mathbf{H}, \mathbf{T}_{-i}, \mathbf{X}_{-i}). \tag{3}$$

[1]In this paper, we use non-bold, italicized, and capitalized letters (e.g., X_i) to denote random variables; non-bold lowercase letters (e.g., t_i) to denote observed values of a scalar; bold lowercase letters (e.g., $\mathbf{x}_i$) to denote observed values of a vector; bold capitalized letters (e.g., $\mathbf{X}$) to denote observed values of a matrix or a set.

We use H, X, T to denote the random variables for the hypergraph structure, features, and treatment assignment for any node. Then our first assumption can be formalized as below.

Assumption 1. *(Expressiveness of summary function) For any node i, any values of H, X_{-i}, and T_{-i}, if the output of summary function $\mathbf{o}_i$ is determined, then the value of the potential outcomes y_i^1 and y_i^0 with feature $\mathbf{x}_i$ are also determined.*

Our second assumption extends the unconfoundedness assumption [33] to the hypergraph interference setting. That is we assume conditioned on the above summary function, the observed features can capture all possible confounders.

Assumption 2. *(Unconfoundedness) For any node i, given the node features, the potential outcomes are independent with the treatment assignment and summary of neighbors, i.e., $Y_i^1, Y_i^0 \perp\!\!\!\perp T_i, O_i | X_i$.*

Based on the above assumptions, the identification of the expectation of potential outcomes Y_i^1 and Y_i^0 can be proved (here we take Y_i^1 as an example):

$$\mathbb{E}[Y_i^1 | T_i = 1, X_i = \mathbf{x}_i, T_{-i} = \mathbf{T}_{-i}, X_{-i} = \mathbf{X}_{-i}, H = \mathbf{H}] \tag{4}$$

$$\overset{(a)}{=\!=} \mathbb{E}[\Phi_Y(T_i = 1, X_i = \mathbf{x}_i, T_{-i} = \mathbf{T}_{-i}, X_{-i} = \mathbf{X}_{-i}, H = \mathbf{H})] \tag{5}$$

$$\overset{(b)}{=\!=} \mathbb{E}[\Phi_Y(T_i = 1, X_i = \mathbf{x}_i, O_i = \mathbf{o}_i)] \tag{6}$$

$$\overset{(c)}{=\!=} \mathbb{E}[\Phi_Y(T_i = 1, X_i = \mathbf{x}_i, O_i = \mathbf{o}_i) | X_i = \mathbf{x}_i] \tag{7}$$

$$\overset{(d)}{=\!=} \mathbb{E}[\Phi_Y(T_i = 1, X_i = \mathbf{x}_i, O_i = \mathbf{o}_i) | X_i = \mathbf{x}_i, T_i = 1, O_i = \mathbf{o}_i] \tag{8}$$

$$\overset{(e)}{=\!=} \mathbb{E}[Y_i | X_i = \mathbf{x}_i, T_i = 1, O_i = \mathbf{o}_i]. \tag{9}$$

Here, the equation (a) is based on the definition of potential outcome in this setting; (b) is inferred from Assumption (1); (c) is a straightforward derivation; (d) is based on Assumption (2); and (e) is based on the widely used consistency assumption [33]. Based on the above proof for the identification of potential outcomes, the identification of ITE can be straightforwardly derived.

3 THE PROPOSED FRAMEWORK

Inspired by the previous theoretical analysis, we propose a novel framework **HyperSCI** to address the studied problem. This framework contains three components: confounder representation learning, interference modeling, and outcome prediction. Holistically, we aim to learn an expressive transformation to summarize high-order interferences (Assumption 1), then take the interference representation, the confounder representation as well as the treatment assignment to estimate the expected potential outcome (Assumption 2). The illustration of **HyperSCI** is shown in Fig. 2.

3.1 Confounder Representation Learning

We first encode the node features $\mathbf{x}_i$ into a latent space via a multi-layer perceptron (MLP) module, i.e., $\mathbf{z}_i = \mathrm{MLP}(\mathbf{x}_i)$. This results in a set of representations $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$, which is expected to capture all potential confounders, so the model can mitigate the confounding biases by controlling for the learned representation $\mathbf{z}_i$.

Representation Balancing. Note a discrepancy may exist between the distributions of confounder representation $\mathbf{Z}$ in the treatment

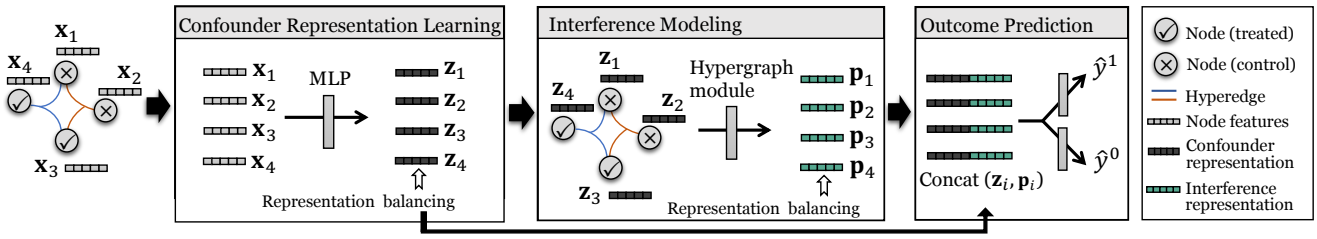


Figure 2: An illustration of the proposed framework HyperSCI, which includes three key components: confounder representation learning, interference modeling, and outcome prediction.

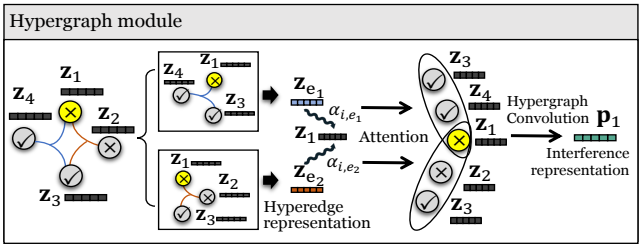


Figure 3: A detailed illustration of the hypergraph module in the interference modeling component of HyperSCI. Here we use the node v_1 (highlighted in yellow) as an example.

group and the control group, incurring biases in causal effect estimation, as shown in [34, 46]. To minimize this discrepancy, we leverage the representation balancing technique by adding a discrepancy penalty to the loss function, where this discrepancy penalty can be calculated with any distribution distance metrics. In our implementation, we use the Wasserstein-1 distance [34] between the representation distributions of treatment group and control group.

