Quantifying Causal Contribution in Rare Event Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce a framework for causal discovery and attribution of causal influence for rare events in time series data–where the interest is in identifying causal links and root causes of individual discrete events rather than the types of these events. Specifically, we build on the theory of temporal point processes, and describe a discrete-time analogue of Hawkes processes to model the occurrence of self-exciting rare events with instantaneous effects. We then introduce several scores to measure causal influence among individual events. These statistics are drawn from causal inference and temporal point process theories, describe complementary aspects of causality in temporal event data, and obey commonly used axioms for feature attribution. We demonstrate the efficacy of our model and the proposed influence scores on real and synthetic data.

1 Introduction

The field of causal inference studies causal links among random variables of interest, disentangling causal effects from simple statistical associations [25, 27]. For example, quantifying the causal effects of a medical treatment on patient outcomes concerns two primary random variables—treatment and outcome—potentially along with other covariates to consider. In causal discovery, the aim is to recover causal links among finitely many well-defined random variables from which a finite sample is observed. However, many causal questions in real world applications take on a different form that do not appeal to these descriptions. Many applications in *root cause analysis* comprise singular discrete events that unfold in time, and the objective is to recover causal links and chains among these individual events [40]. For example, in system administration and operations (recently, *AIOps*) it is often required to establish root causes of some adverse events such as failures and outages to other events in the data such as deployments and failures in dependencies. In the study of electronic health records, one may be interested in causally tracing changes in a patient's trajectory to treatments. These examples can be viewed as establishing causal links among individual rare events unfolding in time.

In multivariate event streams, where events can be identified as members of finitely many types, these questions extend to whether one type of event Granger-causes another [1, 8]. However, there exists no framework for defining this problem in the language of causal discovery and for attribution of causal effects among *individual events* as opposed to *types of events*. We aim to address this problem in this work, paving the way to a unified and consistent methodology for identifying root causes in event streams.

In this paper, our objectives are twofold. We will first introduce a novel time series model for rare "event" data where occurrences of events will be represented as binary random variables in discrete time. Our model is inspired by the rich theory on temporal point processes (TPP) [7] and self-exciting point processes [12, 13]; and will serve to represent event data in a statistical framework that is amenable to causal analysis. We will then use this model to utilize the tools of time-series causal

Submitted to *A causal view on dynamical systems* workshop at NeurIPS 2022. Do not distribute.

inference and discovery, introducing two measures of causal contribution among events that are analogous to causal effects as defined by Pearl[25]. Finally, we will link these quantities to existing results on Hawkes processes and Granger causality. Together, our model and causal attribution scores constitute a framework for fitting rare event processes and attributing causal influence among individual events.

2 Preliminaries

Causal Inference Causal inference focuses on drawing causal conclusions from data, establishing some variables as the *causes* of others as opposed to simply recovering associational relationships (*i.e.*, correlations) among them. One formalism often used in causal inference is the *causal Bayesian network* (CBN) [25, 30], which is a Bayesian network that obeys the causal Markov condition, *i.e.*, that the joint distribution of random variables $X_1, X_2, \cdots, X_N$ decomposes as

$$P(x_1, x_2, \cdots, x_N) = \prod_i P(x_i | PA_i = pa_i),$$

where PA_i denotes the set of variables X_j that are the parents of X_i in the CBN, and pa_i the corresponding random variates. Moreover, each conditional $P(x_i|PA_i = pa_i)$ represents an independent *causal* mechanism. That is, to obtain the *interventional* distribution $P(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_{N-1}|do(X_j = x_j))$ it suffices to replace the term $P(X_j|PA_j)$ with δ_{X_j, x_j} where $\delta_{a,b}$ denotes Kronecker's delta. Note that this distribution is different from the conditional that would result from simply observing that $X_j = x_j$, and corresponds to the distributions of $\{X_i | i \neq j\}$ when X_j is actively determined (*i.e.*, intervened on).

