Robust Event Forecasting with Spatiotemporal Confounder Learning

Songgaojun Deng
sdeng4@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Huzefa Rangwala
rangwala@cs.gmu.edu
George Mason University
Fairfax, VA, USA

Yue Ning
yue.ning@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

ABSTRACT

Data-driven societal event forecasting methods exploit relevant historical information to predict future events. These methods rely on historical labeled data and cannot accurately predict events when data are limited or of poor quality. Studying causal effects between events goes beyond correlation analysis and can contribute to a more robust prediction of events. However, incorporating causality analysis in data-driven event forecasting is challenging due to several factors: (i) Events occur in a complex and dynamic social environment. Many unobserved variables, i.e., hidden confounders, affect both potential causes and outcomes. (ii) Given spatiotemporal non-independent and identically distributed (non-IID) data, modeling hidden confounders for accurate causal effect estimation is not trivial. In this work, we introduce a deep learning framework that integrates causal effect estimation into event forecasting. We first study the problem of Individual Treatment Effect (ITE) estimation from observational event data with spatiotemporal attributes and present a novel causal inference model to estimate ITEs. We then incorporate the learned event-related causal information into event prediction as prior knowledge. Two robust learning modules, including a feature reweighting module and an approximate constraint loss, are introduced to enable prior knowledge injection. We evaluate the proposed causal inference model on real-world event datasets and validate the effectiveness of proposed robust learning modules in event prediction by feeding learned causal information into different deep learning methods. Experimental results demonstrate the strengths of the proposed causal inference model for ITE estimation in societal events and showcase the beneficial properties of robust learning modules in societal event forecasting.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Mathematics of computing** → *Causal networks*; • **Computing methodologies** → *Neural networks*.

KEYWORDS

Event Forecasting; Representation Learning; Causal Inference

1 INTRODUCTION

Predicting societal events such as disease outbreaks, organized crime, and civil unrest movements from social media streams and news media is of great significance for decision-making and resource allocation. Previous approaches mainly focus on learning from historical data to improve the prediction accuracy of events [48]. Recently, to enhance model explainability, many approaches identify salient features or supporting evidence, such as precursor documents [28], relationships represented as graphs [10], and major actors participating in the events [11]. However, these works explain the occurrence of events based on correlation-based indicators.

Attempts to study causality in event analysis and prediction have focused on extracting pairs of causal events from unstructured text [31], or using human-defined causally related historical events to predict events of interest [32]. Causal effect learning has shown advantages in improving predictions in various machine learning problems, such as recommender systems [5], disease diagnosis prediction [24], and computer vision tasks [7]. This suggests the potential of causal effect learning for better prediction of societal events. Leveraging causal effects can presumably provide new insights into causal-level interpretation and improve the robustness of event prediction, e.g., less susceptible to noise in data. In this study, we explore societal event forecasting methods with the help of causal effect learning.

Traditionally, learning causal effects (aka treatment effects) from observational data involves estimating the causal effects of a treatment variable (e.g., medication) on an outcome variable (e.g., recovery) given observable covariates (e.g., gender). In practice, there are also unobserved covariates, i.e., *hidden confounders*, that affect both treatment and outcome variables. For instance, consider a study to evaluate the effectiveness of a medication. Gender as a covariate affects whether a patient chooses to take the medication and the corresponding outcome. The patient's living habits can be hidden confounders that affect both the patient's medication and outcome. Exploring hidden confounders allows for more accurate estimations of treatment effects [17, 25, 26].

In this work, we formulate the problem of estimating treatment effects in the context of societal events. Societal events can be classified into different types. Given a time window, we look at multiple types of events (e.g., "appeal", "investigation") at a location and define treatment variables to be the detection of increased counts

of these events compared to the previous time window. If the sudden and frequent occurrence of such events triggers some event of interest, the implied causal effect can be used to guide and interpret event predictions. We define the outcome as the occurrence of an event of interest (e.g., "protest") at a future time. Both treatment and outcome variables can be affected by hidden social factors (i.e., hidden confounders) that are difficult to explicitly capture due to complex dependencies. Intuitively, exploring hidden confounders can allow us to estimate causal effects more accurately. To this end, we formulate our main research question as: *can we build a robust event predictive model by incorporating treatment effect estimation with hidden confounder learning?* There are some challenges in solving this problem:

- Societal events have geographical characteristics and exhibit a high degree of temporal dependency [11, 28]. Modeling spatiotemporal information requires an in-depth investigation of the dynamic spatial dependencies of societal events. However, few studies have focused on modeling spatiotemporal dependencies in causal effect learning, which poses a challenge for learning causal effects from societal events.
- Events occur in a complex and evolving social environment. Many unknown factors increase the difficulty of accurately estimating causal effects of events. Moreover, events are often caused by a variety of factors rather than a single determinant. Utilizing causal effects to assist in event prediction is a new challenge.

We address the above challenges by first introducing the task of Individual Treatment Effect (ITE) estimation from societal events. ITE is defined as the expected difference between the treated outcome and control outcome, where the outcome is the occurrence of a future event (e.g., *protest*) at a specific place and time, and the treatment is a change in some event (e.g., *appeal*) in the past. We consider multiple treatments (e.g., *appeal*, *investigation*, etc.) with the motivation that the underlying causes of societal events are often complex. We model the spatiotemporal dependencies in learning the representations of hidden confounders to estimate ITEs. We then present an approach to inject the learned causal information into a data-driven predictive model to improve its predictive power. Our contributions are summarized as follows:

- We introduce a novel causal inference model for ITE estimation, which learns the representation of hidden confounders by capturing spatiotemporal dependencies of events in different locations.
- We propose two robust learning modules for event prediction that take as prior knowledge the information learned from the causal inference model. Incorporating such modules can enable event prediction models to be more robust to data noise.

We evaluate the proposed method against other state-of-the-art methods on several real-world event datasets. Through extensive experiments, we demonstrate the strengths of the proposed method in treatment effect learning and robust event prediction.

2 RELATED WORK

2.1 Event Prediction

Event prediction focuses on forecasting events that have not yet happened based on various social indicators, such as event occurrence rates and news reports. Related research has been conducted

in various fields and applications, such as disease outbreak simulation [1], crime prediction [41] and civil unrest prediction [10, 28]. Machine learning models such as linear regression [4] and random forests [19] were investigated to predict events of interest. Time-series methods such as autoregressive models were studied to capture the temporal evolution of event-related indicators [1]. With the increased availability of various data, more sophisticated features have been shown effective in predicting societal events such as topic-related keywords [48], document embedding [28], word graphs [10] and knowledge graphs [11, 12]. More advanced machine learning and deep learning-based models have emerged, such as multi-instance learning [28], multi-task learning [49] and graph neural networks [10, 11]. Given the spatiotemporal dependencies of events, some existing research work studied spatiotemporal correlations in event prediction [15, 40, 47]. However, few studies explored the causality in event prediction. Our proposed model incorporates causal effect learning in a spatiotemporal event prediction framework. This gives us the benefit of discovering the effects of different potential causes on predicting future events.

2.2 Individual Treatment Effect Estimation

Individual treatment effect (ITE) estimation refers to estimating the causal effect of a treatment variable on its outcome. A wealth of observational data facilitates treatment effect estimation in many fields, such as health care [2] recommender systems [5], and computer vision [7]. Existing methods for ITE estimation includes regression and tree based model [8], counterfactual inference [18], and representation learning [35]. These approaches rely on the Ignorability assumption [33], which is often untenable in real-world studies. A deep latent variable model, CEVAE [25] learns representations of confounders through variational inference. Recent work relaxed the Ignorability assumption and studied ITE estimation from observational data with an auxiliary network in a static [17] or dynamic environment [26]. In addition to the traditional causal effect estimation, a new study of causal inference, including multiple treatments and a single outcome, has emerged, namely, Multiple Causal Inference. Researchers have shown that compared with traditional causal inference, it requires weaker assumptions [42]. ITE estimation would considerably benefit decision-making as it can provide potential outcomes with different treatment options. Our work introduces ITE estimation to societal event studies and exploits event-related causal information for event forecasting.