3.2 Interference Modeling

In this interference modeling component, we take the confounder representation ($\mathbf{Z}$), the treatment assignment ($\mathbf{T}$), and the relational information on the hypergraph ($\mathbf{H} = \{h_{i,e}\}$) as input, to capture the high-order interference for each individual. More specifically, we learn a transformation function $\Psi(\cdot)$ through a hypergraph module to generate the interference representations ($\mathbf{p}_i$) for each node i, i.e., $\mathbf{p}_i = \Psi(\mathbf{Z}, \mathbf{H}, \mathbf{T}_{-i}, t_i)$. As shown in Fig. 3, this module is implemented with a hypergraph convolutional network [5, 44] and a hypergraph attention mechanism [5, 11, 52], where the convolutional operator forms the skeleton of interference from hyperedges, and the attention operator enhances this mechanism by allowing flexible node contributions to each hyperedge.

Learning interference representations. To learn the representations which encode the interference in the hypergraph for each node, we propagate the treatment assignment and confounder representations with a hypergraph convoluntional layer. We first introduce a vanilla Laplacian matrix for the hypergraph $\mathcal{H}$:

$$\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{H}\mathbf{B}^{-1}\mathbf{H}^\top\mathbf{D}^{-1/2}. \tag{10}$$

Here $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix where each element represents the node degree (i.e., $\sum_{e=1}^{m} h_{i,e}$). $\mathbf{B} \in \mathbb{R}^{m\times m}$ is another diagonal matrix, where each element is the size of each hyperedge ($\sum_{i=1}^{n} h_{i,e}$). Then we can define the hypergraph convolution operator as:

$$\mathbf{P}^{(l+1)} = \text{LeakyReLU}\left(\mathbf{L}\mathbf{P}^{(l)}\mathbf{W}^{(l+1)}\right), \tag{11}$$

where $\mathbf{P}^{(l)}$ represents the representations from the l-th layer in the hypergraph module. We feed the first layer with the previous confounder representation masked by the treatment assignment, i.e., $\mathbf{p}_i^{(0)} = t_i * \mathbf{z}_i$, where $*$ denotes element-wise multiplication. $\mathbf{W}^{(l+1)} \in \mathbb{R}^{d^{(l)}\times d^{(l+1)}}$ is the parameter matrix in the $(l+1)$-th layer, where $d^{(l)}$ and $d^{(l+1)}$ refer to the dimensionality of the interference representations in the l-th and $(l+1)$-th layers, respectively.

Modeling interference with different significance. Although the above convolution layer can pass interferences through hyperedges, it does not provide much flexibility to account for the significance of interference for different nodes through different hyperedges. In the aforementioned COVID-19 example, intuitively, those individuals who are active in certain group gathering events are more likely to influence or be influenced by others in these groups. To better capture this intrinsic relationship between nodes and hyperedges on a hypergraph, we leverage a hypergraph attention mechanism [5, 11, 52] to learn attention weights for each node and the corresponding hyperedges that contain this node.

More specifically, we compute a representation for each hyperedge (e) by aggregating across its associated nodes ($\mathcal{N}_e$): $\mathbf{z}_e = \text{Agg}(\{\mathbf{z}_i \mid i \in \mathcal{N}_e\})$. Here, $\text{Agg}(\cdot)$ can be any aggregation functions (e.g., the mean aggregation). For each node i and its associated hyperedge e, the attention score between a node i and a hyperedge e can be calculated as:

$$\alpha_{i,e} = \frac{\exp(\sigma(\text{sim}(\mathbf{z}_i\mathbf{W}_a, \mathbf{z}_e\mathbf{W}_a)))}{\sum_{k\in\mathcal{E}_i}\exp(\sigma(\text{sim}(\mathbf{z}_i\mathbf{W}_a, \mathbf{z}_k\mathbf{W}_a)))}, \tag{12}$$

where $\sigma(\cdot)$ is a non-linear activation function, $\mathcal{E}_i$ denotes the set of hyperedges associated with the node i. Here we use $\mathbf{W}_a$ to denote a parameter matrix to compute the node-hyperedge attention. $\text{sim}(\cdot)$ is a similarity function, which can be implemented as:

$$\text{sim}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{a}^\top[\mathbf{x}_i \| \mathbf{x}_j], \tag{13}$$

where $\mathbf{a}$ is a weight vector, $[\cdot\|\cdot]$ is a concatenation operation.

Then we use the attention scores to model the interference with different significance. Specifically, we replace the original incidence matrix $\mathbf{H}$ in Eq. 10 with an enhanced matrix $\widetilde{\mathbf{H}} = \{\widetilde{h}_{i,e}\}$, where $\widetilde{h}_{i,e} = \alpha_{i,e}h_{i,e}$. In this way, the interference from different nodes on the same hyperedge can be assigned with different importance weights, indicating different levels of contribution for interference modeling. We denote the final representations from the last convolution layer as $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{n}$ and expect it to capture the high-order interference for each node.

Representation Balancing. Similar to the coufounder representation learning module, we calculate a discrepancy penalty to reflect the difference between the distributions of interference representations in treatment and control groups. We sum up these two discrepancy penalties together to compute a representation balancing loss, denoted by $\mathcal{L}_b$.

3.3 Outcome Prediction

With the confounder representation $\mathbf{z}_i$ and the interference representation $\mathbf{p}_i$, we model the potential outcomes as:

$$\widehat{y}_i^1 = f_1([\mathbf{z}_i \| \mathbf{p}_i]),\ \widehat{y}_i^0 = f_0([\mathbf{z}_i \| \mathbf{p}_i]), \tag{14}$$

where $f_1(\cdot)$ and $f_0(\cdot)$ are learnable functions to predict the potential outcome w.r.t. $t = 1$ and $t = 0$. We implement $f_1(\cdot)$ and $f_0(\cdot)$ with two MLP modules. Then the prediction for the observed outcome is obtained by $\widehat{y}_i = \widehat{y}_i^{t_i}$. We optimize the model to minimize the following loss function:

$$\mathcal{L} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \alpha \mathcal{L}_b + \lambda \|\Theta\|^2, \tag{15}$$

where the first term is the standard mean squared error, $\mathcal{L}_b$ is the representation balancing loss, Θ represents the parameters in this neural network model. α and λ are two hyperparameters which control the weights for the representation balancing loss and the parameter regularization term. The ITE for each instance i can be estimated as: $\widehat{\tau}_i = \widehat{y}_i^1 - \widehat{y}_i^0$.

3.4 Discussion

Here we revisit some implicit assumptions in the proposed framework. First, we assume that the interference for each node i comes from its neighbors through the hypergraph structure. Here, the interference from neighbors that are multiple hops away can also be captured by stacking more hypergraph convoluntional layers. Second, for simplicity, we assume that the interference for each node only comes from other nodes with non-zero treatment assignment. Third, we assume that the representations of nodes in the same hyperedge are similar in the latent space. Besides, following [5], we assume that the representations of hyperedges are homogeneous with node representations. Nevertheless, we should still mention that this proposed framework is general and extendable, where the above assumptions can be further relaxed by enriching the hypergraph processing module.

