Published at ProbML 2026

Causal Temporal Graphs for Counterfactual Validation of Temporal Link Prediction

Aniq Ur Rahman[1] **Justin P. Coon**[1]

[1]Department of Engineering Science, University of Oxford
[†]Correspondence to aniq.rahman@eng.ox.ac.uk

Abstract

Temporal link prediction (TLP) models are commonly evaluated based on predictive accuracy, yet such evaluations do not assess whether these models capture the causal mechanisms that govern temporal interactions. In this work, we propose a framework for counterfactual validation of TLP models by generating causal temporal interaction graphs (CTIGs) with known ground-truth causal structure. We first introduce a structural equation model for continuous-time event sequences that supports both excitatory and inhibitory effects, and then extend this mechanism to temporal interaction graphs. To compare causal models, we propose a divergence metric based on cross-model predictive error, and empirically validate the hypothesis that predictors trained on one causal model degrade when evaluated on sufficiently distant models. Finally, we instantiate counterfactual evaluation under (i) controlled causal shifts between generating models and (ii) timestamp shuffling as a stochastic distortion with measurable causal divergence. Our framework provides a foundation for causality-aware benchmarking.

1. Introduction

Temporal Interaction Graphs (TIGs) are used to model real-world phenomena in which entities interact for a certain duration, which may be instantaneous. A social network serves as a prime example, where edges form between users interacting through messaging or commenting on each other's posts.

In graph learning, *link prediction* refers to the task of predicting the existence of edges in a static graph after having observed other parts of the graph. In the dynamic setting, we have *temporal link prediction* (TLP), wherein after observing a TIG up to a certain timestamp, a model makes predictions on the existence of edges at specific timestamps in the future.

While the problem of TLP itself has gained popularity in the past years (Kumar et al., 2019; Rossi et al., 2020; Yu et al., 2023), the question of whether these models capture the underlying causal mechanisms in the data remains largely unexplored (Rahman et al., 2025a). Recently, we have evaluated TLP models in a counterfactual setting (Rahman et al., 2025b), concluding that the models cannot distinguish between TIGs if the order in which the edges occur is changed, or the frequency with which they appear is changed. However, the real-world datasets are not certifiably causal (Pearl, 2009); in particular, the causal influence of past temporal links on future interactions is generally unknown. To address this limitation, we propose a method for generating causal temporal interaction graphs (CTIGs), enabling counterfactual evaluation of TLP models using data with known causal structure.

Although research on CTIGs is limited, insights can be drawn from the well-established adjacent field of *causal event sequences* (CES) (Kim et al., 2011; Noorbakhsh and Rodriguez, 2022; Jalaldoust et al., 2022; Cüppers et al., 2024) and from work on *event prediction* (Zhao, 2021). Accordingly, we first develop a model to generate CESs and then naturally extend it to generate CTIGs, where each event corresponds to an edge event. Once a CTIG is generated from a causal model, we are interested in quantifying how it differs from a CTIG generated by a different causal model, which leads to the following question:

> *How do we measure the divergence between two CTIG-generating causal models?*

This question also lays the foundation for testing the following *hypothesis*:

> *A predictive model trained on data generated by a causal model will perform worse on data generated by another causal model, provided that the divergence between the two models is sufficiently large.*

To test this hypothesis, we generate CTIGs from different causal models and query an oracle with access to the true causal parameters to make predictions. We then extend this framework to incorporate causal distortions and repeat the counterfactual experiments of Rahman et al. (2025b) using some well-known temporal link prediction (TLP) models.

Organisation In § 2, we introduce a structural equation model for generating causal event sequences and discuss the properties of the resulting causal model. Then, in § 3, we propose a divergence measure to quantify the difference between two causal models. Next, in § 4, we extend the CES generation procedure to create causal temporal interaction graphs, beginning with the generation of node features and using them as the basis for defining the underlying causal graph. In § 5, we present the counterfactual setup, motivate it through an empirical study, and apply it to TLP models. We propose a method to benchmark TLP models under controlled causal shifts and timestamp shuffling. Finally, we conclude in § 6. Due to page limitations, the preliminaries, related works, and proofs of theoretical properties of the causal model, are deferred to the appendix.

Notation We use $\emptyset$ for the empty set, $|\mathcal{A}|$ for the cardinality of a collection $\mathcal{A}$, and $|\mathcal{A}^{(y)}|$ for the number of element y occurrences in $\mathcal{A}$. We use $\mathbb{N}^+$ and $\mathbb{R}^+$ for sets of positive natural and positive real numbers, respectively. For $n \in \mathbb{N}^+$, we let $[n] = \{1, 2, \ldots, n\}$, and for $a, b \in \mathbb{R}$ such that $a < b$, we let $[a, b) = \{c \in \mathbb{R} \mid a \leq c < b\}$. We use uppercase letters, for example U, for variables, lowercase letters, for example u, for values, and $\varnothing$ for the null value. The indicator function is represented by $\mathbb{I}\{\cdot\} \in \{0, 1\}$ the argument of which is a logical statement. Expectation and variance is denoted by $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$, respectively.

2. Causal Event Sequence

Consider a set of events, $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$, that occur continuously over time. Each data point in this sequence can be represented as a pair, where one component identifies the event, and the other specifies the time of occurrence. For instance, given an observed event sequence: $(\boldsymbol{a}, t_1)$, $(\boldsymbol{c}, t_2)$, $(\boldsymbol{a}, t_3), (\boldsymbol{b}, t_4), (\boldsymbol{c}, t_5), \ldots, (\boldsymbol{a}, t_m)$, the following predictive questions can be posed: (1) Can we predict when the next event will occur? (2) If the timing of the next event, t_{m+1}, is known, can we predict which event will occur at that time? (3) Can we estimate the probability that an event $e \in \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ occurs within a specified future time interval? (4) Can we predict the number of times an event $e \in \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ will occur within a future time interval?

These questions highlight fundamental challenges in forecasting events that occur in continuous time. Now suppose there exists a function f that takes an event sequence as input and produces answers to these predictive queries. A natural question then arises:

Is there reason to believe that the occurrence of past events influences which events will occur in the future, and when they will occur?

If past events have no impact on future events, then providing them as input to f is futile. However, Gong et al. (2023) note that *when data arrives sequentially, it is often assumed that temporal dependencies exist.* Therefore, we propose a causal mechanism to generate event sequences in continuous time so that the assumption holds. Moreover, we derive the following *design principles* for a causal mechanism based on (Gong et al., 2023):

(*D1*) Causal relationships take time to propagate, i.e., suppose an event $\boldsymbol{a}$ causes another event $\boldsymbol{b}$, then, if $\boldsymbol{a}$ occurs at time t, then event $\boldsymbol{b}$ might occur after some delay.

(*D2*) To infer causal relationships among various events, one must consider both the temporal order in which they occur and the delay between them.

(*D3*) The cause and effect events unfold on a single continuous timeline on a rolling basis.

Proposed Model Consider a sequence of events with n distinct event types that occur repeatedly over time. Each event type has an associated *trigger event sequence*, which determines when it may be activated. We define $U_i(t) \in \{0, 1\}$ as an indicator of whether an event of type i is triggered at time t. Given past events, an event of type i may or may not occur at time t, which we denote as $X_i(t) \in \{0, 1\}$. To capture historical influence, we define $X'_i(t) \in \{0, 1\}$ as the presence of event i within the time window $(t - \bar{\tau}, t)$. Therefore, the causal model can be functionally expressed as[1]:

$$x_i(t) = f\Big(\underbrace{\{x'_j(t) : \forall j \in pa_i\}}_{\text{past events}}, \underbrace{u_i(t)}_{\text{trigger}}\Big). \tag{1}$$

For each event of type i, we define a *structural equation model* (SEM) as follows[2]:

$$\begin{aligned} x_i(t) = f(pa_i(t), u_i(t)) &= u_i(t) \cdot \hat{f}(pa_i(t)) \\ &= u_i(t) \cdot \mathbb{I}\left\{\sum_{j \in pa_i(t)} \Theta_{i,j} \geq 0\right\} \qquad (2) \\ &= u_i(t) \cdot \mathbb{I}\left\{\sum_{j \in pa_i} \Theta_{i,j}\, x'_j(t) \geq 0\right\}, \qquad (3) \end{aligned}$$

where $pa_i(t) \subseteq pa_i \subseteq [n]$ denotes the parents of $X_i(t)$ in a causal graph[3], excluding $U_i(t)$, and $\Theta_{i,j} \in [-1, 1]$ is a parameter which we refer to as the *inter-event causal influence.* If $\Theta_{i,j} > 0$, we say that event of type j *excites* the event i, and if $\Theta_{i,j} < 0$, we say that event j *inhibits* event i. The excitatory and inhibitory influence of one event on another was studied in (Kim et al., 2011) while modelling neural spiking activity. To avoid confusion, we refer to the set pa_i as the structural parents of i defined as $pa_i = \{j : \Theta_{i,j} \neq 0, \forall j \in [n]\}$, and $pa_i(t) \subseteq pa_i$ as the active parents of i at time t defined as $pa_i(t) = \{j : x'_j(t) = 1, \forall j \in pa_i\}$. *With slight abuse of notation, when we say that j is a parent of i, we mean that $X'_j(t)$ is a parent of $X_i(t)$.* Therefore, even when $i = j$, the causal graph remains acyclic, as $X'_i(t)$ and $X_i(t)$ represent distinct nodes.