2.3 Knowledge Guided Machine Learning

Purely data-driven approaches might lead to unsatisfactory results when limited data are available to train well-performing and sufficiently generalized models. Such models may also break natural laws or other guidelines [37]. These problems have led to an increasing amount of research that focuses on incorporating additional prior knowledge into the learning process to improve machine learning models. For example, logic rules [14] or algebraic equations [27] have been added as constraints to loss functions. Knowledge graphs are adopted to enhance neural networks with information about relations between instances [3]. The growth of this research suggests that the combination of data and knowledge-driven approaches is

becoming relevant and showing benefits in a growing number of areas. Existing work has typically focused on pre-existing knowledge obtained by human experts. However, such approaches fail when prior knowledge is not available, e.g., for societal events. Some researchers explored causal knowledge-guided methods in health prediction [24] and image-to-video adaptation [7]. In this work, we study causal effects between societal events and use the learned causal information as prior knowledge for event prediction.

3 PROBLEM FORMULATION

The objective of this study is two-fold: (1) given multiple pre-defined treatment events (e.g., appeal, investigation, etc.), estimate their causal effect on a target event (i.e., protest) individually; (2) predict the probability of the target event occurring in the future with the help of estimated causal information. In the following, we will introduce some important concepts, definitions and assumptions.

3.1 Observational Event Data

In this work, we focus on modeling the occurrence of one type of societal event (i.e., "protest") by exploring the possible effects it might receive from other types of events (e.g., "appeals" and "investigation"). A total of E categories of societal events are studied. These events happen at different locations and times. We use M, T to denote the sets of locations and timestamps of interest, respectively. The observational event data can be denoted as $\mathcal{D} = \left\{ \{X_i^{\leq t}, c_i^{\leq t}, y_i^{t+\delta}, \}_{i=1}^M, A \right\}_{t \in T}$, where $X_i^{\leq t}, c_i^{\leq t}, y_i^{t+\delta}$ denote the pre-treatment covariates/features, observed treatments, and outcome, respectively. $A \in \mathbb{R}^{M \times M}$ represents the connectivity of M locations, where each element can denote a fixed geographic distance or the degree of influence of events between locations.

Covariates: We define the covariates $X_i^{\leq t} = (x_i^{t-\Delta+1}, ..., x_i^t) \in \mathbb{R}^{E \times \Delta}$ to be the historical events at location i with size Δ up to time t. $x_i^t \in \mathbb{R}^E$ is a vector representing the frequencies of E types of events that occurred at location i at time t.

Treatments: The treatments $c_i^{\leq t} \in \{0, 1\}^E$ can be represented by a binary vector with dimension E where each element indicates the occurrence state of a type of event (e.g., appeal). Specifically, the j-th element $c_{i(j)}^{\leq t} = 1$ indicates a notable (i.e., 50%) increase of the j-th event type at window $[t - \Delta + 1, t]$ from the previous window $[t - 2\Delta + 1, t - \Delta]$. A value of 1 means getting treated and 0 means getting controlled. For convenience, we refer to each element in the treatment vector as a treatment event.[1]

Observed Outcome: The observed/factual outcome $y_i^{t+\delta} \in \{0, 1\}$ is a binary variable denoting if a target event (i.e., protest) occurs at location i in the future $(t + \delta)$. $\delta \geq 1$ is the lead time indicating the number of timestamps in advance for a prediction.

3.2 Individual Treatment Effects Learning

We first define potential outcomes in observational event data following well-studied causal inference frameworks [34]. We ignore the location subscript i for simplicity unless otherwise stated.

Potential Outcomes: In general, the potential outcome $Y(C)$ denotes what the outcome an instance would receive, if the instance

---

[1]Our setup differs from multiple causal inference [42], which estimates the potential outcome of a combination of multiple treatments. We are more interested in studying the potential outcome of each element in the treatment vector.
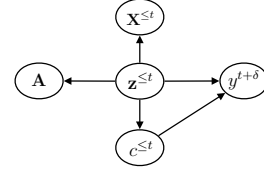

Figure 1: Causal graph defined for ITE estimation on observational event data. Edges represent causal relations. $X^{\leq t}, c^{\leq t}, z^{\leq t}, y^{t+\delta}$ denote covariates, assignment for a treatment event, hidden confounders before time t, and outcome at time $t + \delta$, respectively. A denotes the connectivity of locations.
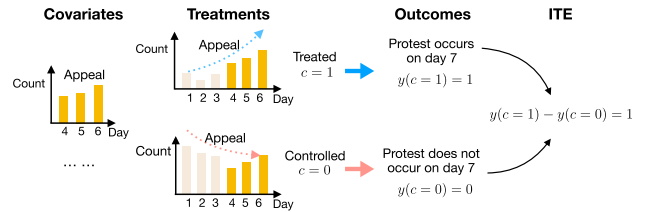

Figure 2: An example of ITE estimation on observational event data. X-axis indicates time and bars indicate the counts of events in a location. Assuming *appeal* is the treatment variable, the treated and controlled scenarios denote whether the amount of appeal events in the current window (on average) is greater or less than the previous window. Orange bars indicate counts of appeal events. Outcomes denote whether the protest will occur at a future time given each treatment assignment. ITE is estimated from the two potential outcomes.

were to take treatment C. A potential outcome $Y(C)$ is distinct from the observed/factual outcome Y in that not all potential outcomes are observed in the real world. In our problem, there are two potential outcomes for each treatment event. Given a location at time $t + \delta$ and the j-th treatment event, we denote by $y_{(j)}^{t+\delta}(1)$ the potential outcome (i.e., occurrence of protest) if the j-th treatment event is getting treated, i.e., $c_{(j)}^{\leq t} = 1$. Similarly, we denote by $y_{(j)}^{t+\delta}(0)$ the potential outcome we would observe if the treatment event is under control, i.e., $c_{(j)}^{\leq t} = 0$.

The factual outcome is $y^{t+\delta}$ when the location has already received the treatment assignment before time t. The counterfactual outcome is defined if the location obtains the opposite treatment assignment. In the observational study, only the factual outcomes are available, while the counterfactual outcomes can never be observed.

The Individual Treatment Effect (ITE) is the difference between two potential outcomes of an instance, examining whether the treatment affects the outcome of the instance. In observational event data, for the j-th treatment event, we formulate the ITE for the location at time $t + \delta$ in the form of the Conditional Average Treatment Effect (CATE) [26, 35]:

$$\tau_{(j)}^{t+\delta} = \mathbb{E}[y_{(j)}^{t+\delta}(1) - y_{(j)}^{t+\delta}(0)|X^{\leq t}, A]. \quad (1)$$

We provide a toy example to illustrate ITE estimation on observational event data, as shown in Fig. 2.

In this study, we aim to estimate ITEs and then use them for event prediction. The challenge of ITE estimation lies in how to estimate the missing counterfactual outcome. Our estimation of ITE is built upon some essential assumptions. For simplicity and readability, we omit the subscripts for the location i and the treatment event j and use $c^{\leq t}$ to represent the j-th treatment event.

ASSUMPTION 1. **No Interference**. *Assuming that one instance is defined as a location at a time in observational event data, the potential outcome on one instance should be unaffected by the particular assignment of treatments on other instances.*

ASSUMPTION 2. **Consistency**. *The potential outcome of treatment $c^{\leq t}$ equals to the observed outcome if the actual treatment received $c^{\leq t}$, i.e., $y^{t+\delta}(c^{\leq t}) = y^{t+\delta}$.*

ASSUMPTION 3. **Positivity**. *If the probability $\mathbb{P}(X^{\leq t}, A) \neq 0$, then the probability to receive treatment assignment 0 or 1 is positive, i.e., $0 < \mathbb{P}(c^{\leq t} = 1 | X^{\leq t}, A) < 1$, $c^{\leq t} \in \{0, 1\}$.*

The Positivity assumption indicates that before time t, each treatment assignment has a non-zero probability of being given to a location. This assumption is testable in practice. In addition to these assumptions, most existing work [25, 35, 39] relies on the Ignorability assumption, which assumes that all confounding variables are observed and reliably measured by a set of features for each instance, i.e., hidden confounders do not exist.