4 EXPERIMENTS

It is typically very hard to obtain the ground-truth counterfactual data as only one of the two potential outcomes can be obtained in the observational data. Hence, in this section, we follow a standard practice to evaluate the proposed framework and the alternative approaches on three semi-synthetic datasets. We aim to leverage as much real-world information as possible in the simulated environment. Our datasets are all based on real-world hypergraph data and we retain the treatment allocations as well as node features (covariates) if they are available. We simulate the outcome generation process to assess the true individual treatment effect (ITE), which eventually allows us to evaluate the performance of the ITE estimation from different causal inference approaches.

4.1 Dataset and Simulation

We obtain the semi-synthetic data based on two publicly available hypergraph datasets (**Contact** [7, 29], **Goodreads** [40, 41]) and one large-scale proprietary web application dataset (**Microsoft Teams**). We do not account for the temporal information of each hyperedge in our experiments and leave this as a future research direction instead. In all three datasets, we discard extremely large hyperedges and keep those with no more than 50 nodes only.[2]

4.1.1 Outcome Simulation. Given the treatment allocations $\mathbf{T}$, node features $\mathbf{X}$, and the hypergraph structure $\mathbf{H}$, the potential outcome of an individual i can be simulated via

$$y_i = f_{y,0}(\mathbf{x}_i) + \overbrace{\gamma f_t(t_i, \mathbf{x}_i)}^{\text{individual treatment effect (ITE)}} + \underbrace{\beta f_s(\mathbf{T}, \mathbf{X}, \mathbf{H})}_{\text{hypergraph spillover effect}} + \epsilon_{y_i}, \tag{16}$$

where $f_{y,0}(\mathbf{x}_i)$ describes the outcome of instance i when $t_i = 0$ and without network interference, $f_t(\cdot)$ calculates the ITE of each instance, $f_s(\cdot)$ calculates the spillover effect, and ϵ_{y_i} denotes the random noise from a Gaussian distribution $\mathcal{N}(0, 1)$. We specify $f_{y,0}(\mathbf{x}_i)$ as a linear transformation of $\mathbf{x}_i$:

$$f_{y,0} = \mathbf{w}_0 \mathbf{x}_i, \tag{17}$$

where $\mathbf{w}_0 \sim \mathcal{N}(0, \mathbf{I}), \mathbf{w}_0 \in \mathbb{R}^d$. Then we control the individual treatment effect ($f_t(t_i, \mathbf{x}_i)$) and the hypergraph spillover effect ($f_s(\mathbf{T}, \mathbf{X}, \mathbf{H})$) under two different settings:

(1) **Linear.**

$$f_t(t_i, \mathbf{x}_i) = \begin{cases} \mathbf{w}_1 \mathbf{x}_i + \epsilon & \text{if } t_i = 1 \\ 0 & \text{if } t_i = 0 \end{cases} \tag{18}$$

Here $\mathbf{w}_1 \in \mathbb{R}^d$, and each element in $\mathbf{w}_1$ follows a Gaussian distribution. We generate f_s as:

$$f_s(\mathbf{T}, \mathbf{X}, \mathbf{H}) = \frac{1}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} \sigma'\Big(\frac{1}{|\mathcal{N}_e|} \sum_{j \in \mathcal{N}_e} t_j \times f_t(t_j, \mathbf{x}_j)\Big). \tag{19}$$

Here, $\sigma'(\cdot)$ is a function on the aggregation over each hyperedge. We implement it with an identity function by default.

(2) **Quadratic.**

$$f_t(t_i, \mathbf{x}_i) = \begin{cases} \mathbf{x}_i^\top \mathbf{W}_t \mathbf{x}_i + \epsilon & \text{if } t_i = 1 \\ 0 & \text{if } t_i = 0 \end{cases} \tag{20}$$

Here $\mathbf{W}_t \in \mathbb{R}^{d \times d}$, and each element in $\mathbf{W}_t$ follows a Gaussian distribution. We generate f_s as:

$$f_s(\mathbf{T}, \mathbf{X}, \mathbf{H}) = \frac{1}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} \sigma'\Big(\frac{1}{|\mathcal{N}_e|^2} (\mathbf{T}_e * \mathbf{X}_e) \mathbf{W}_t (\mathbf{T}_e * \mathbf{X}_e)^\top\Big). \tag{21}$$

Here $\mathbf{X}_e$ and $\mathbf{T}_e$ are the feature matrix and treatment assignment of nodes contained in hyperedge e, respectively. Here $*$ denotes element-wise multiplication.

[2] Note hyperedges with large size of nodes are usually less meaningful [7].

4.1.2 Dataset Details. We follow the above process to generate potential outcomes on all three datasets. Additional details about each dataset are provided as the follows.

Contact. This dataset collects interactions recorded by wearable sensors among students at a high school [7, 29], and includes 327 nodes and 7,818 hyperedges. Each node represents a person, and each hyperedge stands for a group of individuals are in close physical proximity to each other. This contact hypergraph data allows us to simulate a hypothetical question: "how does one's face covering practice (treatment) causally affect their infection risk of an infectious disease (outcome)?". In each group contact, one may bring the virus to the surrounding environment, and thus affect other people's infection risk. Due to the lack of detailed information about each individual, apart from the potential outcome, we also generate the treatment (t_i) and the covariates ($\mathbf{x}_i$) as the follows:

$$\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}), t_i \sim Ber(\text{sigmoid}(\mathbf{x}_i \mathbf{v}_t)), \tag{22}$$

where $\mathbf{I}$ is an $d \times d$ identity matrix, here we set $d = 50$. $\mathbf{v}_t$ is a d-dimensional vector where each element inside follows a Gaussian distribution. Eventually about 50% $\sim$ 60% of the nodes are treated ($t_i = 1$) in our experiments.

GoodReads. This dataset collects book information from the book review website GoodReads[3], including the book title, authors, descriptions, reviews, and ratings [40, 41]. We take each book in the *Children* category as an instance. The bag-of-words of the book descriptions are used as the covariates of each book. Each hyperedge corresponds to each author and all books sharing the same author are in the same hyperedge. The real-world book ratings are considered as treatment assignments: for each node i, we define $t_i = 1$ if the rating score is larger than 3 and $t_i = 0$ otherwise. We aim to study the causal effect of the rating score on the sales of each book. The ratings of each author's books can establish this author's overall reputation, and thus influence the sales of other books from the same author. The final processed dataset includes 57,031 nodes (where 40% are treated) and 12,709 hyperedges. Note each book may have more than one author, and each author may have published multiple books.