To specify the causal model, we describe how the causal graph, trigger event sequences, and causal influence parameters are generated:

1. While the variables are written in uppercase, their values are written in lowercase.
2. In contrast to equation 1, writing the input as $pa_i(t)$ implicitly supplies the indicators $x'_j(t)$ for all $j \in pa_i(t)$.
3. In a causal graph, a directed edge from variable A to variable B indicates that A is a cause of B.

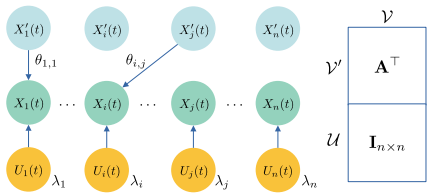


Figure 1. In this figure, we depict a sample causal graph $\mathscr{G}$ with three types of nodes: event nodes X_i, past-event indicators X'_j, and trigger nodes U_k. The causal arrow from $X'_j(t)$ to $X_i(t)$ is annotated with the causal influence parameter $\Theta_{i,j}$. For the triggers, we indicate their generation rate λ_k, following a homogeneous Poisson process.

- For the causal graph, we focus on the bipartite graph from $\mathcal{V}' = \{X'_i(t) : i \in [n]\}$ to $\mathcal{V} = \{X_i(t) : i \in [n]\}$, whose adjacency matrix $\mathbf{A}$ is sampled from a parametric random graph model.[4]
- For each event of type i, its trigger event sequence Φ_i is generated by sampling a homogeneous Poisson point process (Haenggi, 2012, § 2.4.1) with intensity λ_i which can be sampled randomly from another distribution whose support is $\mathbb{R}^+$.
- Lastly, the *inter-event causal influence* parameters $\{\Theta_{i,j} : \mathbf{A}_{i,j} = 1, \forall i, j \in [n]\}$ are sampled from a distribution with support $[-1, 1]$, filtering out $\{0\}$.

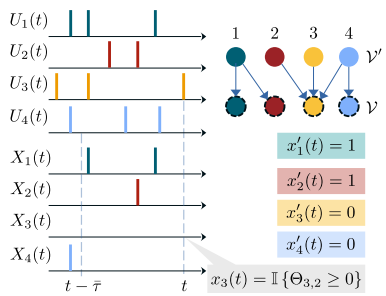


Figure 2. In this figure, we depict the generation process visually for a causal graph with 4 event nodes. At time t, $u_3(t) = 1$ so we evaluate $x_3(t) = f(pa_3(t), u_3(t))$ and find that only $x'_2(t) = 1$ from among its parents $pa_3(t)$. This allows us to determine $x_3(t) = \mathbb{I}\{\Theta_{3,2} \geq 0\}$. To determine the value of $x'_j(t)$ we only consider the time window $[t - \bar{\tau}, t)$.

3. Comparing Causal Models

Having discussed how to generate a causal model and from it, an event sequence, we now turn to quantify the divergence between two causal models. Specifically, we evaluate how a model performs on CESs generated from another model, and vice versa. Finally, we report the average error across multiple CES realisations from both models.

We denote a causal model by $\mathfrak{C}$, parameterised by the triple $(\Lambda, \Theta, \bar{\tau})$. Given a time horizon $T \in \mathbb{R}^+$, the model generates a random trigger event sequence $\Phi \in \mathbb{S}$ and an event sequence $\mathcal{S} \in \mathbb{S}$ up to time T, which we write as $(\Phi, \mathcal{S}) \sim \mathfrak{C}(T)$, with $\mathcal{S} \subseteq \Phi$.

To facilitate comparisons between models, we introduce a consistent notation to distinguish their parameters and generated sequences. Consider two causal models, $\mathfrak{C}_A$ and $\mathfrak{C}_B$, with $(\Phi^A, \mathcal{S}^A) \sim \mathfrak{C}_A(T)$, and $(\Phi^B, \mathcal{S}^B) \sim \mathfrak{C}_B(T)$. We then define the function $\tilde{f}_B(i, t, \mathcal{S}^A, \Phi^A)$, which uses model $\mathfrak{C}_B$ to make predictions on the event sequence $\mathcal{S}^A$, as[5]

$$\tilde{f}_B(i, t, \mathcal{S}^A, \Phi^A) = u_i^A(t) \cdot \mathbb{I}\left\{ \sum_{j \in pa_i^B} \Theta_{i,j}^B \, x_j'^A(t) \geq 0 \right\}. \tag{4}$$

4. Standard random graph models such as the Erdős–Rényi (ER) model (Erdos, 1961), the stochastic block model (SBM) (Abbe, 2018), or the Barabási–Albert (BA) model (Barabási, 2014, § 5) can be used.
5. The superscript A in $u_i^A(t)$ and $x_j'^A(t)$ indicates that these quantities are derived from Φ_i^A and $\mathcal{S}^A$, respectively.

We define the following divergence to quantify the predictive error of $\mathfrak{C}_B$ on $\mathcal{S}^A$:

$$d_B(\mathcal{S}^A, \Phi^A) = \frac{1}{n} \sum_{i \in [n]} \mathop{\mathbb{E}}_{t \in \Phi_i^A} \left[\left| \tilde{f}_B(i, t, \mathcal{S}^A, \Phi^A) - \tilde{f}_A(i, t, \mathcal{S}^A, \Phi^A) \right| \right]. \tag{5}$$

To make the divergence measure between the causal models $\mathfrak{C}_A$ and $\mathfrak{C}_B$ symmetric, we define[6]

$$d_{A,B}(\mathcal{S}^A, \Phi^A, \mathcal{S}^B, \Phi^B) = \sqrt{d_B(\mathcal{S}^A, \Phi^A) \cdot d_A(\mathcal{S}^B, \Phi^B)}. \tag{6}$$

In order to make the divergence metric invariant to the sampled sequences, we calculate the mean over multiple samples, defining the final divergence metric as

$$\bar{d}_{A,B} = \mathop{\mathbb{E}}_{\substack{(\Phi^A, \mathcal{S}^A) \sim \mathfrak{C}_A(T) \\ (\Phi^B, \mathcal{S}^B) \sim \mathfrak{C}_B(T)}} \left[d_{A,B}(\mathcal{S}^A, \Phi^A, \mathcal{S}^B, \Phi^B) \right]. \tag{7}$$

To empirically assess the statistical stability of the proposed divergence measure $\bar{d}_{A,B}$, we examine the decay of its variance as the effective sample size increases. Fig. 3 provides empirical evidence that the estimator becomes more stable as more data are observed.

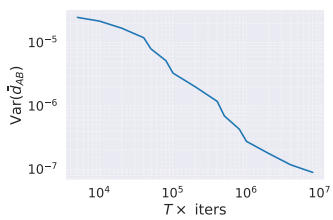


Figure 3. Empirical variance of the estimated divergence $\bar{d}_{AB}$ between two independent causal models, plotted as a function of the effective sample size $T \times$ iters.

4. Causal Temporal Interaction Graphs

In this section, we consider events as interactions between nodes, which together form a temporal interaction graph. Accordingly, instead of n distinct events, we now have a graph with n nodes, leading to a total of $E = \binom{n}{2}$ possible edges.