DEFINITION 1. **Ignorability Assumption**. *Given pre-treatment covariates $X^{\leq t}$, the outcome variables are independent of its treatment assignment, $y^{t+\delta}(0), y^{t+\delta}(1) \perp\!\!\!\perp c^{\leq t} | X^{\leq t}$.*

However, this assumption is untenable in societal event studies due to the complex environment in which societal events occur. We relax this assumption by introducing the existence of hidden confounders [17]. Note that hidden confounders are unobserved in observational event data but will be learned in our approach through a spatiotemporal model. We define a causal graph, as shown in Fig. 1. The hidden confounders $z^{\leq t}$ causally affect the treatment and outcome.[2] The potential outcomes are independent of the observed treatment, given the hidden confounders: $y^{t+\delta}(0), y^{t+\delta}(1) \perp\!\!\!\perp c^{\leq t} | z^{\leq t}$. In addition, we assume the features $X^{\leq t}$ and the connectivity of locations A are proxy variables for hidden confounders $z^{\leq t}$. Unobservable hidden confounders can be measured with $X^{\leq t}$ and A. Based on the temporal and spatial characteristics of our observational event data. We introduce the following assumption [26]:

ASSUMPTION 4. **Spatiotemporal Dependencies in Hidden Confounders**. *In observational event data, hidden confounders capture spatial information among locations, reflected by A, and show temporal dependencies of events across multiple historical steps (i.e., Δ).*

Note that this assumption does not contradict the *No Interference* assumption. We focus on the scenario in which spatiotemporal information can be exploited to control confounding bias.

---

[2]For the j-th treatment event, the hidden confounders can be written as $z_{(j)}^{\leq t}$.

3.3 Event Prediction

DEFINITION 2. **Traditional Event Prediction**. *Learn a classifier that predicts the probability of the target event occurring at a location at time $t + \delta$ based on available data: $\mathbb{P}(y^{t+\delta} | X^{\leq t}, A)$.*

Instead of learning a mapping function from input features to event labels, we are interested in estimating treatment effects under different treatment events individually and exploiting such causal information to enhance event prediction.

DEFINITION 3. **Event Prediction with Causal Knowledge**. *Build an event forecaster using available data with causal information as prior knowledge: $\mathbb{P}(y^{t+\delta} | X^{\leq t}, A, C(X^{\leq t}, c^{\leq t}, A))$, where C is the trained causal inference model that takes features, multiple treatments and the connectivity information of locations as input and outputs potential outcomes.*

The multiple treatment setting (i.e., $c^{\leq t}$) aims to produce informative causal knowledge to assist event prediction. We will discuss the proposed method of event prediction with causal knowledge in the following sections.

4 METHODOLOGY

We propose a novel framework **CAPE**, which incorporates **ca**usal inference into the **p**rediction of future **e**vent occurrences in a spatiotemporal environment. In our framework, ITEs with different treatment events are jointly modeled in a spatiotemporal causal inference model. It will contribute to the final event prediction by feeding the causal output (e.g., potential outcomes) to a non-causal data-driven prediction model. The overall framework, as illustrated by Fig. 3, consists of two parts: (1) causal inference and (2) event prediction. The causal inference component is designed to estimate ITEs, including two essential modules: hidden confounder learning and potential outcome prediction. For each treatment event, it learns the representation of hidden confounders by capturing spatiotemporal dependencies and outputs the potential outcomes under different treatment assignments. The event prediction part comprises two robust learning modules, a feature reweighting module and an approximation constraint loss. They take the causal information learned from the causal inference model as prior knowledge to assist the training of a data-driven event prediction model. Next, we will elaborate on these components.

4.1 Causal Inference

4.1.1 *Hidden Confounder Learning*. Hidden confounders are common in real-world observational data [30]. Assuming spatiotemporal dependencies exist in hidden confounders, we introduce a novel and effective network that models spatial and temporal information for each location at each time step. It consists of several temporal feature learning layers and spatial feature learning layers. Our network is based on the success of previous work [9, 29]. It is designed to be adaptable to a multi-task setting to learn hidden confounders of multiple treatments.

Temporal Feature Learning. Dilated casual convolution networks [46] handle long sequences in a non-recursive manner, which facilitates parallel computation and alleviates the gradient explosion problem. Gating mechanisms have shown benefits to control information flow through layers for convolution networks [9, 29, 43].
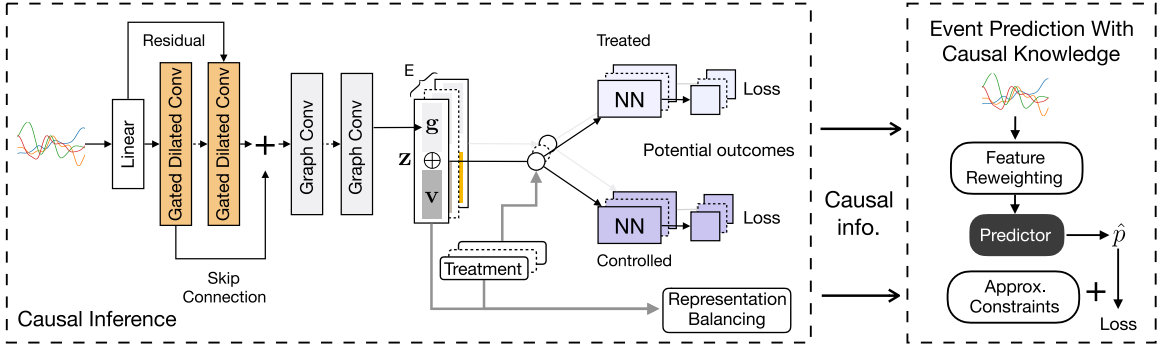
Figure 3: The overall architecture of the proposed framework. The proposed causal inference method estimates ITEs in the presence of multiple treatment events. The representation of hidden confounders is learned from a spatiotemporal model. Then, the representation of hidden confounders corresponding to each treatment event is fed into two neural networks (NNs) for ITE estimation. Next, the framework takes the output of the causal inference model as prior knowledge to forecast events.

We employ the dilated causal convolution with a gating mechanism in temporal feature learning to capture a location's temporal dependencies. For a location before time t, the multivariate time series of historical event occurrences is a matrix $\mathbf{X}^{\leq t} \in \mathbb{R}^{E \times \Delta}$, where each row indicates the frequency sequence of one type of event in the historical window with size Δ. We use a linear transformation to map the event frequency matrix into a latent space, i.e., $\mathbf{X}'^{\leq t} \in \mathbb{R}^{d_s \times \Delta}$. d_s indicates the feature dimension in the latent space. Then, we apply the dilated convolution to the sequence. For simplicity, we use $\mathbf{r} \in \mathbb{R}^{\Delta}$ to denote a row in the matrix $\mathbf{X}'^{\leq t}$. Formally, for the 1-D sequence input $\mathbf{r}$ and a filter $\boldsymbol{\Gamma} \in \mathbb{R}^K$, the dilated causal convolution operation on element s of the sequence is defined as:

$$(\mathbf{r} *_d \boldsymbol{\Gamma})(s) = \sum_{k=0}^{K-1} \boldsymbol{\Gamma}(k)\mathbf{r}(s - d \cdot k), \tag{2}$$

where $*_d$ is a d-dilated convolution. K is the filter size. $(s), (k), (s - d \cdot k)$ are indices of vectors. $(\mathbf{r} *_d \boldsymbol{\Gamma})$ is the output vector.