Structural Causal Models While CBNs suffice to completely specify all possible interventional distributions of a set of variables, a stricter formalism is needed to answer so called *counterfactual* queries that allow answering "what if?" questions for individual observations. Under *structural causal models* (SCMs, also referred to as functional causal or structural equation models), every variable is written as a function of its parents and an unobserved noise variable, $X_i = f_i(PA_i, U_i)$, where U_i are statistically mutually independent. $f_i(PA_i, U_i)$ is an SCM for the conditional $P(X_i|PA_i)$ if $f_i(pa_i, U_i)$ is distributed according to $P(X_i|PA_i = pa_i)$ for almost all pa_i.

Granger causality Time ordering of data significantly facilitates reasoning about causal relations, as causal effects can only act forward in time. However, variables measured *simultaneously* in time, up to the temporal granularity available, still present an issue as the causal ordering among these variables are not determined [27, Ch. 10]. However, in the absence of causal influence among simultaneous measurements of variables, or so called *instantaneous effects*, causal influence can be captured using the formalism of *Granger causality* [11]. Let $\{\mathbf{X}_t\}_{t=1}^T$ denote a discrete-time vector-valued stochastic process where $\mathbf{X}_t = [X_t^{(1)}, X_t^{(2)}, \cdots, X_t^{(d)}]$. A time-series $X^{(i)}$ is said to Granger-cause another time series $X^{(j)}$ if the past of $X^{(i)}$ improves the predictions of $X^{(j)}$ given all past information about $\{X^{(j')} | j' \neq i\}$.

Temporal Point Processes A TPP specifies the full generative model for random sequences of points $(t_1, t_2, \cdots, t_n)$ on a bounded subset of the real line, where $0 < t_1 < t_2 < \cdots \leq T$ and the variable n is also random [7]. The *conditional intensity* function of the TPP

$$\lambda^*(t)dt = \mathbb{P}\{\text{next event is in } [t, t + dt) | \mathcal{H}_t\},$$

completely determines the process and is often used to characterize TPP models. Here $\mathcal{H}_t$ denotes the history (filtration) up to time t—specified by the set of points up to time t, $\{t_i | t_i < t\}$. Intuitively, the conditional intensity function specifies the arrival rate of events per unit time, in the infinitesimal interval after t.

Hawkes process [13, 12]. A (univariate) Hawkes process is given by the conditional intensity[1]

$$\lambda^*(t) = \mu + \sum_{t_i < t} \alpha \varphi(t - t_i). \tag{1}$$

Here $\mu > 0$ is a *background intensity*—the arrival rate of events if previous events had no effect on the present. The *delay density* $\varphi(s)$ determines the temporal profile of interactions between points—with $\int \varphi(s)ds = 1$ w.l.o.g. Moreover, the function φ is always "causal," $\varphi(s) = 0, \forall s < 0$, nonnegative $\varphi(s) \geq 0$, and is often monotonically decreasing over all $s > 0$. The parameter $\alpha > 0$ is the so-called *infectivity* or *branching* parameter.

Multivariate TPPs model *marks* $y_i \in \{1, \cdots, d\}$ for each event t_i. That is, *events* are now observed as ordered pairs (t_i, y_i). Practically, y_i often represent membership to an entity, such as a user on a social network, host on a computer network, etc. The conditional intensity of the multivariate Hawkes process (MHP) is written separately for each mark k as

$$\lambda_k^*(t) = \mu_k + \sum_m \sum_{t_i < t | y_i = m} \alpha_{km} \varphi(t - t_i).$$

Note that the background intensity of each mark is now different, and the infectivity parameters can now be arranged along a matrix $\mathbf{A}_{km} = \alpha_{km}$, where each element describes the directional *infectivity* of one mark over the other. Eichler *et al.* [8] show that $\alpha_{km} > 0$ implies that the process m Granger-causes k; while Achab *et al.* have shown how to recover $\mathbf{A}$ via moment-matching estimators [1]. While Hawkes processes are defined in continuous time ($t_i \in \mathbb{R}$), in this paper we will explore their discrete time analogues ($t_i \in \mathbb{N}$) starting from the next section.