We begin by generating node feature vectors $\mathbf{x}_a \in \mathbb{R}^r, \forall a \in [n]$, which in practice are obtained by randomly sampling from a distribution with support on $\mathbb{R}^r$, such as a multivariate normal distribution. We then encode the relationship between two nodes a and b through edge feature vector $\mathbf{z}_{(a,b)} \in \mathbb{R}^r$, computed from their respective node features $\mathbf{x}_a$ and $\mathbf{x}_b$ via a *symmetric*[7] function $\mathbf{z}_{(a,b)} = g(\mathbf{x}_a, \mathbf{x}_b)$ ensuring $\mathbf{z}_{(a,b)} = \mathbf{z}_{(b,a)}$. As a simple choice, we define

$$g(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{x}_a \odot \mathbf{x}_b. \tag{8}$$

Having defined edge features, we now model causal influence between edges as a function of these features. Because causal influence between edge events is inherently directional (e.g., one interaction can influence another without reciprocity), we require this function, denoted by h to be *asymmetric*. Let the features of two edges be $\mathbf{z}$ and $\mathbf{z}'$, then

$$h(\mathbf{z}, \mathbf{z}') = \sin\left(\nu_0 \tanh\left(\mathbf{z}^\top \mathbf{B}\, \mathbf{z}'\right)\right), \qquad \nu_0 \in \mathbb{R}, \quad \mathbf{B} \neq \mathbf{B}^\top. \tag{9}$$

The sine function bounds the output to $[-1, 1]$, and ν_0 adds a tunable scaling that increases the flexibility of h. To ensure asymmetry of h, we construct $\mathbf{B}$ as a skew-symmetric matrix, as follows:

$$\hat{\mathbf{B}} = \left[\hat{\mathbf{b}}_i\right]_{i=1}^r, \; \hat{\mathbf{b}}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_{r \times r}); \quad \mathbf{B} = \hat{\mathbf{B}} - \hat{\mathbf{B}}^\top. \tag{10}$$

6. The shorthand for $d_B(\mathcal{S}^A, \Phi^A)$ is d_B^A.
7. We assume undirected interactions; directed interactions would require $g(\cdot, \cdot)$ to be asymmetric.

The causal influence of edge j on i can be defined directly as $\Theta_{i,j} = h(\mathbf{z}_i, \mathbf{z}_j)$. However, this formulation is restrictive from a design perspective. To further generalise the model, we introduce thresholding of the output of h to control the *sparsity* of the resulting causal graph. Moreover, we allow certain edges to be designated as *non-causal*, meaning they neither exert causal influence on other edges nor are influenced by them, thereby behaving as spurious noise in the data.

We first apply thresholding to the output of the influence function h in order to control sparsity. Specifically, for a threshold $\nu_1 \in (0, 1)$, we define the function $\psi(x) = x \cdot \mathbb{I}\{|x| \geq \nu_1\}$. We then define the causal influence matrix $\tilde{\Theta}$, prior to introducing non-causal edges, with entries

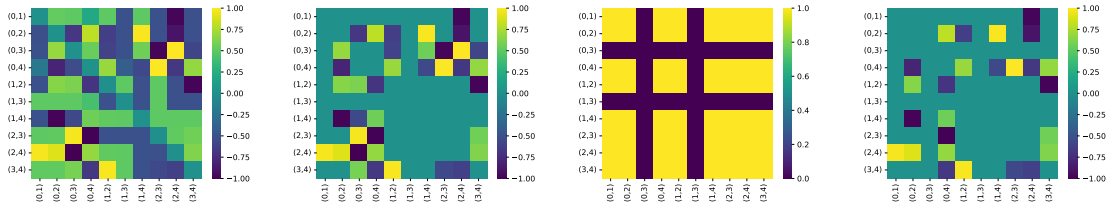
$$\tilde{\Theta}_{i,j} = \psi\left(h(\mathbf{z}_i, \mathbf{z}_j)\right). \tag{11}$$

In real-world scenarios, some edges correspond to spurious or unrelated events that have no causal relation with other edges. To model this, we designate $l < E$ edges as non-causal by sampling a subset of l distinct edges uniformly at random from $[E]$. To this end, we construct a symmetric binary mask $\mathbf{M}_l \in \{0, 1\}^{E \times E}$ with elements in the corresponding l rows and columns set to 0. The causal influence matrix Θ, is then obtained as:

$$\Theta = \tilde{\Theta} \odot \mathbf{M}_l. \tag{12}$$

Lastly, the causal graph $\mathbf{A}$ can be obtained as $\mathbf{A}_{i,j} = \mathbb{I}\{\Theta_{i,j} \neq 0\}$.

In Fig. 4, we illustrate the step-by-step construction[8] of the causal influence matrix Θ for a CTIG model, instantiated on a graph with 5 nodes.



(a) $\mathbf{H}, \nu_0 = 100$ (b) $\tilde{\Theta}, \nu_1 = 0.55$ (c) $\mathbf{M}_l, l = 2$ (d) Θ

Figure 4: Construction of the CTIG causal parameters for a graph with $n = 5$ and $r = 5$.

5. Counterfactual Analysis

Motivation Borrowing notation from § 3, we denote the *true* causal model by $\mathfrak{C}_0 = (\Lambda_0, \Theta_0)$, the *estimated* model by $\mathfrak{C}_\star = (\Lambda_\star, \Theta_\star)$, and the *distorted* causal model by $\mathfrak{C}_\dagger = (\Lambda_\dagger, \Theta_\dagger)$.

The divergence function of the estimated model is $d_\star : \mathbb{S} \times \mathbb{S} \to [0, 1]$, which maps an event sequence $\mathcal{S} \in \mathbb{S}$ and its trigger sequence $\Phi \in \mathbb{S}$ generated by a causal model $\mathfrak{C}$ to a scalar performance error. Accordingly, the performance of the estimated model on a sequence generated by the true causal model is $1 - d_\star(\mathcal{S}^0, \Phi^0)$, while its performance on a sequence generated by the distorted model is $1 - d_\star(\mathcal{S}^\dagger, \Phi^\dagger)$.

Comparing the performance achieved by the estimated model on the distorted sequence with the true sequence, we measure the gap $\Delta_\star(\mathcal{S}^0, \Phi^0, \mathcal{S}^\dagger, \Phi^\dagger)$ as

$$\underbrace{\Delta_\star(\mathcal{S}^0, \Phi^0, \mathcal{S}^\dagger, \Phi^\dagger)}_{\Delta_\star^{0,\dagger}} = \left(1 - d_\star(\mathcal{S}^0, \Phi^0)\right) - \left(1 - d_\star(\mathcal{S}^\dagger, \Phi^\dagger)\right) = \underbrace{d_\star(\mathcal{S}^\dagger, \Phi^\dagger)}_{d_\star^\dagger} - \underbrace{d_\star(\mathcal{S}^0, \Phi^0)}_{d_\star^0}.$$

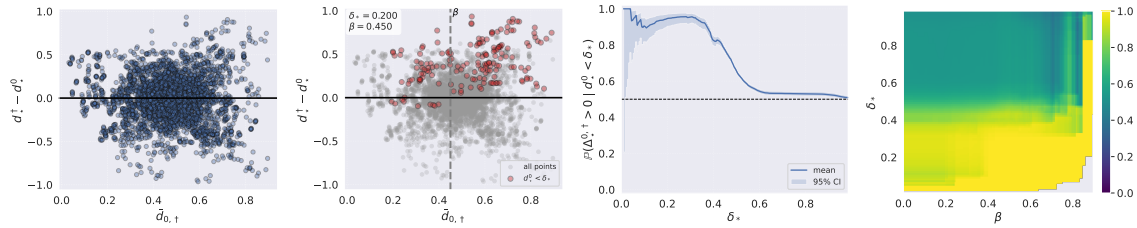
We then define the mapping $\Delta_\star(\mathcal{S}^0, \Phi^0, \mathcal{S}^\dagger, \Phi^\dagger) \mapsto \bar{d}_{0,\dagger}$, for realisations $(\Phi^0, \mathcal{S}^0) \sim \mathfrak{C}_0(T)$, and $(\Phi^\dagger, \mathcal{S}^\dagger) \sim \mathfrak{C}_\dagger(T)$, where $\bar{d}_{0,\dagger} \in [0, 1]$ is the divergence between the true and distorted causal models (see equation 7). This mapping can be used to analyse the relation between the performance gap and the divergence between the causal models $\mathfrak{C}_0$ and $\mathfrak{C}_\dagger$. We then introduce the following hypothesis.