We further incorporate a gated dilated convolutional layer which consists of two parallel dilated convolution layers:

$$\mathbf{h} = \tanh(\mathbf{r} *_d \boldsymbol{\Gamma}_1) \odot \sigma(\mathbf{r} *_d \boldsymbol{\Gamma}_2), \tag{3}$$

where $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$ are filters for dilated convolutional layers, and $\odot$ is the Hadamard product. $\tanh(\cdot)$ is to regularize the features. $\sigma(\cdot)$ is the sigmoid function that determines the ratio of information passed to the next layer. Specifically, we stack multiple gated dilated convolutional layers (Eq. 3) with increasing dilation factors (e.g., $d = 1, 2, 4$). Residual and skip connections are applied to avoid the vanishing gradient problem [29, 43]. To this end, the temporal dependencies are captured, and we use $\mathbf{H}^t \in \mathbb{R}^{M \times d_s}$ to denote the learned temporal features for M locations at time t.

Spatial Feature Learning. Graph convolution is a powerful operation to learn representations of nodes given the graph structure. To capture the spatial dependencies, we adopt the graph convolutional network (GCN) [22] to learn the spatial influence from locations by treating each location as a node in graph:

$$\mathbf{G}^t = \text{ReLU}(\mathbf{A}'\mathbf{H}^t\mathbf{W}^g), \tag{4}$$

where $\mathbf{W}^g$ is the weight matrix for a GCN layer. $\mathbf{G}^t$ denotes the spatiotemporal feature matrix referring to all locations, where each row $\mathbf{g}^t$ captures the historical information of a specific location as well as the neighboring locations. $\mathbf{A}'$ is a learnable adjacency

matrix. The geographical adjacency matrix of locations usually cannot represent the connectivity of locations in the context of societal event forecasting. Therefore, we adopt the self-adaptive adjacency matrix [43], which does not require any prior knowledge and is learned through training. We randomly initialize two node embedding matrices with learnable parameters $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{M \times d_a}$. The self-adaptive adjacency matrix is defined as:

$$\mathbf{A}' = \text{Softmax}(\text{ReLU}(\mathbf{E}_1\mathbf{E}_2^T)), \tag{5}$$

where the ReLU activation function eliminates weak connections and the Softmax applies normalization.

Hidden Confounder Representation Learning. To learn the representation of hidden confounders, we utilize the spatiotemporal feature and a learnable embedding specific to each treatment event (i.e., $\mathbf{v}_{(j)}, 1 \leq j \leq E$). It is worth pointing out that in our framework, we include multiple treatment events and expect to estimate the ITE corresponding to each treatment event. The treatment-specific embedding aims to capture latent information of each treatment event and distinguish the hidden confounder representations learned for each treatment effect learning task. Similar ideas of task embeddings are studied in prior work [38]. Given a location and a time t, the representation of hidden confounders for the j-th treatment is:

$$\mathbf{z}_{(j)}^{\leq t} = \mathbf{g}^t \oplus \mathbf{v}_{(j)}, \tag{6}$$

where $\oplus$ is concatenation.

*4.1.2 **Potential Outcome Prediction.*** Using the above components, we obtain the representation of hidden confounders $\mathbf{z}_{(j)}^{\leq t}$. Following the predefined causal graph in Fig. 1, the learned hidden confounders can be used to estimate potential outcomes. We use two networks that output two potential outcomes of the j-th treatment event, respectively:

$$\hat{y}_{(j)}^{t+\delta}(1) = \Phi_{(j)}^1(\mathbf{z}_{(j)}^{\leq t}), \quad \hat{y}_{(j)}^{t+\delta}(0) = \Phi_{(j)}^0(\mathbf{z}_{(j)}^{\leq t}), \tag{7}$$

where $\hat{y}_{(j)}^{t+\delta}(1/0)$ denotes the inferred potential outcome when the j-th treatment event is getting treated or controlled. $\Phi_{(j)}^1(\cdot), \Phi_{(j)}^0(\cdot)$ are parameterized by deep neural networks with a sigmoid function at the last layer. The networks are trained end-to-end, and one can estimate the potential outcomes under multiple treatment events.

*4.1.3 **Objective Function.*** We use the binary cross-entropy loss as the objective factual loss for predicting potential outcomes. When

only the j-th treatment event is considered (i.e., the general case for treatment effect learning [17, 26, 35]), the factual loss is:

$$\mathcal{L}^{\text{fact}}_{(j)} = - \sum_{t \in T} \sum_{i \in M} y_i^{t+\delta} \log \hat{y}_{i(j)}^{t+\delta} + (1 - y_i^{t+\delta}) \log \left(1 - \hat{y}_{i(j)}^{t+\delta}\right), \quad (8)$$

where $y_i^{t+\delta}$ is the observed outcome for location i at time $t + \delta$. $\hat{y}_{i(j)}^{t+\delta} = \hat{y}_{i(j)}^{t+\delta}(c_{i(j)}^{\leq t})$ is the predicted outcome given the observed treatment $c_{i(j)}^{\leq t} \in \{0, 1\}$. Since our model predicts potential outcomes for multiple treatment events, we express the total factual loss as follows:

$$\mathcal{L}^{\text{fact}} = \sum_{1 \leq j \leq E} \mathcal{L}^{\text{fact}}_{(j)} + \eta \cdot \Omega(\Theta), \quad (9)$$

where $\Omega(\Theta)$ stands for the ℓ_2-norm regularization for all training parameters and η is the weight for scaling the regularization term.
Representation Balancing. Studies have proved that balancing representations of treated and control groups would help mitigate the confounding bias and minimize the upper bound of outcome inference errors [18, 35]. Therefore, we incorporate a representation balancing layer to force the distributions of hidden confounders of treated and controlled groups to be similar. Specifically, we adopt the integral probability metric (IPM) [35] to measure the difference between the distributions of treated instances and controlled instances in terms of their hidden confounder representations:

$$\mathcal{L}^{\text{disc}} = \alpha \cdot \text{IPM}(\mathcal{Z}_1, \mathcal{Z}_0), \quad (10)$$

where $\mathcal{Z}_1 = \{z_{i(j)}^{\leq t}\}_{i,t,j:c_{i(j)}^{\leq t}=1}$, $\mathcal{Z}_0 = \{z_{i(j)}^{\leq t}\}_{i,t,j:c_{i(j)}^{\leq t}=0}$ indicate the sets of hidden confounders for samples (in a batch) in the treated group and the controlled group, respectively. The IPM can be Wasserstein or Maximum Mean Discrepancy (MMD) distances. α is a hyperparameter that indicates the imbalance penalty.

Formally, we present the loss function of the proposed causal inference model as $\mathcal{L}^{\text{cau}} = \mathcal{L}^{\text{fact}} + \mathcal{L}^{\text{disc}}$.

4.2 Event Prediction with Causal Knowledge

To improve the robustness of event predictions with imperfect real-world data, we incorporate causal information output by the causal inference model as priors to forecast future events. We introduce two robust learning modules into the training of event predictors: (1) feature reweighting, which involves causal information to weight the original input features to obtain causally enhanced features, and (2) approximation constraints, which use the predicted potential outcomes as value range constraints applied to event prediction scores. Next, we introduce these two modules in detail.

4.2.1 *Feature Reweighting.* Feature reweighting was introduced in object detection [20], where a reweighting vector is learned to indicate the importance of meta features for detecting objects. Here, we introduce a new feature reweighting method that leverages causal information. We use ITEs estimated from the causal inference model to reweight the event frequency features to predict events.
Causal Feature Gates. We define a feature gate based on ITEs calculated using predicted potential outcomes from the causal inference model. For the j-th treatment event, the estimated ITE of a location at time $t + \delta$ is as follows:

$$\hat{\tau}_{(j)}^{t+\delta} = \hat{y}_{(j)}^{t+\delta}(1) - \hat{y}_{(j)}^{t+\delta}(0). \quad (11)$$

When considering multiple treatment events, we obtain the ITE vector $\hat{\tau}^{t+\delta} \in \mathbb{R}^E$, where each dimension refers to a treatment event.