3 Quantifying Causal Contribution with Discrete-Time Hawkes Processes

Basic Observations General TPPs model a wide range of occurrence patterns such as self-excitation [12], self-inhibition [16], quasi-periodicity [6], etc. for discrete events in continuous time. However, much of the established literature in causal inference deals with a finite set of random variables as opposed to continuous time processes, leading to conceptual difficulties in analyzing cause-effect relationships in continuous-time stochastic processes. Similarly, in many application domains time is inherently *quantized*, *i.e.*, the data is sampled in discrete time—events can often "co-occur" with no temporal ordering implied among them—and a continuous-time process serves as an approximation. For example, neural spike trains are recorded with finite sampling rates, or many rare events in computer systems logs are recorded in a predetermined time resolution. Therefore, in this section, we start with the introduction of a discrete-time analogue of self-exciting temporal point processes which will serve primarily to reconcile notation between causal inference and TPPs, as well as having the added benefit of removing any statistical bias that results from using continuous-time models for discrete data.

Discrete-Time Hawkes Processes In our formalism, the occurrences of "events"[2] or "points" are interpreted as those times $t \in \mathbb{Z}_{>0}$ of X_t where $X_t = 1$. Such models have been called discrete-time point processes[38], such as in determinantal point processes [19] or discrete-time renewal processes [9]. In addition to modeling discretely sampled events, our model builds on Hawkes processes to model excitation patterns among them. We introduce the discrete-time Hawkes process (DTHP) below.

Definition 1. *(Discrete-time Hawkes Process (DTHP)) A binary-valued stochastic process* $\{X_t \in \{0, 1\}\}_{t \in \mathbb{Z}_{>0}}$ *is a discrete-time Hawkes process if, for all* t*,*

$$p_t := \mathbb{P}\{X_t = 1 | X_{1:t-1}\} = 1 - \exp\left(-\mu - \sum_{s=1}^{t-1} X_s g(t - s)\right).$$

where $g(\tau) : \mathbb{Z} \to \mathbb{R}_{\geq 0}$ *is a nonnegative function that satisfies* $g(\tau) = 0, \forall \tau < 0$*.*

We observe that for all $s < t$, $\mathbb{E}[X_t | X_s = 1] > \mathbb{E}[X_t | X_s = 0]$, therefore the process preserves the self-excitation property of Hawkes processes, *i.e.*, that events only increase the probability of future event occurrences.

[2]Not to be confused with the events of the underlying probability space, we reserve this term exclusively to refer to occurrences of 1 in a discrete-time binary process.

Our construction of DTHP admits the continuous-time Hawkes process as a limit case, *i.e.*, it tends to a continuous-time Hawkes process as events become "infinitely" rare. In the same light, we can examine a *rare event limit* or how the probability of events behaves as events become increasingly rare. We will use these limits to derive approximations to the true causal effects that are expressed simply in terms of the learned parameters of our model. Concretely, bounding the probability of occurrence of points such that $\forall t, p_t \leq \bar{p}$ we observe as $\bar{p} \to 0$, these probabilities also admit a linear approximation in the effects of past points.

Proposition 1. $p_t = \mu + \sum_{s=1}^{t-1} X_s g(t-s) + O(\bar{p}^2)$ *as* $\bar{p} \to 0$.

Another benefit of casting event occurrences in discrete time is that it enables the use of concepts from traditional time-series analysis and the well-established literature of causal inference; specifically causal Bayesian networks [25], and causality in time series [26, 27]. Moreover, in order to set our new model in this framework, we can write an SCM that results in joint distributions equivalent to the DTHP, which follows from observing that each X_t can be written as a function of an independent source of noise and parent variables $X_{1:t-1}$.

Definition 2. *(DTHP SCM) Let* $\{X_t\}$ *are determined by the structural equations*

$$X_t = [\![U_t \leq \lambda(X_{1:t-1})]\!] \qquad \textit{where } \lambda(X_{1:t-1}) = \mu + \alpha \sum_{s<t} X_s g(t-s),$$

U_t *are independent standard exponential random variables and* $[\![.]\!]$ *denotes the indicator function.*

Apart from rendering the mathematical objects conceptually simpler, DTHP enables using the language of causal graphical models. Note that our choice of $1 - \exp(-x)$ as a link function in Definition 1 is one of many possible that would yield similar and tighter approximations. However, for the purposes, this function suffices to demonstrate the key links between self-exciting point processes and measures of causal contribution.