8. The code will be made publicly available on Github.com/Aniq55/CES.

Hypothesis 5.1. *For any causal model estimate $\mathfrak{C}_\star$ satisfying $d_\star^0 \in [0, \delta_\star)$, there exists $\beta \in (0,1)$ such that for all distorted models $\mathfrak{C}_\dagger$ with $\bar{d}_{0,\dagger} > \beta$, the performance gap satisfies $\Delta_\star^{0,\dagger} > 0$.*

In other words, the estimated model performs better on the true sequence compared to the distorted sequence, if the divergence of the distorted model $\mathfrak{C}_\dagger$ from the true causal model $\mathfrak{C}_0$ is large enough, and if $\mathfrak{C}_\star$ is sufficiently accurate, i.e., its error on the true sequence satisfies $0 \leq d_\star^0 < \delta_\star$.

We empirically examine the relationship between the performance gap $\Delta_\star^{0,\dagger} = d_\star^\dagger - d_\star^0$ and the divergence between the true and distorted causal models $\bar{d}_{0,\dagger}$ in Fig. 5 (see Appendix D).



(a) $\left(\bar{d}_{0,\dagger}, d_\star^\dagger - d_\star^0\right)$ (b) $\left(\bar{d}_{0,\dagger}, d_\star^\dagger - d_\star^0\right) : d_\star^0 < \delta_\star$ (c) $\mathbb{P}\left(\Delta_\star^{0,\dagger} > 0 \mid d_\star^0 < \delta_\star\right)$ (d) $\mathbb{P}\left(\Delta_\star^{0,\dagger} > 0 \mid \delta_\star, \beta\right)$

Figure 5: Empirical evaluation of the performance gap under causal distortion. The panels visualize how the performance gap $\Delta_\star^{0,\dagger}$ and its associated probabilities vary with the causal divergence $\bar{d}_{0,\dagger}$ and the accuracy threshold $\delta_\star$, providing support for Hypothesis 5.1. $n = 7, T = 10^3$.

Fig. 5(a) shows a scatter plot of the performance gap against $\bar{d}_{0,\dagger}$. In Fig. 5(b), we restrict attention to settings where the estimated model is sufficiently accurate, i.e., $d_\star^0 < \delta_\star$ with $\delta_\star = 0.2$, and highlight that for distortions satisfying $\bar{d}_{0,\dagger} > 0.45$, the performance gap is strictly positive. In Fig. 5(c), we instead vary $\delta_\star$ and observe that the probability $\mathbb{P}(\Delta_\star^{0,\dagger} > 0 \mid d_\star^0 < \delta_\star)$ decreases from near 1 toward 0.5 as the accuracy constraint is relaxed. Finally, Fig. 5(d) presents a heat map $\mathbb{P}\left(\Delta_\star^{0,\dagger} > 0 \mid d_\star^0 < \delta_\star, \bar{d}_{0,\dagger} > \beta\right)$ over $(\beta, \delta_\star)$, illustrating that for larger $\delta_\star$, increasing β alone is insufficient to guarantee an all-positive performance gap.

Application Hypothesis 5.1 states that if the estimator $\mathfrak{C}_\star$ is sufficiently accurate, i.e., $d_\star^0 \leq \delta_\star$, then there exists $\beta \in (0,1)$ such that for every distorted causal model $\mathfrak{C}_\dagger$ with $\bar{d}_{0,\dagger} > \beta$, the corresponding performance gap satisfies $\Delta_\star^{0,\dagger} > 0$ with probability 1 under the joint sampling $(\mathcal{S}^0, \Phi^0) \sim \mathfrak{C}_0(T)$ and $(\mathcal{S}^\dagger, \Phi^\dagger) \sim \mathfrak{C}_\dagger(T)$. Formally, this can be written as

$$\mathbb{P}\left(\Delta_\star^{0,\dagger} > 0 \mid d_\star^0 < \delta_\star, \bar{d}_{0,\dagger} > \beta\right) = 1. \tag{13}$$

In many practical settings, computing $\bar{d}_{0,\dagger}$ may be infeasible. For instance, one may only observe event sequences, while the parameters of the causal models that generate them remain unknown. Moreover, the true causal model $\mathfrak{C}_0$ may not be available, or the causally distorted sequences $\mathcal{S}^\dagger$ may be produced by an alternative mechanism, such as corruption of the true sequence $\mathcal{S}^0$ via an external procedure, rendering the direct computation of $d_\dagger^0$, and consequently $\bar{d}_{0,\dagger}$, impossible. Consequently, the statement in equation 13 is not directly verifiable in practice.

However, we can still measure $d_\star^0$ as the prediction error of the estimated model on sequences generated by the true causal model. In the absence of $\bar{d}_{0,\dagger}$, we therefore use the probability

$$p_\star^{0,\dagger}(\delta_\star) \triangleq \mathbb{P}\left(\Delta_\star^{0,\dagger} > 0 \mid d_\star^0 < \delta_\star\right)$$

as a *proxy* for quantifying the strength of the distortions in $\mathcal{S}^\dagger$. Moreover, for a fixed $\delta_\star$, this quantity enables comparison between predictors: given identical data, the predictor attaining a higher value of $p_\star^{0,\dagger}(\delta_\star)$ is deemed to perform better.[9]

9. after comparing the $1 - d_\star^0$ values of the predictors

Negative Sampling In practice, trigger events are typically unobserved[10]; accordingly, we construct a set of negative events via *negative sampling*.[11] We then define

$$\Phi \;=\; \mathcal{S} \;\cup\; \{(j,t) : (i,t) \in \mathcal{S},\; j \in [E] \setminus \{i\}\}\,. \tag{14}$$

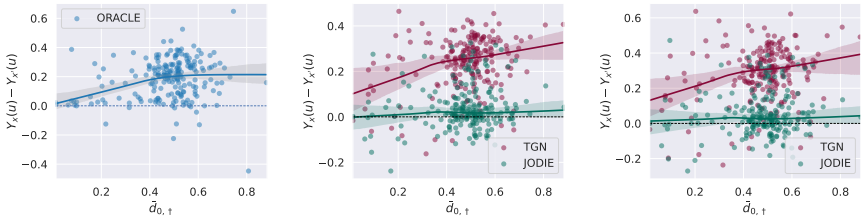
Note that events in $\Phi \setminus \mathcal{S}$ are not guaranteed to be infeasible for the causal model under consideration.[12] Consequently, Φ should not be interpreted as the set of true trigger events. In the *transductive* setting the negative samples are only drawn uniformly from the observed set of edges so that $j \sim \mathrm{Unif}\left(\{k : (k,t) \in \mathcal{S}\} \setminus \{i\}\right)$.

Experiment A We consider two causal event sequences: $\mathcal{S}^0$ generated by a causal model $\mathfrak{C}_0(T)$, and $\mathcal{S}^\dagger$ generated by a causal model $\mathfrak{C}_\dagger(T)$. We split[13] the sequence $\mathcal{S}^0$ chronologically into training and test sets as $\mathcal{S}^{\text{train}} = \mathcal{S}^0(0,\tau_\star)$ and $\mathcal{S}^{\text{test}} = \mathcal{S}^0(\tau_\star, T)$ for some $\tau_\star \in (0,T)$. We then construct a counterfactual test set as $\bar{\mathcal{S}}^{\text{test}} = \mathcal{S}^\dagger(\tau_\star, T)$. Negative samples are generated in the transductive setting for both training and testing.

Let Y denote a performance metric of a predictive model. Let U represent the training condition, i.e., the dataset used for training, and let X denote the test dataset. We use counterfactual-style notation and write $Y_{X=x}(U=u) = y$, which is interpreted as: *a model trained under condition u and evaluated on dataset x achieves performance y*. For brevity, it is also written as $Y_x(u) = y$.

To instantiate this notation in our setting, we fix the training condition $u \equiv \mathcal{S}^{\text{train}}$ and consider two alternative test conditions, namely $x \equiv \mathcal{S}^{\text{test}}$ and $x' \equiv \bar{\mathcal{S}}^{\text{test}}$. With u fixed, the effect of data from another causal model is captured by the performance difference $Y_x(u) - Y_{x'}(u)$. We focus in particular on its sign, quantified by $\mathbb{P}\left(Y_x(u) - Y_{x'}(u) > 0\right)$, which aligns with $\mathbb{P}\left(\Delta_\star^{0,\dagger} > 0\right)$.