The linear layer f_τ is then applied to model the association between the effects of different treatment events:

$$\boldsymbol{\beta}^{t+\delta} = \sigma\left(f_\tau(\hat{\tau}^{t+\delta})\right), \quad (12)$$

where $\boldsymbol{\beta}^{t+\delta} \in \mathbb{R}^E$ is the gating variables that will be applied to the original event frequency features. The sigmoid function σ converts the gating variable into a soft gated signal with a range of $(0, 1)$.
Reweighting Feature. We reweight the event frequency features using the gating variables defined above. It is worth emphasizing that the event frequency vector x^t has the same dimension as $\boldsymbol{\beta}^{t+\delta}$, and their corresponding elements represent the same event type. Nevertheless, we prefer not to apply the gating variables directly to the feature vector. ITE examines whether the binary treatment variable affects the outcome of an instance, while the event frequency vector includes discrete numbers. Thus, we transform the event frequency feature into a latent vector using a position-wise feed-forward network (FFN) [36]. It maps the features into a continuous space, assuming that the gating variables can be aligned with the variables in this space. The formal procedure is as follows:

$$\text{FFN}(x^t) = \text{ReLU}(x^t W^{r_0} + b^{r_0}) W^{r_1} + b^{r_1}, \quad (13)$$

$$\widetilde{x}^t = \text{FFN}(x^t) \odot \boldsymbol{\beta}^{t+\delta} + x^t, \quad (14)$$

where $\{W, b\}^{\{r_0, r_1\}}$ are learnable parameters. A residual connection is added to ensure that the causally weighted elements still contain some original information. We denote the causality enhanced features across Δ historical steps as $\widetilde{X}^{\leq t} \in \mathbb{R}^{E \times \Delta}$. Such features are fed into a predictor to perform event prediction. We denote the estimated event probability at time $t + \delta$ as $\hat{p}^{t+\delta} = \mathcal{P}(\widetilde{X}^{\leq t}, A)$.[3]

4.2.2 *Approximation Constraints.* The approximation constraints method was proposed to limit the reasonable range of the target variable during the model training process to generate a more robust model [27]. We follow this idea and propose a new method of integrating learned causal information into variable constraints. We first obtain the event prediction $\hat{p}^{t+\delta}$ from $\mathcal{P}$. Then, we assume that the causal range of the target variable, i.e., the event prediction, is $\hat{p}^{t+\delta} \in [l^{t+\delta}, u^{t+\delta}]$. The sample-wise boundaries are defined as:

$$l^{t+\delta} = \text{Min}(\hat{y}^{t+\delta}(\cdot)), \quad u^{t+\delta} = \text{Max}(\hat{y}^{t+\delta}(\cdot)), \quad (15)$$

where $\hat{y}^{t+\delta}(\cdot) = \{\hat{y}_{(j)}^{t+\delta}(1), \hat{y}_{(j)}^{t+\delta}(0) | 1 \leq j \leq E\}$ is the set of potential outcomes for all treatment events. The minimum and maximum values are considered the lower and upper limits of the target variable for a given sample. Based on the range obtained from causal knowledge, we define a constraint loss term:

$$\mathcal{L}^{\text{cstr}} = \sum_{t \in T} \sum_{i \in M} \text{ReLU}(l_i^{t+\delta} - \hat{p}_i^{t+\delta}) + \text{ReLU}(\hat{p}_i^{t+\delta} - u_i^{t+\delta}). \quad (16)$$

The loss term can be involved during the training of the predictor $\mathcal{P}$. Given the proposed robust learning modules for event prediction, we train the predictor by minimizing the following loss function:

$$\mathcal{L}^{\text{evt}} = \mathcal{L}^{\text{pred}} + \mu \cdot \mathcal{L}^{\text{cstr}}, \quad (17)$$

where $\mathcal{L}^{\text{pred}}$ is the loss function defined by the predictor $\mathcal{P}$ that minimizes the difference in event prediction $\hat{p}_i^{t+\delta}$ and ground truth event occurrence $y_i^{t+\delta}$. μ is a hyperparameter.

5 EXPERIMENTAL EVALUATION

We aim to answer the following research questions: **RQ1**: How well does **CAPE** estimate ITEs in observational event data? **RQ2**: Can

---

[3] $\hat{p}^{t+\delta} = \mathcal{P}(X^{\leq t}, A)$ when feature reweighting is not applied.

Table 1: Dataset statistics. M **is the number of locations. Positive indicates the ratio of positive samples, i.e., the protest event has occurred. Location represents the geographical level of events. For India, we select top locations based on the total number of events. Time Units is the time granularity of each timestamp when constructing the sample.**

Dataset	M	Positive	Location	Time	Time Unit	Source
India	14	30.1%	State	2000-2017	3 days	ICEWS
Nigeria	6	65.7%	Geopolitical zone	2015-2020	1 day	GDELT
Australia	8	44.4%	State	2015-2020	1 day	GDELT
Canada	13	26.8%	State	2015-2020	1 day	GDELT

CAPE improve the robustness of event prediction? **RQ3**: What causal information can we learn from studies of causally related event prediction?[4] We will introduce the experimental setup and discuss the experimental results for addressing the above questions.

5.1 Datasets

Experimental evaluation is conducted on two data sources: Integrated Conflict Early Warning System (ICEWS) [6], and Global Database of Events, Language, and Tone (GDELT) [23]. These two data sources include daily events encoded from news reports. For event data from GDELT, we only select root events identified in news reports. We construct event datasets for four countries, i.e., **India**, **Nigeria**, **Australia** and **Canada**, based on their large volume of events. Events are categorized into 20 main categories (e.g., appeal, protest, etc.) according to Conflict and Mediation Event Observations (CAMEO) [6]. Each event is encoded with geolocation, time, category, etc. We focus on predicting one category of event: *protest*, as the target variable, and using historical data of all event types as feature variables. Data statistics are shown in Table 1.

5.2 Evaluation Metrics

For the ITE estimation, since there are no ground truth counterfactual outcomes, we report the ATT error $\epsilon_{\text{ATT}} = |\text{ATT} - \mathbb{E}_{\mathcal{S}}[\hat{y}^{t+\delta}(1) - \hat{y}^{t+\delta}(0)|c^{\leq t} = 1]|$ [35], where ATT is the true average treatment effect on the treated, i.e., $\text{ATT} = \mathbb{E}_{\mathcal{S}}[y^{t+\delta}|c^{\leq t} = 1] - \mathbb{E}_{\mathcal{S}}[y^{t+\delta}|c^{\leq t} = 0]$. $\mathcal{S}$ denotes the subset of samples simulating a randomized controlled trial. Specifically, given the treatment event, we employ a 1-nearest neighbor algorithm [44] to find a matching control instance (without replacement) for each treated instance. Euclidean distance is adopted to measure feature vectors. The matching process is performed for each location.

We quantify the performance of event predictions based on Balanced Accuracy (BACC), i.e, BACC = (TPR + TNR)/2. TPR and TNR are the true positive rate and true negative rate, respectively. BACC is a good metric when the classes are imbalanced.

5.3 Comparative Methods

For the ITE estimation, we compare our causal inference model, notated as CAPE_C, with two groups of baselines: (i) Tree based methods: Bayesian Additive Regression Trees (**BART**) [8] and Causal

[4]Results are in the appendix.