Microsoft Teams. We sampled 91,391 anonymized employees of a multinational technology company and collected their aggregated telemetry data on Microsoft Teams[4]. Microsoft Teams is a workplace communication platform where users are allowed to create a group space (i.e., "team" or "channel") to enable public communication within each group. We are interested in how a user's usage of these group spaces causally affects their productivity. We process the treatment assignment into binary values by taking it as 1 if the employee has sent out at least one message in any of these group spaces during the first week of March, 2021; otherwise the treatment is assigned as 0. Each group space can be regarded as a hyperedge, where information can be shared via group discussions thus one's activeness on this platform may affect other individuals' outcomes in the same group. Employee demographics (e.g., office location, job description, work experience) were leveraged as the covariates.

[3]https://www.goodreads.com/
[4]https://www.microsoft.com/en-us/microsoft-teams

Table 1: ITE estimation performance (mean ± standard error). "CT", "GR" and "MS" stand for Contact, GoodReads and Microsoft Teams datasets, respectively.

Data	Method	Linear		Quadratic	
		$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
CT	LR	25.41 ±0.04	9.11 ±0.09	38.22 ±0.77	20.28 ±0.38
	CEVAE	22.88 ±1.07	8.29 ±0.69	35.28 ±0.75	18.22 ±0.76
	CFR	24.04 ±0.75	7.17 ±0.43	32.24 ±1.01	17.28 ±0.75
	Netdeconf	10.22 ±0.47	4.29 ±0.13	21.23 ±0.72	11.39 ±0.74
	GNN-HSIC	7.42 ±0.39	2.06 ±0.03	16.28 ±0.24	7.28 ±0.39
	GCN-HSIC	7.28 ±0.44	2.08 ±0.04	14.23 ±0.20	6.27 ±0.15
	HyperSCI	**3.45 ±0.27**	**1.39 ±0.03**	**9.20 ±0.09**	**2.24 ±0.07**
GR	LR	23.01 ±0.04	13.42 ±0.12	48.56 ±1.02	31.19 ±0.47
	CEVAE	22.69 ±0.03	12.49 ±0.06	45.21 ±3.10	29.22 ±0.44
	CFR	20.30 ±0.03	13.21 ±0.09	41.72 ±0.72	26.28 ±0.43
	Netdeconf	18.39 ±0.19	12.20 ±0.03	35.18 ±0.78	21.20 ±0.76
	GNN-HSIC	17.20 ±0.23	12.18 ±0.13	27.22 ±0.78	16.87 ±0.47
	GCN-HSIC	16.01 ±0.20	12.06 ±0.15	25.42 ±0.76	16.28 ±0.76
	HyperSCI	**15.68 ±0.21**	**11.81 ±0.15**	**19.23 ±0.44**	**13.33 ±0.27**
MS	LR	22.80 ±0.64	21.41± 0.74	414.17 ±3.94	192.80 ±2.97
	CEVAE	19.36 ±0.80	8.63 ±0.78	315.01 ±2.53	188.47 ±4.27
	CFR	25.23 ±0.01	18.28 ±0.02	392.56 ±4.33	189.75 ±4.80
	Netdeconf	11.11 ±0.01	9.22 ±0.03	241.02 ±2.32	147.29 ±1.04
	GNN-HSIC	9.38 ±0.44	6.91 ±0.38	114.28 ±3.62	81.21 ±2.53
	GCN-HSIC	8.27 ±0.41	6.60 ±0.48	109.57 ±3.85	77.75 ±3.93
	HyperSCI	**5.13 ±0.56**	**4.46 ±0.61**	**81.08 ±0.37**	**74.41 ±0.42**

4.2 Experiment Settings

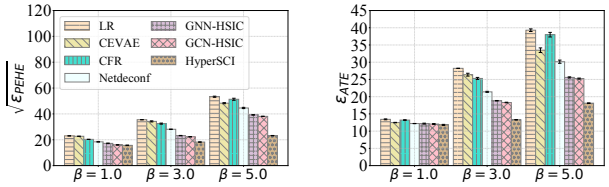
4.2.1 Metrics. We evaluate the performance of causal effect estimation through two standard metrics, including Rooted Precision in Estimation of Heterogeneous Effect ($\sqrt{\epsilon_{PEHE}}$) [18] and Mean Absolute Error (ϵ_{ATE}) [42]. These metrics can be defined as follows:

$$\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{n} \sum_{i \in [n]} (\tau_i - \widehat{\tau}_i)^2}, \ \epsilon_{ATE} = |\frac{1}{n} \sum_{i \in [n]} \tau_i - \frac{1}{n} \sum_{i \in [n]} \widehat{\tau}_i|. \tag{23}$$

Lower $\sqrt{\epsilon_{PEHE}}$ or ϵ_{ATE} indicates better causal effect estimations.

4.2.2 Baselines. To investigate the effectiveness of our framework, we compare it with multiple state-of-the-art ITE estimation baselines. These baselines can be divided into the following categories:

- **No graph.** We compare the estimation results with traditional methods which do not consider graph data and spillover effects. These methods include outcome regression which is implemented by linear regression (**LR**), counterfactual regression (**CFR** [34]), causal effect variational autoencoder (**CEVAE** [27]). By comparing the proposed framework to these methods, we evaluate the effectiveness of modeling interference for ITE estimation.
- **No spillover effect in ordinary graphs.** Although assuming no spillover effect exists, the network deconfounder (**Netdeconf**) [17] captures latent confounders for ITE estimation by utilizing the network structure among instances.
- **Spillover effect in ordinary graphs.** We compare our framework with other ITE estimation baselines which can handle the pairwise spillover effect on ordinary graphs: a node representation learning based method [28] estimates ITE under network interference, including two variants: (a) **GNN + HSIC**, which is



(a) $\sqrt{\epsilon_{PEHE}}$ (b) ϵ_{ATE}

Figure 4: Comparison of the performance of ITE estimation under different values of β in linear setting on GoodReads.

based on graph neural network [30] and Hilbert Schmidt independence criterion (HSIC) [16], and (b) **GCN + HSIC**, which is based on GCN [24].

To utilize the baselines which handle ordinary graphs, we project the original hypergraph $\mathcal{H}$ to an ordinary graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}^p\}$ by setting $(v_i, v_j) \in \mathcal{E}^p$ if v_i and v_j are contained in at least one common hyperedge in $\mathcal{H}$. By comparing **HyperSCI** to the above baselines, we are able to evaluate the benefits of modeling high-order interferences on the original hypergraph.