Fig. 6 illustrates the empirical relationship between the causal shift induced by replacing $\mathcal{S}^{\text{test}}$ with $\bar{\mathcal{S}}^{\text{test}}$ and the resulting performance difference $Y_x(u) - Y_{x'}(u)$, aggregated over multiple independently generated model pairs and sequence realisations.[14] In Fig. 6(a), we report the performance gap achieved by the `Oracle` under counterfactual evaluation. Although the overall trend becomes increasingly positive as the causal divergence grows, a small number of outliers persist. These deviations are expected, as negative sampling does not guarantee that all sampled negative events are infeasible, as discussed previously.



(a) Accuracy (b) AP (c) AUC

Figure 6: Counterfactual evaluation of TLP models. Each panel shows a scatter plot of the performance gap, overlaid with a `LOWESS` smoothing curve (fraction 0.95) to highlight the overall trend.

The remaining panels evaluate two TLP models[15], `TGN` (Rossi et al., 2020) and `JODIE` (Kumar et al., 2019). Fig. 6(b) reports results using average precision (AP), while Fig. 6(c) reports results using area under the ROC curve (AUC). The performance of `JODIE` remains largely flat and close to

10. TLP naturally fits within the positive-unlabelled learning (Mansouri and Ben-David, 2025) setting.
11. For each $(i,t) \in \mathcal{S}$, a negative event (j,t) is generated with $j \in [E] \setminus \{i\}$.
12. An event (i,t) is said to be feasible for a causal model if $\hat{f}(pa_i(t)) = 1$; see equation 3.
13. We denote the restriction of $\mathcal{S}$ to the time interval $[a,b)$ by $\mathcal{S}(a,b) = \{(i,t) \in \mathcal{S} : t \in [a,b)\}$.
14. Each point corresponds to one realisation of $(\mathcal{S}^0, \mathcal{S}^\dagger)$ and one trained predictor; the horizontal axis represents the causal divergence $\bar{d}_{0,\dagger}$, while the vertical axis denotes the performance gap $Y_x(u) - Y_{x'}(u)$.
15. Extending the benchmark to additional TLP models is left to future work.

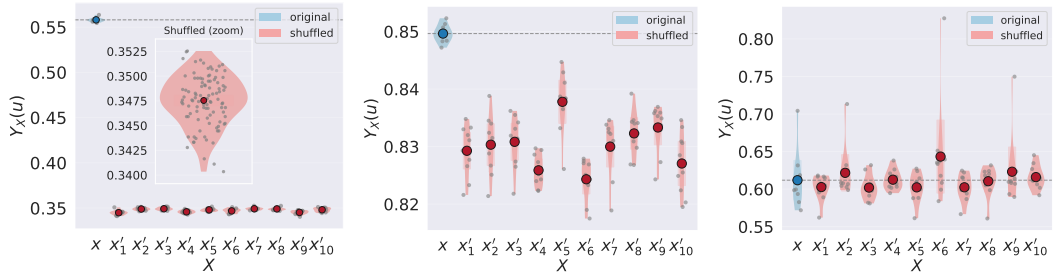
zero across increasing causal divergences, indicating limited sensitivity to causal distortions. In contrast, TGN exhibits a clear positive trend, with the performance gap increasing as the causal divergence grows, suggesting that it consistently performs better on the original test set than on its distorted counterpart. This behaviour supports the conclusion that TGN is causally sensitive under our counterfactual test, while JODIE is not[16].

For completeness, we also report the performance $Y_x(u)$ of the models on the original test set $\mathcal{S}^{\text{test}}$. The Oracle achieves a mean accuracy of 0.693 (95% CI $[0.678, 0.708]$), while TGN and JODIE achieve mean average precision scores of 0.763 (95% CI $[0.751, 0.775]$) and 0.563 (95% CI $[0.551, 0.575]$), respectively.

Experiment B We borrow the setup from the previous experiment. However, instead of constructing the counterfactual test set $\bar{\mathcal{S}}^{\text{test}}$ from an alternative causal model, we directly distort the original test sequence $\mathcal{S}^{\text{test}}$ by randomly shuffling event timestamps, thereby disrupting the temporal structure while preserving the set of observed interactions. We then repeat the same evaluation protocol as in Experiment A, training models on $\mathcal{S}^{\text{train}}$ and evaluating them on both the original and shuffled versions of $\mathcal{S}^{\text{test}}$. A similar counterfactual setup based on timestamp shuffling was previously considered by us in (Rahman et al., 2025b), albeit without access to an underlying causal model. In contrast, our setting allows test sequences to be generated from a known causal model and equips us with a principled divergence measure to quantify the extent to which shuffled sequences deviate from the original sequence.

Since shuffling is a stochastic procedure, we generate multiple shuffled realisations of the test data, denoted by x'_i for the ith realisation. The causal divergence between the original and shuffled sequences, as measured by the causal model $\mathfrak{C}_0$, has a mean value of 0.4322 with a 95% confidence interval of [0.4302, 0.4342]. This lets us treat shuffling as a quantified causal distortion rather than an ad hoc stress test.

In Fig. 7, we report violin plots of the achieved performance metrics $Y_X(u)$, where the horizontal axis corresponds to the test dataset X. A clear separation is observed between the metric values obtained on the original test data x and those obtained on the shuffled test sets $x'_1, \ldots, x'_{10}$, indicating a consistent degradation in performance under temporal distortion. A gap between $Y_x(u)$ and $Y_{x'_i}(u)$ is observed for the Oracle and TGN models, indicating sensitivity to temporal distortion, whereas no such separation is observed for JODIE. These results again support the conclusion that TGN passes the counterfactual test, while JODIE does not[17].



(a) Accuracy, Oracle (b) AP, TGN (c) AP, JODIE

Figure 7: Violin plots of model performance under timestamp shuffling.

16. The evaluation framework is model-agnostic and can be applied to any TLP model.
17. Prior work (Poursafaei et al., 2022, Fig. 6) identifies JODIE, TGN, and CAWN as among the most consistent models in terms of performance correlation. Additionally, in (Rahman et al., 2025b, Table 1), we have shown that TGN passes counterfactual tests on more datasets compared to CAWN. Based on these observations, we selected TGN and JODIE as representative models with contrasting behaviours for a proof-of-concept evaluation.

6. Conclusion

In this work we presented an algorithm to generate causal event sequences, and a metric to measure the divergence between parametric models generating them. We then extended the mechanism to temporal interaction graphs by proposing a construction in which node features induce a causal graph over edge events, yielding CTIGs with known ground-truth structure. Using this setup, we instantiated counterfactual evaluation in two complementary ways: by testing models across CTIGs generated from different causal models, and by applying timestamp shuffling as a stochastic temporal distortion whose severity can be quantified by the proposed divergence. In both settings, we observed degradation in predictive performance under causal distortion; in particular, TGN exhibits clear sensitivity to the induced distortions whereas JODIE remains largely flat. Our evaluation framework offers a controlled testbed for developing and assessing TLP models that can exploit underlying causal structure rather than superficial correlations. Furthermore, the same methodology can be extended to other predictive domains where the performance of a model under causal shifts is of interest.

Practical Usage We describe two practical usages of our proposed causal model:

- Given a TLP model F, the user generates two causal models A and B with divergence $d_{A,B}$, samples CTIGs $\mathcal{S}_A$ and $\mathcal{S}_B$ from each up to a horizon T, and splits both chronologically at τ_0. Training F on $\mathcal{S}_A(0, \tau_0)$ and testing on $\mathcal{S}_A(\tau_0, T)$ and $\mathcal{S}_B(\tau_0, T)$ gives performance measures Y_A and Y_B; the user expects $Y_A > Y_B$ once $d_{A,B}$ is large enough, since F is trained under A but tested on data from a different mechanism B. Repeating this across samples and causal shifts, defined as transitions in the generator's parameters (Θ, Λ), gives a scatter plot of $Y_A - Y_B$ against $d_{A,B}$, from which a cut-off, or a value satisfying $Y_A - Y_B > \epsilon$ for a user-chosen ϵ, is reported as the causal sensitivity of F. Comparing this cut-off across models, after accounting for differences in standalone performance, indicates which model makes greater use of causal information, assuming consistent negative sampling and a shared node set and interaction space between $\mathcal{S}_A$ and $\mathcal{S}_B$.
- Given a candidate causal model A representing a user's hypothesis about the mechanism generating an observed dataset, the structural equation f of A is used to predict the feasibility of an edge event given the history, and these predictions are evaluated against the real dataset. Since the true causal mechanism underlying real-world data is generally unknown, A can only ever be an approximate, hypothesised model rather than an exact one. Strong performance nonetheless suggests that A is a reasonable approximation of the true generative process, offering a way to validate causal assumptions against real data using the same framework.