Table 2: ITE estimation results showing the mean and standard deviation of ATT errors on all datasets with treatment event being *Appeal.* **Lower is better.**

	India	Nigeria	Australia	Canada
BART	.114 ± .012	.024 ± .015	.034 ± .018	.035 ± .017
CF	.112 ± .011	.034 ± .015	.038 ± .021	.038 ± .016
CFR-MMD	.014 ± .012	.020 ± .013	.016 ± .011	.017 ± .010
CFR-WASS	.021 ± .010	.025 ± .016	.019 ± .013	.012 ± .009
CEVAE	.012 ± .007	.018 ± .014	.018 ± .011	.017 ± .016
SITE	.013 ± .009	.019 ± .015	.012 ± .008	.016 ± .013
Net-Deconf	.034 ± .027	.030 ± .017	.022 ± .023	.034 ± .020
CAPE_C	**.011 ± .008**	**.010 ± .007**	**.010 ± .007**	**.008 ± .006**
CAPE_{C-G}	.019 ± .009	.019 ± .011	.019 ± .008	.016 ± .011
CAPE_{C-T}	**.011 ± .009**	.027 ± .021	.013 ± .010	.011 ± .007
CAPE_{C-B}	.018 ± .012	.026 ± .013	.020 ± .013	.019 ± .012

Forest (**CF**) [39]; (ii) Representation learning based methods: Counterfactual regression with MMD (**CFR-MMD**) [35] and Wasserstein metric (**CFR-WASS**) [35], Causal Effect Variational Autoencoder (**CEVAE**) [25], Network Deconfounder (**Net-Deconf**) [17], and Similarity Preserved Individual Treatment Effect (**SITE**) [45].

We study three variants of our model to examine the impact of different components in our model: (i) CAPE_{C-G} which removes the spatial feature learning. (ii) CAPE_{C-T} which replaces the temporal feature learning with a simple linear transformation. (iii) CAPE_{C-B} removes the loss term $\mathcal{L}^{\text{disc}}$.

To evaluate the effectiveness of proposed robust learning modules in event prediction, we adopt two spatiotemporal models as the predictor $\mathcal{P}$: (i) **Cola-GNN** [13]: A graph-based framework for long-term Influenza-like illness prediction; (ii) **GWNet** [43]: A state-of-the-art spatiotemporal graph model for traffic prediction. Given the spatiotemporal characteristics of societal event data, these models can be well applied to our problem. Note that we do not adopt protest event prediction models [10, 12] because they handle more complex data, such as text and knowledge graphs. We leave the causal exploration of such complex data to future work.

6 EXPERIMENTAL RESULTS

6.1 Results of ITE Estimation (RQ1)

To evaluate the effectiveness of our proposed causal inference framework, we limit the number of treatment events to be one and compare our model with other baselines. We report the ATT errors for the treatment event *appeal* in Table 2. The motivation is that *appeal* might be a potential cause of protests, as it expresses a serious or urgent request, typically to the public. The results show that tree-based models performs worse than representational learning-based models. This finding highlights the benefits of representation learning in estimating ITEs for observational event data. **CFR-MMD** and **CFR-WASS** learn a balanced representation such that the induced treated and control distributions look similar. Both models achieved good results in most cases, demonstrating the importance of controlling for representation distributions to predict potential outcomes. **CEVAE** learns latent variables based on variational autoencoders and **SITE** focuses on capturing local similarities to estimate ITEs. These two models present the most stable and relatively small ATT errors in all settings. This suggests that learning latent variables and considering similarity information are useful for estimating

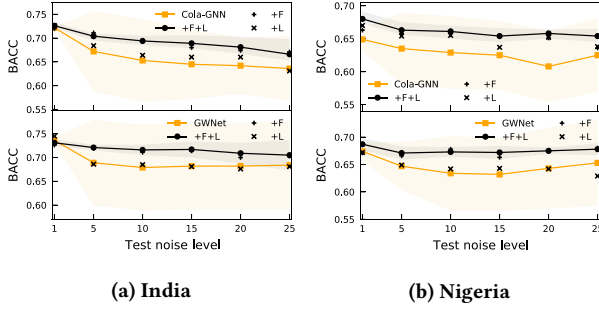**(a) India** **(b) Nigeria**

Figure 4: BACC value of event prediction when varying the noise level in validation and test sets. Higher is better.

ITEs for observational event data. **Net-Deconf** learns hidden confounders by leveraging network/spatial information. However, it does not outperform representation-based baselines. This may be because the model was designed for semi-synthetic datasets and the spatial characteristics of observational event data are different from the network used in the original paper. Our proposed causal inference framework learns hidden confounders while capturing spatial and temporal information and achieves the best performance. For our model variants, we observe that removing the representation balancing makes the results worse. Ignoring the temporal or spatial feature learning can also deteriorate the results. This reflects possible spatiotemporal dependencies underlying the hidden confounders, and demonstrates the capability of our model in capturing the spatiotemporal information from observational event data.

6.2 Robustness Tests in Event Prediction (RQ2)

Here, we perform two robustness tests on event prediction and conduct a case study on the proposed feature reweighting module.

*6.2.1 **Robustness to Test Noise**.* A model is considered to be robust if its output variable is consistently accurate when one or more input variables drastically change due to unforeseen circumstances. In this setting, we add Poisson noise into the validation and test sets while keeping the training data noise-free. We aim to verify whether our method guarantees good prediction performance when test input features are biased. We vary the rate parameter (aka expectation) of the Poisson distribution from 1 to 25 and provide the comparison results for different noise levels in Fig 4. We notice that training with the proposed robust learning module leads to higher average BACC results and lower variance over multiple runs. In most cases, the feature reweighting module (+F) contributes more in improving prediction results. Incorporating these two modules (+F+L) can lead to better overall results. The results suggest that the learned causal information is beneficial for improving the robustness of predictions.

*6.2.2 **Robustness to Training Noise**.* Human errors or machine failures in real-world data collection usually reduce data accuracy. With this motivation, we assume that only training data are biased and test whether our method can achieve decent event prediction results on unbiased test data. As shown in Fig. 5, applying robust learning modules can help the prediction model achieve better
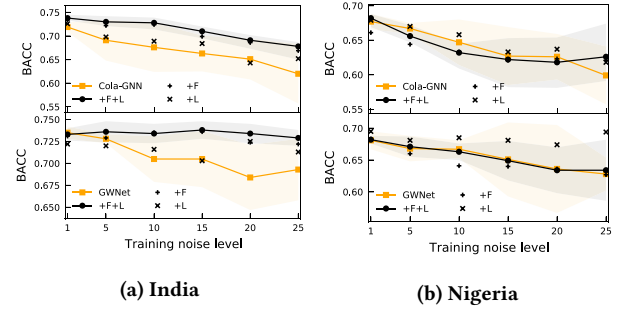

(a) India **(b) Nigeria**

Figure 5: BACC value of event prediction when varying the noise level in the training set. Higher is better.
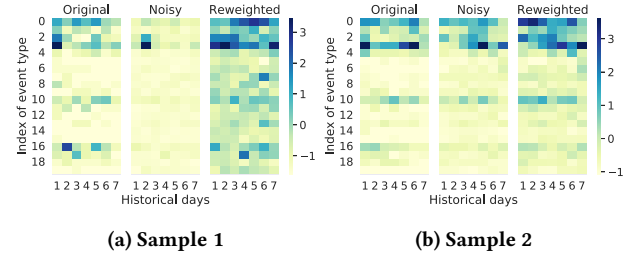

(a) Sample 1 **(b) Sample 2**

Figure 6: Distributions of original, noisy, and reweighted features of corrupted samples on the India dataset.

results in BACC when the noise level increases. Adding the approximation constraint loss (+L) can lead to a higher BACC than adding the two modules (e.g., Fig. 5b). The results also illustrate that even with biased data (with corrupted features), the trained causal inference model learns valuable information that contributes to the event prediction.