4.2.3 Setup. We randomly partition all datasets into 60%-20%-20% training/validation/test splits. All the results are averaged over ten repeated executions. Unless otherwise specified, we set the hyperparameters as $\alpha = 0.001$, $\beta = 1.0$, $\gamma = 1.0$, $\lambda = 0.01$, the dimension for confounder representation and interference representation both as 64. We use ReLU as the activation function, and use an Adam optimizer. By default, the interference modeling component contains one hypergraph convolutional layer.

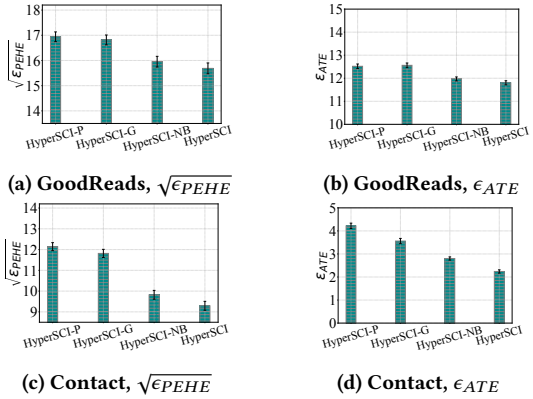
4.3 ITE Estimation Performance

We include the results for the ITE estimation task in Table 1. From this table we observe that the proposed framework outperforms all the baselines under both linear and quadratic outcome simulation settings. We attribute these results to the fact that **HyperSCI** utilizes the relational information in hypergraph to model the high-order interference, and thus mitigates the influence of the spillover effect on ITE estimation performance. Compared with other baselines, the methods which incorporate the pairwise network interference (**GCN-HSIC** and **GNN-HSIC**), as well as **Netdeconf** which utilizes the network structure for ITE estimation, perform better than those baselines which do not take advantage of the relational information (**LR**, **CEVAE**, **CFR**).

We also vary the hyperparameter (β) which controls the significance of hypergraph spillover effect in the outcome simulation and report the ITE estimation results in Fig. 4. As β increases, i.e., the outcome is more heavily influenced by interference, larger performance gains can be observed from the proposed framework (**HyperSCI**) against baselines. This observation further validates the effectiveness of our framework in modeling the interference for enhancing the performance of ITE estimation.

4.4 Ablation Study

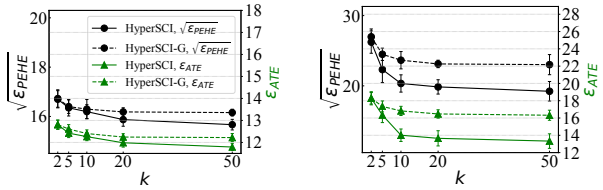
To investigate the effectiveness of different components in the proposed framework, we conduct ablation studies by considering the following variants: 1) we apply the proposed model **HyperSCI** on the projected graph (in a hypergraph structure) (denoted as **HyperSCI-P**); 2) we replace the hypergraph neural network module with a



(a) GoodReads, $\sqrt{\epsilon_{PEHE}}$ (b) GoodReads, ϵ_{ATE}

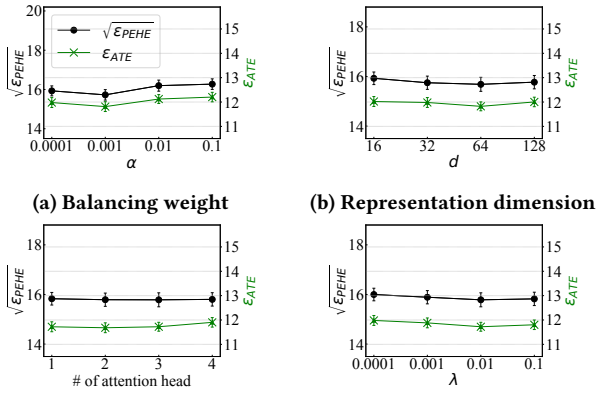
(c) Contact, $\sqrt{\epsilon_{PEHE}}$ (d) Contact, ϵ_{ATE}

Figure 5: Ablation studies of different variants of our framework HyperSCI. Results (mean and standard error) are reported under the linear setting but similar patterns can be found under the quadratic setting and on all datasets.



(a) GoodReads, Linear (b) GoodReads, Quadratic

Figure 6: ITE estimation performance of HyperSCI / HyperSCI-G on hypergraphs with hyperedge size no more than k.



(a) Balancing weight (b) Representation dimension

(c) # of attention head (d) Regularization weight

Figure 7: ITE estimation performance (mean and standard error) of the proposed framework HyperSCI under different parameters or model structures on GoodReads dataset.

graph neural network module with the same number of layers, and then apply it on the projected graph (in an original graph structure) (**HyperSCI-G**). Notice that although both evaluated on the projected graph, **HyperSCI-G** handles ordinary graphs with its graph neural network module, while **HyperSCI-P** handles hypergraphs with its hypergraph neural network module; 3) we remove the balancing techniques in the framework (**HyperSCI-NB**). The ITE

estimation results are reported in Fig. 5, where we notice significant performance gaps between **HyperSCI-P**/**HyperSCI-G** and **HyperSCI**, which imply the effectiveness of modeling the high-order relationships on hypergraphs. We also observe the ITE estimation performance degrades after removing the representation balancing modules, which indicates the effectiveness of the representation balancing techniques on mitigating the biases of ITE estimations.

4.5 A Closer Look at High-Order Interference

In addition to the overall ITE estimation performance, we take a closer look at high-order interference. We investigate how the proposed framework responds to hyperedges with difference sizes. More specifically, we remove the hyperedges with size larger than k, denote the modified hypergraph as $\mathcal{H}^{(k)}$, and vary the value of k. In Fig 6, we compare the ITE estimation performance of the proposed framework **HyperSCI** with its variant on the projected ordinary graph **HyperSCI-G**. We observe that: 1) When $k = 2$ (hyperedge size ≤ 2), the performance of HyperSCI-G is close to HyperSCI. Because when $k = 2$, graph convolution can be regarded as a special case of hypergraph convolution with small differences in the graph Laplacian matrix (as illustrated in [5]). Empirically this leads to a minor performance difference between HyperSCI-G and HyperSCI; 2) When k increases, the performance of ITE estimation from both methods are gradually improved, but such an improvement becomes less significant when k is larger. Besides, we notice **HyperSCI** consistently outperforms **HyperSCI-G** and such a difference becomes larger as k increases, indicating its efficacy on modeling high-order interference especially on large hyperedges.