Limitations Let $E = \mathcal{O}(n^2)$ denote the number of possible edges in a graph with n nodes. For each edge, the causal window captures a binary vector over all E edges, each either present or absent in the preceding window, giving 2^E possible configurations to review when learning its causal dependencies. Across all E edges, the resulting search space is $\mathcal{O}(E\,2^E) = \mathcal{O}(n^2\,2^{n^2})$, which grows rapidly with n. Predictive models therefore become more sample-hungry as n increases, requiring access to longer time horizons to learn the underlying causal mechanism. To present our framework, we use graphs with $n = 7$ nodes, and scaling to larger graphs remains a limitation of our current empirical setup. This also motivates the development of more compute-efficient causal models and causal discovery algorithms capable of handling larger node sets.

Acknowledgement

The authors would like to thank the anonymous reviewers for their valuable comments. They are also grateful to the chairs for handling the submission. The work of A. U. Rahman is supported by the Clarendon Scholarship Fund through Christ Church, University of Oxford, and Oxford University Press, whose support is gratefully acknowledged.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Albert-László Barabási. Network science book. *Network Science*, 625, 2014.

Joscha Cüppers, Sascha Xu, Ahmed Musa, and Jilles Vreeken. Causal discovery from event sequences by local cause-effect attribution. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Paul Erdos. On the evolution of random graphs. *Bulletin of the Institute of International Statistics*, 38:343–347, 1961.

Tianwei Gong, Tobias Gerstenberg, Ralf Mayrhofer, and Neil R. Bramley. Active causal structure learning in continuous time. *Cognitive Psychology*, 140:101542, February 2023. ISSN 0010-0285. doi: 10.1016/j.cogpsych.2022.101542.

Martin Haenggi. *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Amirkasra Jalaldoust, Kateřina Hlaváčková-Schindler, and Claudia Plant. Causal discovery in hawkes processes by minimum description length. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6978–6987, 2022.

Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology*, 7(3): e1001110, 2011.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1269–1278, 2019.

Farnam Mansouri and Shai Ben-David. Learning from positive and unlabeled examples-finite size sample bounds. *arXiv preprint arXiv:2507.07354*, 2025.

Kimia Noorbakhsh and Manuel Rodriguez. Counterfactual temporal point processes. *Advances in Neural Information Processing Systems*, 35:24810–24823, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. *Advances in Neural Information Processing Systems*, 35: 32928–32941, 2022.

Aniq Ur Rahman, Ahmed A Elhag, and Justin P Coon. A primer on temporal graph learning. *ACM Computing Surveys*, 58(5):1–28, 2025a.

Aniq Ur Rahman, Alexander Modell, and Justin Coon. Rethinking evaluation for temporal link prediction through counterfactual analysis. In *I Can't Believe It's Not Better: Challenges in Applied Deep Learning*, 2025b. URL https://openreview.net/forum?id=TKydQh6koc.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. *Advances in Neural Information Processing Systems*, 36:67686–67700, 2023.

Liang Zhao. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)*, 54(5):1–37, 2021.

A. Preliminaries

Definition 1 (*Point Process*, (Haenggi, 2012, § 2.2))**.** *A point process is a countable random collection of points that reside in some measure space, usually the Euclidean space $\mathbb{R}^d$. The associated σ-algebra consists of the Borel sets $\mathcal{B}^d$, and the measure is the Lebesgue measure.*

In simple terms, a point process is a countable random set $\Phi = \{x_1, x_2, \cdots\} \subseteq \mathbb{R}^d$ consisting of random variables $x_i \in \mathbb{R}^d$ as its elements. For a set $B \subseteq \mathbb{R}^d$, the cardinality of $\Phi \cap B$ is denoted as $N(B) \in \mathbb{N}$, and $N(\cdot)$ is called a counting measure.

Definition 2 (*1-D Poisson Point Process*, (Haenggi, 2012, § 2.4.1))**.** *The one-dimensional Poisson point process (PPP) with parameter $\lambda \in \mathbb{R}^+$ is a point process in $\mathbb{R}$ such that*

1. *for every bounded interval $[a, b)$, the number of points in the interval $N([a, b)]) \in \mathbb{N}$ has a Poisson distribution with mean $\lambda(b - a)$, i.e., $P(N([a, b)]) = k) = e^{-\lambda(b-a)} \frac{(\lambda(b-a))^k}{k!}$, and*
2. *if $[a_1, b_1), [a_2, b_2), \cdots, [a_m, b_m)$ are disjoint bounded intervals, then the number of points in those intervals $N([a_1, b_1)), \cdots, N([a_m, b_m))$ are independent random variables.*

We denote the samples drawn from a PPP with parameter λ as $\Phi(\lambda) \subset \mathbb{R}$, and $(\Phi(\lambda), <)$ satisfies the properties of a *total order*, i.e., $\Phi(\lambda) = \{t_1, t_2, \cdots, t_m\}$, such that $t_i < t_{i+1}, \forall i \in [m]$ for some $m \in \mathbb{N}^+$.

Definition 3 (*Causal model*, (Pearl, 2009, § 2.2, § 7.1))**.** *Let $\mathcal{U} = \{U_1, \cdots, U_n\}$ be a set of background variables determined by factors outside the model, and let $\mathcal{V} = \{X_1, \cdots, X_n\}$ be a set of endogenous variables that are determined by variables in the model, i.e., variables in $\mathcal{U} \cup \mathcal{V}$. A causal model is a pair $\mathfrak{M} = (\mathscr{G}, \Theta_{\mathscr{G}})$ consisting of a causal graph[18] $\mathscr{G}$ and a set of parameters $\Theta_{\mathscr{G}}$ compatible[19] with $\mathscr{G}$. The parameters $\Theta_{\mathscr{G}}$ assign a function $x_i = f_i(pa_i(t), u_i)$ to each variable $X_i \in \mathcal{V}$ and a probability measure $\mathbb{P}(u_i)$ to each u_i, where Pa_i are the parents of X_i and where each $U_i \in \mathcal{U}$ is randomly distributed according to $\mathbb{P}(u_i)$.*

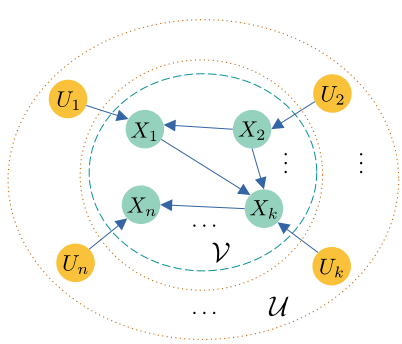


Figure 8: Causal model.

18. Causal graph is a directed graph where an edge from node X to Y indicates that X causes Y.
19. The elements in $\Theta_{\mathscr{G}}$ specify the functions $f_i : U_i \cup Pa_i \mapsto X_i, \forall i \in [n]$.

Definition 4 (*Monotonicity*, (Pearl, 2009, § 9.2))**.** *A variable Y is said to be monotonic relative to variable X in a causal model if and only if*

$$\mathbb{I}\{Y = 0 \mid do(X = 1)\} \cdot \mathbb{I}\{Y = 1 \mid do(X = 0)\} = 0.$$

Definition 5 (*Exogeneity*, (Pearl, 2009, § 7.4))**.** *In a causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ a variable X is exogenous relative to Y if X and Y have no common ancestor in $\mathscr{G}$.*

B. Related Works

Noorbakhsh and Rodriguez (2022) introduced a method to model event sequences by selectively removing events from a point process based on a probabilistic rule. However, their approach does not explicitly define causal relationships between different event types. The authors define a multi-dimensional Hawkes process (MHP) (Hawkes, 1971).

Jalaldoust et al. (2022) propose a causal discovery algorithm to learn the Granger-causal network underlying an MHP. For evaluation, the authors also generate an event sequence by sampling an MHP, where $\lambda_i(t)$ denotes the intensity function of events in the ith dimension.

$$\lambda_i(t) = \boldsymbol{\mu}_i + \textstyle\sum_{j \in pa_i(t)} \int_{-\infty}^{t} \phi_{ij}(t - \tau)\, dU_j(\tau). \tag{15}$$

Here, U_j denotes the counting process for the jth dimension, and $U_j(t)$ corresponds to the number of events in the jth dimension that have occurred before time t. Moreover, $\boldsymbol{\mu}_i$ is the underlying Poisson intensity, and ϕ_{ij} is the kernel function causally linking dimension i to dimension j, i.e., $pa_i(t) = \{j : \phi_{ij} \neq 0\}$.