*6.2.3 **Case Study of Feature Reweighting**.* To illustrate the functionality of the feature reweighting on robust event prediction, we provide two examples in the India dataset, as shown in Fig. 6. We use **Cola-GNN** for analysis, given the more apparent improvements when it is applied with the feature reweighting module. Specifically, we first train an event prediction model on the India dataset using **Cola-GNN** with the feature reweighting module. We select two corrupted test samples with random noise added to their input features (noise level of 5). We visualize the original features, the noisy features, and the ones obtained from the feature reweighting module. We observe that the reweighted features can encode similar patterns of original features. It highlights the advantages of the ITEs used in the feature reweighting module and demonstrates its ability to capture crucial information underlying the data distribution.

7 CONCLUSION AND FUTURE WORK

Learning causal effects of societal events is beneficial to decision-making and helps practitioners understand the underlying dynamics of events. In this paper, we introduce a deep learning framework that can estimate the causal effects of societal events and predict societal events simultaneously. We design a novel spatiotemporal causal inference model for estimating ITEs and propose two robust learning modules that use the learned causal information

as prior knowledge for societal event prediction. We conducted extensive experiments on real-world event datasets and showed that our approach achieves the best results in ITE estimation and robust event prediction. One future direction is to examine other potential causes of event occurrence, such as events with specific themes and potentially biased media coverage.

8 BROADER IMPACTS

This work aims to advance computational social science by investigating causal effects among societal events from observational data. Causal effects among different types of societal events have not been extensively studied. In this work, we provide preliminary results on estimating the individual causal effects of one type of event on another and incorporate this causal information to improve the predictive power of event prediction models. We hope to provide a way to understand human behavior from the societal and causal inference aspects and broaden the possibilities for future work on societal event studies.

ACKNOWLEDGEMENTS

REFERENCES

[1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using twitter data. In *IEEE Conference on Computer Communications Workshops*. IEEE, 702–707.
[2] Andrew Anglemyer, Hacsi T Horvath, and Lisa Bero. 2014. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews* 4 (2014).
[3] Peter W Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. 2016. Interaction networks for learning about objects, relations and physics. *arXiv:1612.00222* (2016).
[4] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science* 2, 1 (2011), 1–8.
[5] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *RecSys*. 104–112.
[6] Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. ICEWS Coded Event Data.
[7] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. 2021. Spatial-temporal Causal Inference for Partial Image-to-video Adaptation. In *AAAI*, Vol. 35. 1027–1035.
[8] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.
[9] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *ICML*. PMLR, 933–941.
[10] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2019. Learning Dynamic Context Graphs for Predicting Social Events. In *KDD*. ACM, 1007–1016.
[11] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2020. Dynamic Knowledge Graph Based Multi-Event Forecasting. Association for Computing Machinery, New York, NY, USA.
[12] Songgaojun Deng, Huzefa Rangwala, and Yue Ning. 2021. Understanding Event Predictions via Contextualized Multilevel Feature Learning. In *CIKM*. 342–351.
[13] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. 2020. Cola-GNN: Cross-location Attention based Graph Neural Networks for Long-term ILI Prediction. In *CIKM*. 245–254.
[14] Michelangelo Diligenti, Soumali Roychowdhury, and Marco Gori. 2017. Integrating prior knowledge into deep learning. In *ICMLA*. IEEE, 920–923.
[15] Matthew S Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.
[16] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.
[17] Ruocheng Guo, Jundong Li, and Huan Liu. 2019. Learning individual treatment effects from networked observational data. *arXiv:1906.03485* (2019).
[18] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*. 3020–3029.
[19] Nathan Kallus. 2014. Predicting crowd behavior with big public data. In *WWW*. 625–630.
[20] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. 2019. Few-shot object detection via feature reweighting. In *ICCV*. 8420–8429.
[21] D Kinga and J Ba Adam. 2015. A method for stochastic optimization. In *ICLR*, Vol. 5.
[22] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.
[23] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, Vol. 2. Citeseer, 1–49.
[24] Jia Li, Xiaowei Jia, Haoyu Yang, Vipin Kumar, Michael Steinbach, and Gyorgy Simon. 2020. Teaching deep learning causal effects improves predictive performance. *arXiv:2011.05466* (2020).
[25] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *NIPS*. 6446–6456.
[26] Jing Ma, Ruocheng Guo, Chen Chen, Aidong Zhang, and Jundong Li. 2021. Deconfounding with Networked Observational Data in a Dynamic Environment. In *WSDM* (Virtual Event, Israel) *(WSDM '21)*. 166–174.
[27] Nikhil Muralidhar, Mohammad Raihanul Islam, Manish Marwah, Anuj Karpatne, and Naren Ramakrishnan. 2018. Incorporating prior domain knowledge into deep neural networks. In *ICBD*. IEEE, 36–45.
[28] Yue Ning, Sathappan Muthiah, Huzefa Rangwala, and Naren Ramakrishnan. 2016. Modeling precursors for event forecasting via nested multi-instance learning. In *KDD*. ACM, 1095–1104.
[29] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv:1609.03499* (2016).
[30] Judea Pearl et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[31] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *WWW*. 909–918.
[32] Kira Radinsky and Eric Horvitz. 2013. Mining the web to predict future events. In *WSDM*. 255–264.
[33] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
[34] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
[35] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. JMLR. org, 3076–3085.
[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*. 5998–6008.
[37] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, et al. 2019. Informed Machine Learning–A Taxonomy and Survey of Integrating Knowledge into Learning Systems. *arXiv:1903.12394* (2019).
[38] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. *arXiv:1910.13616* (2019).
[39] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.
[40] Xiaofeng Wang, Donald E Brown, and Matthew S Gerber. 2012. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *ISI*. IEEE, 36–41.
[41] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 231–238.
[42] Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *J. Amer. Statist. Assoc.* just-accepted (2019), 1–71.
[43] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv:1906.00121* (2019).
[44] Liu Yang and Rong Jin. 2006. Distance metric learning: A comprehensive survey. *Michigan State Universiy* 2, 2 (2006), 4.
[45] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation learning for treatment effect estimation from observational data. *NIPS* 31 (2018).
[46] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122* (2015).
[47] Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. 2014. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS one* 9, 10 (2014).
[48] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Spatiotemporal event forecasting in social media. In *SIAM*. SIAM, 963–971.
[49] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015. Multi-task learning for spatio-temporal event forecasting. In *KDD*. ACM, 1503–1512.

A PSEUDOCODE

The training steps of the proposed method are shown in Algorithm 1.

Algorithm 1: CAPE

Input: Observational event data
$\mathcal{D} = \left\{ \{ X_i^{\leq t}, c_i^{\leq t}, y_i^{t+\delta}, \}_{i=1}^M, \mathbf{A} \right\}_{t \in T}$, a predictor $\mathcal{P}$ with randomly initialized parameters, and initialized model **CAPE**, including causal inference model C and robust learning modules $\mathcal{R}$ for event prediction.

  ▷ Train the causal inference model C.

1 **while** C *has not converged* **do**
2   Input $\mathcal{D}$ to C to obtain predicted potential outcomes. Calculate the loss $\mathcal{L}^{\text{cau}}$.
3   Update C by minimizing $\mathcal{L}^{\text{cau}}$.
4 Freeze the causal inference model C.

  ▷ Train the predictor $\mathcal{P}$ and robust modules $\mathcal{R}$ with C.

5 **while** $\mathcal{P}, \mathcal{R}$ *have not converged* **do**
6   Input $\mathcal{D}$ to C to obtain predicted potential outcomes.
7   Calculate causally reweighed features $\widetilde{X}$ from robust learning modules $\mathcal{R}$ using predicted potential outcomes.
8   Input $\{\widetilde{X}, \mathbf{A}\}$ to the predictor $\mathcal{P}$ to obtain the event prediction. Calculate the prediction loss $\mathcal{L}^{\text{pred}}$.
9   Calculate the constraint loss $\mathcal{L}^{\text{cstr}}$ from $\mathcal{R}$ using predicted potential outcomes.
10   Update the predictor $\mathcal{P}$ and causal modules $\mathcal{R}$ by optimizing Eq. 17.