4.6 Sensitivity Analysis

To evaluate the robustness of the proposed framework, we present the ITE estimation performance of **HyperSCI** under different settings of model hyper-parameters in Fig. 7. More specifically, we vary the value of the balancing weight from $\{0.0001, 0.001, 0.01, 0.1\}$, and vary the representation dimension from $\{16, 32, 64, 128\}$. We also vary the number of attention head from $\{1, 2, 3, 4\}$, then change the parameter of regularization weight from $\{0.0001, 0.001, 0.01, 0.1\}$. As can be observed, our framework is generally robust to different hyper-parameter settings, but proper fine-tuning of these hyper-parameters is still beneficial for the ITE estimation performance.

5 RELATED WORK

Causal studies under network interference. There have been many causal studies [3, 6, 9, 21, 25, 28, 37, 39, 49] which address the existence of network interference. These works mainly include the following categories: (i) *Random assignment strategy under interference* [3, 6, 12, 21, 39]. These works focus on experimental studies under interference (without SUTVA assumption). In some studies [23], strong interference is assumed to exist within each group while there is no interference across different groups; (ii) *Causal effect estimation on observational data with interference* [2, 28, 32, 38]. Different from the experimental studies which can design assignment strategy, another line of works (and also our work) assume interferences exist across individuals in the observational data. They relax the SUTVA assumption and define the potential outcome with a function that takes the instance covariates and treatment assignment of each individual and other interacted individuals as input. Among them, Rakesh et al. [32] propose a Linked Causal Variational Autoencoder (LCVA) framework to estimate the causal effect of a treatment on an outcome with the existence of interference between pairs of instances. Different from these works that focus on pairwise spillover effects, Ma et al. [28] consider the spillover effect in network structure, and propose a graph neural network (GNN) [24] based framework for causal effect estimation under network interference. However, these works are still limited in pairs of individuals or ordinary graphs and lack consideration of high-order interference. Another line of studies is bipartite causal inference [31, 53]. Traditionally, bipartite causal inference involves two types of units: interventional/outcome units. Interventional units are assigned with treatments, and outcomes are observed from outcome units. Although this setup is different from ours, considering that there is a node-hyperedge bipartite corresponding to each hypergraph (and ordinary graph), thus the two modeling approaches (bipartite and hypergraph) are conceptually similar. Nevertheless, we argue hypergraph is a more appropriate framing in many scenarios since: i) hypergraph does not require instantiating edges as additional nodes, or treating these two kinds of nodes differently, thus more computationally efficient; ii) hypergraph has the potential to be more convenient and efficient when generalizing to new hyperedges, while bipartite needs to generate both new nodes for the new hyperedges and their associated new edges.

Hypergraph algorithms and neural networks. To process hypergraph structures for downstream tasks, a line of works simplify the hypergraph structure by taking abstract representations of complicated multi-way interactions [7, 26, 43, 45, 51]. Other works directly tackle the original hypergraph structure [4, 8, 13, 19, 35, 44]. Recently, numerous works have studied on hypergraph neural networks [5, 44]. Feng el al. [13] propose hypergraph neural networks (HGNN) framework to encode high-order data correlation in a hypergraph structure. A hyperedge convolution operation is designed for representation learning. Bai et al. [5] introduce two end-to-end trainable operators hypergraph convolution and hypergraph attention to learn node representations in hypergraphs. Yadati et al. [44] develop a self-attention based hypergraph neural network Hyper-SAGNN, which is applicable to homogeneous or heterogeneous hypergraphs with variable hyperedge sizes. Jiang et al. [22] propose a dynamic hypergraph neural network (DHGNN) which can dynamically update hypergraph structure on each layer.

6 CONCLUSION

In this paper, we study an important research problem of individual treatment effect estimation with the existence of high-order interference on hypergraphs. We identify and analyze the influence of high-order interference in causal effect estimation. To address this problem, we propose a novel framework **HyperSCI**, which estimates the ITEs based on representation learning. More specifically, **HyperSCI** learns the representation of confounders, models the high-order interference with a hypergraph neural network module, then predicts the potential outcomes for each instance with the learned representations. We conduct extensive experiments to evaluate the proposed framework, where the results consistently validate the effectiveness of **HyperSCI** in ITE estimation under different interference scenarios.

REFERENCES

[1] Rohini Ahluwalia, H Rao Unnava, and Robert E Burnkrant. 2001. The moderating role of commitment on the spillover effect of marketing communications. *Journal of Marketing research* 38, 4 (2001), 458–470.

[2] David Arbour, Dan Garant, and David Jensen. 2016. Inferring network effects from observational data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 715–724.

[3] Peter M Aronow and Cyrus Samii. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* 11, 4 (2017), 1912–1947.

[4] Devanshu Arya and Marcel Worring. 2018. Exploiting relational information in social networks using geometric deep learning on hypergraphs. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 117–125.

[5] Song Bai, Feihu Zhang, and Philip HS Torr. 2021. Hypergraph convolution and hypergraph attention. *Pattern Recognition* 110 (2021), 107637.

[6] Guillaume Basse and Avi Feller. 2018. Analyzing two-stage experiments in the presence of interference. *J. Amer. Statist. Assoc.* 113, 521 (2018), 41–55.

[7] Austin R Benson, Rediet Abebe, Michael T Schaub, Ali Jadbabaie, and Jon Kleinberg. 2018. Simplicial closure and higher-order link prediction. *Proceedings of the National Academy of Sciences* 115, 48 (2018), E11221–E11230.

[8] Austin R Benson, Ravi Kumar, and Andrew Tomkins. 2018. Sequences of sets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1148–1157.

[9] Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. 2020. Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence*.

[10] Ulrik Brandes, Linton C Freeman, and Dorothea Wagner. 2013. *Social networks*.

[11] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be more with less: Hypergraph attention networks for inductive text classification. *arXiv preprint* (2020).

[12] Zahra Fatemi and Elena Zheleva. 2020. Minimizing interference and selection bias in network experiment design. In *International AAAI Conference on Web and Social Media*.

[13] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3558–3565.

[14] Ronald Aylmer Fisher. 1936. Design of experiments. *Br Med J* 1, 3923 (1936), 554–554.

[15] Cory E Goldstein, Charles Weijer, Jamie C Brehaut, Dean A Fergusson, Jeremy M Grimshaw, Austin R Horn, and Monica Taljaard. 2018. Ethical issues in pragmatic randomized controlled trials: a review of the recent literature identifies gaps in ethical argumentation. *BMC Medical Ethics* (2018).

[16] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[17] Ruocheng Guo, Jundong Li, and Huan Liu. 2020. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 232–240.

[18] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011).

[19] Jin Huang, Rui Zhang, and Jeffrey Xu Yu. 2015. Scalable hypergraph learning and processing. In *2015 IEEE International Conference on Data Mining*. IEEE, 775–780.