Cüppers et al. (2024) introduce a causal model to generate event sequences wherein individual events of the cause *trigger* events of the effect with dynamic delays. The causal graph $\mathscr{G}$ consists of a set of source nodes $\mathcal{U}$, and effect nodes $\mathcal{V}$ with $\mathcal{U} \cap \mathcal{V} = \emptyset$. The model follows the assumption that an event of type $i \in \mathcal{U}$ (cause) occurring at time t *causes* an event of type $j \in \mathcal{V}$ (effect) at some time $t' \geq t$ with probability $\alpha_{i,j}$. It is further assumed that the events of type $i \in \mathcal{U}$ are sampled from a homogeneous Poisson process with intensity λ_i, i.e., $\mathcal{S}^i = \Phi(\lambda_i)$. Some events of type $j \in \mathcal{V}$ occur due to noise, described by another Poisson process with intensity λ_j, i.e., $\mathcal{W}_j = \Phi(\lambda_j)$.

The events in $\mathcal{S}^i$ result in delayed events in $\mathcal{S}^{i,j}$ which is constructed[20] as follows[21]:

$$\mathcal{S}^{i,j} = \left\{ v \cdot (t + d) + (1 - v) \cdot \xi : t \in \mathcal{S}^i, d \sim \text{Exp}(\Theta_{i,j}), v \sim \text{Bern}(\alpha_{i,j}) \right\} \setminus \{\xi\}. \tag{16}$$

Finally, all the events of type $j \in \mathcal{V}$ is defined as the set $\mathcal{S}^j$ (see (Cüppers et al., 2024, eq. 3)):

$$\mathcal{S}^j = \left(\bigcup_{i \in pa_j} \mathcal{S}^{i,j} \right) \cup \mathcal{W}_j, \tag{17}$$

where different events in pa_j work independently to cause events of type $j \in \mathcal{V}$, i.e., the structural equation groups the parent variables through a simple "or" operation. Most prior works assume that causes only increase the likelihood of an effect occurring. However, real-world systems often include inhibitory effects, where an event can reduce the likelihood of another event happening. For instance, Kim et al. (2011) explored cases where one event could either excite or suppress another. Our model incorporates both positive and negative influences, making it more flexible for diverse applications.

Table 1 highlights key limitations in existing literature and explains how our proposed causal model resolves them.

20. We have reframed (Cüppers et al., 2024, eq. 2) and its description as equation 16.
21. For practical purposes, ξ is a very large finite value which works as a proxy for ∞.

Table 1: Limitations in the literature and their resolution through our proposed causal model.

	Limitation	*Resolution*
Noorbakhsh and Rodriguez (2022):		
(*L1*)	The thinning procedure of a variable is independent of other variables.	We define the acceptance probability as a function of other variables, through structural equations.
Cüppers et al. (2024):		
(*L2*)	An event of type i at time t can at most cause a single event of type j at some time $t' > t$.	The influence of an event lasts for a finite duration within which it can cause an event i multiple times.
(*L3*)	An event is the result of a single cause, and a more realistic scenario of multiple causes working towards an effect is unexplored.	In the SEM equation 3 we can see that multiple events jointly cause an event.
Cüppers et al. (2024); Noorbakhsh and Rodriguez (2022); Jalaldoust et al. (2022):		
(*L4*)	The cause-effect relationships are strictly positive, i.e., a cause can only result in the occurrence of the effect.	In this work, we look at both positive and negative cause-effect relationships, i.e., excitatory and inhibitory.

C. Properties of the Causal Model

Summary We now summarise several key theoretical properties of the proposed causal model, the details of which can be found in the remainder of this appendix.

- The causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is
  - <u>semi-Markovian</u> for all $\mathbf{A} \in \{0,1\}^{n\times n}$, (Proposition 6).
  - <u>not Markovian</u> for all $\mathbf{A} \in \{0,1\}^{n\times n}$, (Proposition 7).
  - <u>Markovian</u> if $\mathbf{A} = \mathbf{I}$, (Corollary 8).
- $X_i(t)$ is <u>monotonic</u> relative to $X'_j(t)$ iff $\Theta_{i,j} \in [0,1]$, (Proposition 9).
- $X'_j(t)$ is <u>exogenous</u> relative to $X_i(t)\, \forall i,j \in [n]$, (Proposition 11).
- If for some $j \in pa_i$, $X'_j(t)$ exogenous relative to $X_i(t)$, and $X_i(t)$ is monotonic relative to $X'_j(t)$, then the probability of necessity and sufficiency (PNS) is <u>identifiable</u>, (Theorem 12).

Proposition 6. *The causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is <u>semi-Markovian</u> for all $\mathbf{A} \in \{0,1\}^{n\times n}$.*

Proof. The causal graph $\mathscr{G} \subset \{\mathcal{U} \times \mathcal{V}\} \cup \{\mathcal{V}' \times \mathcal{V}\}$ consists of two unidirectional bipartite subgraphs, one from $\mathcal{U}$ to $\mathcal{V}$, and another from $\mathcal{V}'$ to $\mathcal{V}$. Starting from any node, there is no way of reaching nodes in $\mathcal{U}$ or $\mathcal{V}'$. Moreover, starting from a node in $\mathcal{V}$ there is no way to return to it, as there are no edges directed away from the nodes in $\mathcal{V}$. Therefore, it is impossible to form a directed cycle in $\mathscr{G}$ allowing us to conclude that $\mathscr{G}$ is *acyclic*, and thus the causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is semi-Markovian (Pearl, 2009, § 3.2). □

Proposition 7. *The causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is <u>not Markovian</u> for all $\mathbf{A} \in \{0,1\}^{n\times n}$.*

Proof. In the proposed model $(\mathscr{G}, \Theta_{\mathscr{G}})$ the set $\mathcal{U} \cup \mathcal{V}'$ represents the background variables, while $\mathcal{V}$ constitutes the endogenous variables. For a model to be Markovian, it should be semi-Markovian, and in addition the background variables should be *independent* (Pearl, 2009, § 3.2). Therefore, the semi-Markovian model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is also Markovian if

(C1) $U_i(t) \perp\!\!\!\perp U_j(t)$, $\forall i,j \in [n]$, $i \neq j$,

(C2) $X'_i(t) \perp\!\!\!\perp U_j(t)$, $\forall i,j \in [n]$,

(C3) $X'_i(t) \perp\!\!\!\perp X'_j(t)$, $\forall i, j \in [n]$, $i \neq j$.

Verifying (C1): We express $U_i(t) \perp\!\!\!\perp U_j(t)$ as

$$\mathbb{I}\{t \in \Phi(\lambda_i)\} \perp\!\!\!\perp \mathbb{I}\{t \in \Phi(\lambda_j)\},$$

which is equivalent to verifying whether $\Phi(\lambda_i) \perp\!\!\!\perp \Phi(\lambda_j)$. Since the two PPPs are independent, (C1) holds *true*.

Verifying (C2): We express $X'_i(t)$ and $U_j(t)$ in terms of the PPPs:

$$\begin{aligned}
X'_i(t) &= \mathbb{I}\left\{\textstyle\sum_{\bar{t}\in(t-\bar{\tau},t)} X_i(\bar{t}) > 0\right\} \\
&= \mathbb{I}\left\{\sum_{\bar{t}\in(t-\bar{\tau},t)} f\left(\bigcup_{j\in pa_i(\bar{t})} \mathbb{I}\{|\Phi(\lambda_j)\cap(\bar{t}-\bar{\tau},\bar{t})| > 0\}, U_i(\bar{t})\right) > 0\right\}. \\
U_j(t) &= \mathbb{I}\{t \in \Phi(\lambda_j)\} = \mathbb{I}\{\lim_{\epsilon\downarrow 0} |\Phi(\lambda_j)\cap[t, t+\epsilon)| > 0\}.
\end{aligned}$$

The value of $X'_i(t)$ depends upon $\Phi(\lambda_k) \cap (t - 2\bar{\tau}, t)$, $\forall k \in pa_i$ and $\Phi(\lambda_i) \cap (t - \bar{\tau}, t)$. Similarly, the value of $U_j(t)$ depends upon $\Phi(\lambda_j) \cap [t, t + \epsilon)$ for some $\epsilon \downarrow 0$. Since there is no overlap between $(t - 2\bar{\tau}, t)$ and $\lim_{\epsilon\downarrow 0}[t, t + \epsilon)$, (C2) holds *true*.