Output: the updated model **CAPE**.

B IMPLEMENTATION DETAILS

For the causal inference model, we use three gated temporal convolutional layers with dilation factors $d = 1, 2, 4$, and two graph convolutional layers. The dimension d_a is set to 10. The feature dimensions of all other hidden layers including d_s are set to be equal and searched from $\{16, 32, 64\}$.

The number of treatment events E is 20, where each treatment event corresponds to an event type, such as appeal and protest. Following previous work [10, 11], we set the historical window size Δ to 7 and the lead time δ to 1. The hyperparameter η used for parameter regularization is fixed to 1e-5. We use the squared linear MMD for representation balancing [35]. The imbalance penalty α is searched from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The scaling term μ in Eq. 17 is searched from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. All parameters are initialized with Glorot initialization [16] and trained using the Adam [21] optimizer with learning rate 10^{-3} and dropout rate 0.5. The batch size is set to 64. We use the objective value on the validation set for early stopping.

For causal inference baselines, **CF** [5] and **BART** [6] are implemented using R packages. We implement the causal inference models **CFR-MMD**, **CFR-WASS**, **SITE** by ourselves and use the source code of **CEVAE** [7] and **Net-Deconf** [8]. We apply parameter searching

---

[5] https://rdrr.io/cran/grf/man/causal_forest.html
[6] https://rdrr.io/cran/BART/
[7] https://github.com/rik-helwegen/CEVAE_pytorch
[8] https://github.com/rguo12/network-deconfounder-wsdm20

on all baseline models. For representation learning based approach, the dimension of hidden layers are searched from $\{32, 64, 128\}$ and the number of hidden layers are searched from $\{1, 2\}$. For models that introduce balancing representation learning, we search the hyperparameter from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. The model **Net-Deconf** involves an auxiliary network and we use the geographic adjacency matrix of locations.

For the experiments on event forecasting, we run the source code of **Cola-GNN** [9] and **GWNet** [10]. For event prediction models, we fix the dimension of hidden layers to 32. **Cola-GNN** takes the geographic adjacency matrix as input and **GWNet** learns the adaptive adjacency matrix.

We report the average of 5 randomized trials for all experiments. At each training, we randomly split the data into training, validation, and test sets at a ratio of 70%-15%-15% with a fixed seed value. All python code is implemented using Python 3.7.7 and Pytorch 1.5.0 with CUDA 9.2.

C MORE EXPERIMENTAL RESULTS

Table 3: ITE estimation results showing the mean and standard deviation of ATT errors on all datasets with treatment event being *Reject*. Lower is better.

|	India	Nigeria	Australia	Canada
BART	.184 ± .012	.072 ± .021	.045 ± .030	.070 ± .021
CF	.181 ± .013	.062 ± .021	.039 ± .031	.067 ± .022
CFR-MMD	.022 ± .011	.021 ± .020	.024 ± .013	.018 ± .009
CFR-WASS	.016 ± .012	.018 ± .009	.026 ± .021	.016 ± .011
CEVAE	.015 ± .010	.021 ± .018	.020 ± .012	.019 ± .015
SITE	**.010 ± .008**	.019 ± .010	.022 ± .019	.024 ± .012
Net-Deconf	.026 ± .021	.020 ± .018	.024 ± .019	.023 ± .014
CAPE$_C$	.016 ± .011	.016 ± .005	.012 ± .009	.015 ± .012
CAPE$_{C-G}$	**.010 ± .013**	.020 ± .012	.015 ± .010	**.011 ± .011**
CAPE$_{C-T}$	.015 ± .014	**.014 ± .013**	**.007 ± .004**	.016 ± .011
CAPE$_{C-B}$	.017 ± .013	.028 ± .013	.018 ± .014	.018 ± .009

(a) Australia   **(b) Canada**
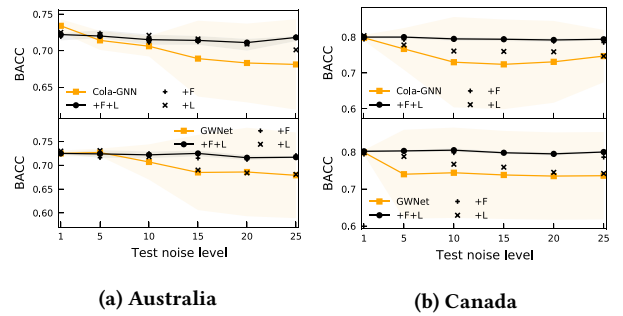
Figure 7: BACC value of event prediction when varying the noise level in validation and test sets. Higher is better.

---

[9] https://github.com/amy-deng/colagnn
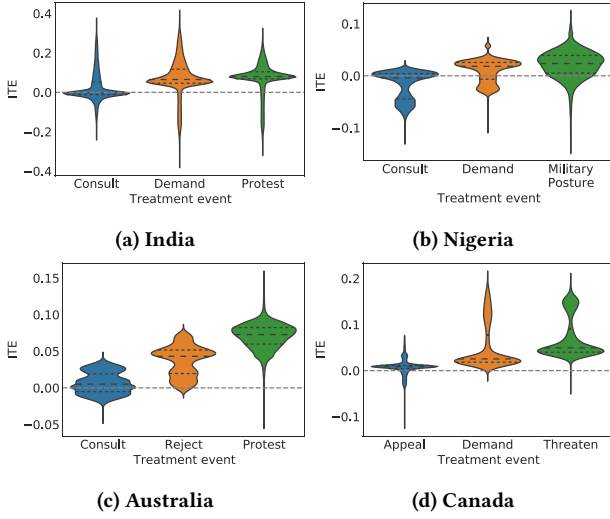[10] https://github.com/nnzhan/Graph-WaveNet

Figure 9: ITE distribution of different treatment events on the outcome *protest* in violin plot. The loosely dashed line represents the median.
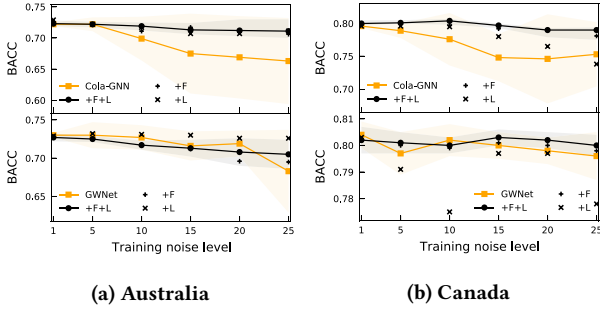

Figure 8: BACC value of event prediction when varying the noise level in the training set. Higher is better.

D CAUSAL EFFECT IN SOCIETAL EVENTS (RQ3)

In our study, whether there is a significant increase in certain types of events (e.g., appeal) over the past window is defined as the treatment of a location. The outcome is the future occurrence of a target event, i.e., protest. In this case, the ITE measures the difference in the outcome of the protest occurring between the two scenarios of the treatment event (i.e., increased or not). Thus, when the necessary assumptions hold, it implies a causal effect of the treatment event on the protest. A higher ITE suggests that an increase in a treatment event will be more influential in the occurrence of future protests, compared to a decrease or no change.

To better illustrate the effect of treatment events on future protests, we visualize the predicted ITEs based on Eq. 11. Violin plots for the four datasets are shown in Fig. 9. We select three treatment events for each dataset. They have relatively low, moderate, and high ITEs on average, respectively. The results vary from datasets due to different social environments. In India and Australia, massive historical protests may lead to future protests. In Nigeria and Canada, events related to military posture and threats, respectively, are likely to be more dominant factors in future protests. Nevertheless, we hardly conclude that protests will occur when the treatment event increases substantially because both types of events (i.e., treatment and outcome events) can be affected by hidden variables (i.e., unknown social factors such as economic and demographic factors). These results can provide supporting evidence for conjectures on protest triggers and generate hypotheses for future experiments.