[20] Michael G Hudgens and M Elizabeth Halloran. 2008. Toward causal inference with interference. *J. Amer. Statist. Assoc.* 103, 482 (2008), 832–842.

[21] Kosuke Imai, Zhichao Jiang, and Anup Malani. 2020. Causal inference with interference and noncompliance in two-stage randomized experiments. *J. Amer. Statist. Assoc.* (2020), 1–13.

[22] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. 2019. Dynamic Hypergraph Neural Networks.. In *IJCAI*. 2635–2641.

[23] Brian Karrer, Liang Shi, Monica Bhole, Matt Goldman, Tyrone Palmer, Charlie Gelman, Mikael Konutgan, and Feng Sun. 2021. Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3106–3116.

[24] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint* (2016).

[25] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1168–1176.

[26] Dong Li, Zhiming Xu, Sheng Li, and Xin Sun. 2013. Link prediction in social networks based on hypergraph. In *Proceedings of the 22nd International Conference on World Wide Web*. 41–42.

[27] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*.

[28] Yunpu Ma and Volker Tresp. 2021. Causal Inference under Networked Interference and Intervention Policy Enhancement. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3700–3708.

[29] Rossana Mastrandrea, Julie Fournet, and Alain Barrat. 2015. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one* 10, 9 (2015), e0136497.

[30] Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4602–4609.

[31] Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. 2019. Variance reduction in bipartite experiments through correlation clustering. (2019).

[32] Vineeth Rakesh, Ruocheng Guo, Raha Moraffah, Nitin Agarwal, and Huan Liu. 2018. Linked causal variational autoencoder for inferring paired spillover effects. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1679–1682.

[33] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* 75, 371 (1980), 591–593.

[34] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*.

[35] Ankit Sharma, Jaideep Srivastava, and Abhishek Chandra. 2014. Predicting multi-actor collaborations using hypergraphs. *arXiv preprint* (2014).

[36] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* (1990), 465–472.

[37] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. 2012. On causal inference in the presence of interference. *Statistical methods in medical research* 21, 1 (2012), 55–75.

[38] Eric J Tchetgen Tchetgen, Isabel R Fulcher, and Ilya Shpitser. 2021. Auto-g-computation of causal effects on a network. *J. Amer. Statist. Assoc.* 116, 534 (2021), 833–844.

[39] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 329–337.

[40] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*. 86–94.

[41] Mengting Wan, Rishabh Misra, Ndapandula Nakashole, and Julian McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2605–2610.

[42] Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 1 (2005), 79–82.

[43] Ye Xu, Dan Rockmore, and Adam M Kleinbaum. 2013. Hyperlink prediction in hypernetworks using latent social features. In *International Conference on Discovery Science*. Springer, 324–339.

[44] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Anand Louis, and Partha Talukdar. 2018. Hypergcn: Hypergraph convolutional networks for semi-supervised classification. *arXiv preprint* 22 (2018).

[45] Naganand Yadati, Vikram Nitin, Madhav Nimishakavi, Prateek Yadav, Anand Louis, and Partha Talukdar. 2018. Link prediction in hypergraphs using graph convolutional networks. (2018).

[46] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems* 31 (2018).

[47] Serdar Yilmaz, Kingley E Haynes, and Mustafa Dinc. 2002. Geographic and network neighbors: Spillover effects of telecommunications infrastructure. *Journal of Regional Science* 42, 2 (2002), 339–360.

[48] Se-eun Yoon, Hyungseok Song, Kijung Shin, and Yung Yi. 2020. How Much and When Do We Need Higher-order Information in Hypergraphs? A Case Study on Hyperedge Prediction. In *Proceedings of The Web Conference 2020*. 2627–2633.

[49] Yuan Yuan, Kristen Altenburger, and Farshad Kooti. 2021. Causal Network Motifs: Identifying Heterogeneous Spillover Effects in A/B Tests. In *Proceedings of the Web Conference 2021*. 3359–3370.

[50] Chen Zeng, Yan Song, Dawei Cai, Peiying Hu, Huatai Cui, Jing Yang, and Hongxia Zhang. 2019. Exploration on the spatial spillover effect of infrastructure network on urbanization: A case study in Wuhan urban agglomeration. *Sustainable Cities and Society* 47 (2019), 101476.

[51] Muhan Zhang, Zhicheng Cui, Shali Jiang, and Yixin Chen. 2018. Beyond link prediction: Predicting hyperlinks in adjacency space. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[52] Ruochi Zhang, Yuesong Zou, and Jian Ma. 2019. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs. *arXiv preprint* (2019).

[53] Corwin M Zigler and Georgia Papadogeorgou. 2021. Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics* 36, 1 (2021), 109.

A NOTATION TABLE

We use Table 2 to list the most important notations used in this paper.

Table 2: Notation.

Notation	Definition
$\mathcal{H}$	hypergraph
$\mathcal{V}, \mathcal{E}$	the set of nodes/hyperedges
$\mathbf{H}$	hypergraph structure matrix
n	the number of nodes
m	the number of hyperedges
$\mathcal{N}_i$	the set of neighboring nodes for the i-th node
$\mathcal{N}_e$	the set of nodes on the hyperedge e
$\mathcal{E}_i$	the set of hyperedges which contains the i-th node
$\mathbf{X}, \mathbf{x}_i$	features of all nodes/the i-th node
$\mathbf{T}, t_i$	treatment assignment of all nodes/the i-th node
$\mathbf{Y}, y_i$	observed outcome of all nodes/the i-th node
y_i^1, y_i^0	potential outcomes of the i-th node
$\Phi_Y(\cdot)$	potential outcome function
$\tau_i, \widehat{\tau}_i$	true/predicted ITE for the i-th node
$y_i, \widehat{y}_i$	true/predicted outcome for the i-th node
$(\cdot)_{-i}$	variables for all the nodes except the i-th node
δ_i	the spillover effect of the i-th node
$\texttt{SMR}(\cdot)$	summary function
$\mathbf{O}, \mathbf{o}_i$	the environment information of the i-th node
$\mathbf{Z}, \mathbf{z}_i$	confounder representations of all nodes/the i-th node
$\Psi(\cdot)$	interference representation learner
$f_1(\cdot), f_0(\cdot)$	potential outcome prediction functions
$\mathbf{P}, \mathbf{p}_i$	interference representations of all nodes/the i-th node
$r(i)$	the ratio of the treatment assignment of the i-node in its neighborhood

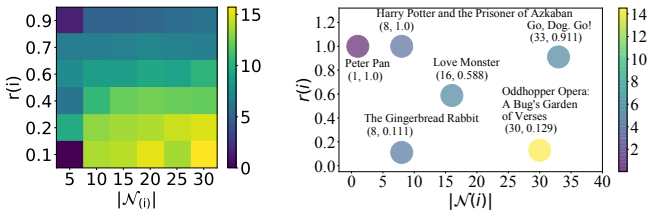
B MORE EXPERIMENTAL RESULTS

B.1 ITE Estimation Performance under Different Settings on All the Datasets

In this section, we show the ITE estimation performance under different settings (including the linear and quadratic settings with β among $\{1.0, 3.0, 5.0\}$) on all the datasets in Fig. 9. We can observe that the proposed **HyperSCI** consistently outperforms the baselines under different settings on all the datasets. The superiority of our framework against baselines becomes more obvious when β is larger (i.e., the interference is stronger), because our framework can better handle the interference in the hypergraph.