Verifying (C3): From the previous paragraph, we can say that $X'_i(t)$ depends on $\Phi(\lambda_k) \cap (t - 2\bar{\tau}, t)$, $\forall k \in pa_i(t)$ and $\Phi(\lambda_i) \cap (t - \bar{\tau}, t)$, and $X'_j(t)$ depends on $\Phi(\lambda_m) \cap (t - 2\bar{\tau}, t)$, $\forall m \in pa_j(t)$ and $\Phi(\lambda_j) \cap (t - \bar{\tau}, t)$.

Intuitively, this means that each $X'_i(t)$ is influenced by the recent activity of its parent events and its own recent occurrences within their respective temporal windows. For $X'_i(t)$ and $X'_j(t)$ to be independent, their corresponding sets of influencing events must not overlap at any time t. In other words, their dependency domains must remain disjoint over time.

Therefore, $X'_i(t) \perp\!\!\!\perp X'_j(t)$ only if

$$\begin{aligned}
&\{pa_i(t) \cup \{i : X'_i(t) = 1\}\} \cap \{pa_j(t) \cup \{j : X'_j(t) = 1\}\} = \emptyset, \quad \forall i \neq j, \forall t \in \mathbb{R}^+ \\
&\implies \{pa_i \cup \{i\}\} \cap \{pa_j \cup \{j\}\} = \emptyset, \quad \forall i \neq j \\
&\implies pa_i \subseteq \{i\}, \forall i \in [n].
\end{aligned}$$

The implication follows under the assumption that, for some t, the time-varying parent sets $pa_i(t)$ and $pa_j(t)$ coincide with their structural counterparts pa_i and pa_j, and that the indicators $X'_i(t) = 1$ and $X'_j(t) = 1$ can be simultaneously satisfied. This condition effectively requires that the sets of potential causal influencers remain completely non-overlapping. However, such disjointness is highly restrictive and can only occur for $\mathbf{A} = \mathbf{I}$. Therefore, we conclude that (C3) is *not true* for all realizations of $\mathbf{A}$.

Since (C3) does not always hold, we conclude that the causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is not Markovian for all $\mathbf{A}$. □

Corollary 8. *The causal model $(\mathscr{G}, \Theta_{\mathscr{G}})$ is* <u>*Markovian*</u> *if $\mathbf{A} = \mathbf{I}$.*

Proposition 9. *$X_i(t)$ is* <u>*monotonic*</u> *relative to $X'_j(t)$ iff $\Theta_{i,j} \in [0, 1]$.*

Proof. From def. 4, we can say that $X_i(t)$ is monotonic relative to $X'_j(t)$ if and only if

$$\mathbb{I}\big\{X_i(t) = 0 \mid do\big(X'_j(t) = 1\big)\big\}\, \mathbb{I}\big\{X_i(t) = 1 \mid do\big(X'_j(t) = 0\big)\big\} = 0.$$

The above equation must hold for all subsets $\mathcal{C}_i \subseteq pa_i \setminus \{j\}$, i.e.,

$$\mathbb{I}\left\{\Theta_{i,j} + \sum_{k\in\mathcal{C}_i} \Theta_{i,k} < 0\right\} \mathbb{I}\left\{\sum_{k\in\mathcal{C}_i} \Theta_{i,k} \geq 0\right\} = 0, \quad \forall \mathcal{C}_i \subseteq pa_i \setminus \{j\}. \tag{18}$$

Case 1: $\sum_{k\in\mathcal{C}_i}\Theta_{i,k}\geq 0$. To ensure that equation 18 holds, we require

$$\Theta_{i,j}+\sum_{k\in\mathcal{C}_i}\Theta_{i,k}\geq 0.$$

Taking the intersection over all such $\mathcal{C}_i$, we obtain

$$\begin{aligned}\Theta_{i,j}&\in\bigcap_{\mathcal{C}_i:\sum_{k\in\mathcal{C}_i}\Theta_{i,k}\geq 0}\left[-\sum_{k\in\mathcal{C}_i}\Theta_{i,k},\,1\right]\\&=\left[\max\left\{-\sum_{k\in\mathcal{C}_i}\Theta_{i,k}:\sum_{k\in\mathcal{C}_i}\Theta_{i,k}\geq 0,\,\forall\mathcal{C}_i\subseteq pa_i\setminus\{j\}\right\},1\right].\end{aligned}$$

For $\mathcal{C}_i=\emptyset$ we have $\sum_{k\in\mathcal{C}_i}\Theta_{i,k}=0$, yielding

$$\Theta_{i,j}\in[0,1].$$

Case 2: $\sum_{k\in\mathcal{C}_i}\Theta_{i,k}<0$. In this case, $\mathbb{I}\left\{\sum_{k\in\mathcal{C}_i}\Theta_{i,k}\geq 0\right\}=0$, so equation 18 is automatically satisfied for all $\Theta_{i,j}$. Hence, these subsets impose no additional restriction beyond the assumed domain $\Theta_{i,j}\in[-1,1]$.

Intersecting the results from both cases gives $\Theta_{i,j}\in[0,1]$, which completes the proof. □

Corollary 10. *If for some $i\in[n]$, $\Theta_{i,j}\in[0,1],\forall j\in pa_i$ then the variable $X_i(t)$ is monotonic relative to all $X'_j(t),\forall j\in pa_i$. However, the structural equation is reduced to $x_i(t)=u_i(t)$ and the causal impact of $X'_j(t)$ on $X_i(t)$ vanishes*[22] *$\forall j\in pa_i$.*

Proposition 11. *$X'_j(t)$ is <u>exogenous</u> relative to $X_i(t)\,\forall i,j\in[n]$.*

Proof. Firstly, for all $\mathbf{A}\in\{0,1\}^{n\times n}$ the subgraph with edges $\mathcal{V}'\times\mathcal{V}$ is a unidirectional bipartite graph. Secondly, $X'_j(t)\in\mathcal{V}'$ has no ancestors. Therefore, $X_i(t)$ and $X'_j(t)$ have no common ancestor $\forall i,j\in[n]$ which concludes the proof (see def. 5). □

Theorem 12 (Identifiability)**.** *If for some $j\in pa_i$ $X'_j(t)$ exogenous relative to $X_i(t)$, and $X_i(t)$ is monotonic relative to $X'_j(t)$, then the probability of necessity and sufficiency* (PNS) *is <u>identifiable</u>, and determined through:*

$$\mathsf{PNS}_{i,j}=\mathbb{P}\left(X_i=1\mid X'_j=1\right)-\mathbb{P}\left(X_i=1\mid X'_j=0\right).$$

Proof. For the detailed proof, please see (Pearl, 2009, § 9.2.3). □

D. Causal Parameter Generation

For the experiments in Fig. 5, causal models are instantiated by randomly sampling the following parameters. We fix $n=7$ nodes and a time horizon $T=10^3$, giving $E=\binom{n}{2}=21$ possible edges. Node feature vectors $\mathbf{x}_a\in\mathbb{R}^r$ are sampled with dimensionality $r\sim\text{UnifInt}(5,20)$. The number of non-causal edges is set to $l\sim\text{UnifInt}\left(1,\left\lfloor\frac{E}{5}\right\rfloor\right)$, ensuring that at most one-fifth of all edges are designated as noise. The scaling parameter in the influence function h is drawn as $\nu_0\sim\text{Unif}(0,200)$, and the sparsity threshold as $\nu_1\sim\text{Unif}(0,1)$. Finally, the PPP rate for each edge $i\in[E]$ is sampled as $\lambda_i\sim\mathcal{N}(\mu_\lambda,\sigma_\lambda^2)$, where $\mu_\lambda\sim\text{Unif}(1,4)$ and $\sigma_\lambda\sim\text{Unif}(0.1,2)$ are themselves randomly drawn for each causal model instantiation.

Although we have not conducted a dedicated parameter-wise sensitivity analysis, the experiments in Fig. 5 already sample causal models under this randomised procedure, and the observed trends, namely larger performance gaps at higher causal divergence, are consistent across these realisations.

22. To rule out trivial feasibility, we verify that $|S|\neq|\Phi|$ for all datasets (equality $|S|=|\Phi|$ would indicate degeneracy).