B.2 Case Studies

We further conduct case studies to investigate how the proposed method acts on individuals in responding to their neighboring nodes (i.e., the size of one's neighborhood and the homophily of treatment assignments within one's neighborhood). The neighborhood of i is defined as the set of nodes which are connected with i via any hyperedges, i.e., $\mathcal{N}_i = \bigcup_{e \in \mathcal{E}_i} \{j \in \mathcal{N}_e\}$. The homophily of treatment assignment is defined as the ratio of neighboring nodes which share the same treatment assignment as oneself, i.e., $r(i) = \frac{\sum_{j \in \mathcal{N}_i} \mathbf{1}(t_j = t_i)}{|\mathcal{N}_i|}$. In Fig. 8a, we show the difference between the ITE estimation results made with the original hypergraph and with the projected graph, w.r.t. $\mathcal{N}_i$ and $r(i)$. Overall, we see larger divergences on individuals with a larger neighborhood size but less agreement with their neighbors in terms of treatment assignments. In Fig. 8b, we further showcase the insights by presenting several representative children books on the GoodReads dataset. For example, the author of "Peter Pan" had not published many works but these books all received good rating scores, leading to a "consistent" reputation of the author. Therefore, the outcome of the book "Peter Pan" is less impacted by the high-order interference among its neighbors. On the other hand, the high rating score of the book "Oddhopper Opera" differs from most of its neighbors, leading to a mixed reputation of the author. In this case, the potential outcome is more likely to be affected by the high-order interference on the hypergraph.



(a) Heatmap **(b) Case studies**

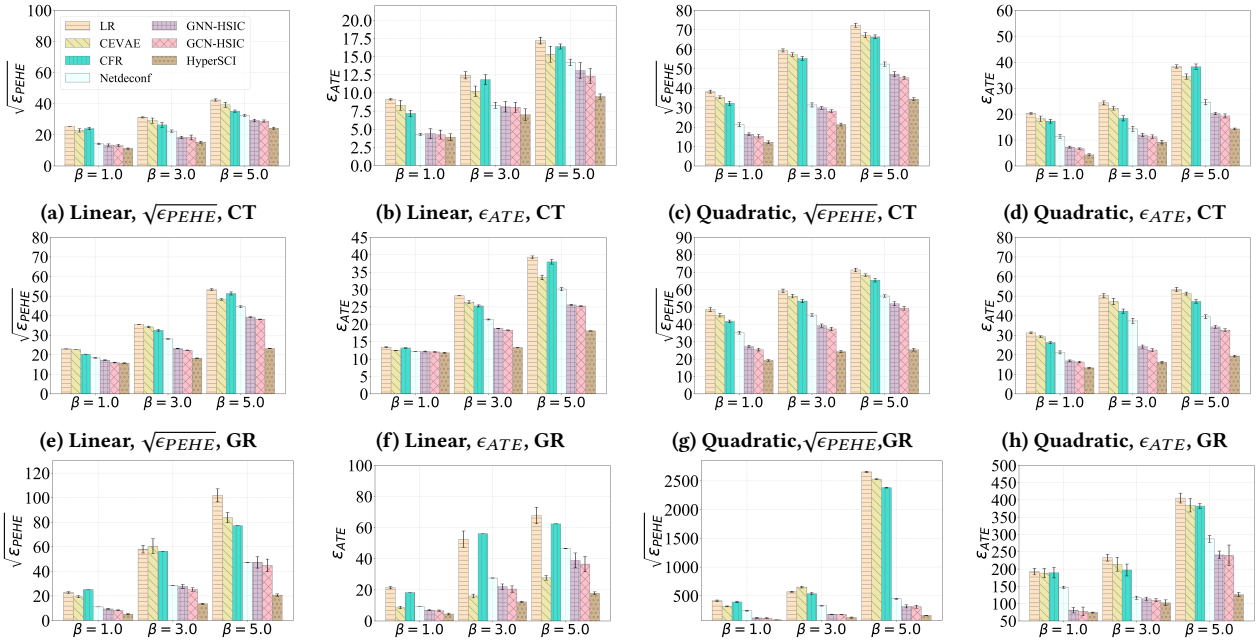
Figure 8: (a) Heatmap: the difference between ITE estimations with hypergraph and with projected ordinary graph on GoodReads. Nodes are divided into 6×6 grids w.r.t. their number of neighbors $|\mathcal{N}_i|$ and the homophily of treatment assignment $r(i)$. (b) Case studies of representative books.

C DETAILS OF EXPERIMENTAL SETTINGS

All the experiments are conducted under the following environment:

- Operating system: Ubuntu 18.04
- GPU memory: 16GB
- Pytorch 1.9.0, Cuda ToolKit 11.1, cuDNN 8.0.5

Baseline parameter settings. For the baselines CEVAE, CFR, Netdeconf, GNN-HSIC, and GCN-HSIC, we set the representation dimension as 32, 32, 100, 64, respectively. The numbers of training epochs for these baselines are set as 500. The number of samples in CEVAE in training is set as 5 by default.



(a) Linear, $\sqrt{\epsilon_{PEHE}}$, CT (b) Linear, ϵ_{ATE}, CT (c) Quadratic, $\sqrt{\epsilon_{PEHE}}$, CT (d) Quadratic, ϵ_{ATE}, CT

(e) Linear, $\sqrt{\epsilon_{PEHE}}$, GR (f) Linear, ϵ_{ATE}, GR (g) Quadratic, $\sqrt{\epsilon_{PEHE}}$, GR (h) Quadratic, ϵ_{ATE}, GR

(i) Linear, $\sqrt{\epsilon_{PEHE}}$, MS (j) Linear, ϵ_{ATE}, MS (k) Quadratic, $\sqrt{\epsilon_{PEHE}}$, MS (l) Quadratic, ϵ_{ATE}, MS

Figure 9: Comparison of the performance of ITE estimation under different settings. "CT", "GR" and "MS" stand for Contact, GoodReads and Microsoft Teams datasets, respectively.