Multivariate DTHP We can now extend the DTHP to multivariate processes, where the interest is in multiple related types of events.

Definition 3. *(Multivariate DTHP) A binary vector valued process* $\mathbf{X}_t = (X_t^{(1)}, \cdots, X_t^{(d)}) \in \{0,1\}^d$ *is a multivariate DTHP if for all* k, t

$$p_t^{(k)} := \mathbb{P}\{X_t^{(k)} = 1 | \mathbf{X}_{1:t-1}\} = 1 - \exp\left(-\mu^{(k)} - \sum_{m=1}^{d} \sum_{s=1}^{t-1} X_s^{(m)} g_{m\to k}(t-s)\right).$$

Here, $g_{m\to k}$ determine the decay profile of effects of events in type m on events of type k. In the remainder of this paper, we will assume a more specific form for this quantity, $g_{m\to k}(t-s) = \mathbf{A}_{km} g(t-s)$ where $\mathbf{A} \in \mathbb{R}^{d\times d}$ and we assume $\sum_{\tau=1}^{\infty} g(\tau) = 1$ without loss of generality.

We can also rely on previous results in time series causal discovery [27] to repeat a result similar to those of [8] and [1] for DTHP.

Proposition 2. *Events* $\{X_t^{(m)}\}$ *Granger-cause events* $\{X_t^{(k)}\}$ *if and only if* $\mathbf{A}_{km} > 0$.

In the discrete-time world, however, we encounter another conceptual difficulty: continuous-time TPPs are built on the *simplicity* assumption [7], that specify that no two points can co-occur on the same point $t' \in \mathbb{R}$ almost surely. This is a somewhat restrictive requirement in discrete time where one may be interested in multiple types of points occurring together while being causally related, *i.e.*, via *instantaneous* effects. Our formulation in Definition 3 disallows any such interactions between 'simultaneous' variables $X_t^{(m)}$ and $X_t^{(k)}$. In order to incorporate such effects for more realistic modeling, we can extend the model as follows. For brevity, we denote

$$\lambda_t^{(k)} = \lambda^{(k)}(\mathbf{X}_{1:t-1}) := \sum_{m=1}^{d} \sum_{s=1}^{t-1} X_s^{(m)} \mathbf{A}_{km} g(t-s),$$

and define

$$p_t^{(k)} = 1 - \exp\left(-\lambda_t^{(k)} - \sum_{X^{(m)} \in PA_k^{(B)}} \mathbf{B}_{km} X_t^{(m)}\right),$$

where we define $\mathbf{B} \in \mathbb{R}^{d \times d}$ as the weighted adjacency matrix of a graph that specifies the instantaneous causal effects among types of events, and $PA_k^{(B)}$ to denote the set of parents of $X^{(k)}$ along this graph.

Quantifying Causal Contribution We can now build on the DTHP to introduce our method for quantifying causal influence among observed events themselves. Specifically, we will focus on quantifying causal contributions given a fitted multivariate DTHP model (Definition 3), where we will currently ignore instantaneous effects for notational brevity. However, extensions of our arguments to the case with instantaneous effects and implications for continuous-time Hawkes processes can be derived from our framework.

Our problem can be formulated as follows. Given a finite realization of the process $\{\mathbf{X}_t = \mathbf{x}_t\}_{t=1}^T$, we seek to quantify the causal contribution of $X_s^{(m)}$ on $X_t^{(k)}$, where $s < t$, and when such events are "rare." In practice, such quantification is only relevant when $x_s^{(m)} = x_t^{(k)} = 1$ as, under our model, events are assumed to be mutually exciting and we do not intuitively expect that the absence of an event is the cause of another.

Such a notion of causal influence, of $X_s^{(m)}$ on $X_t^{(k)}$, can be built on several familiar quantities in causal inference. For example, one could consider the *average causal effect* $\text{ACE}(X_s^{(m)} \to X_t^{(k)}) = \mathbb{E}\left[X_t^{(k)} \mid \text{do}(X_s^{(m)} = 1)\right] - \mathbb{E}\left[X_t^{(k)} \mid \text{do}(X_s^{(m)} = 0)\right]$, measuring the added probability of an event on $X_t^{(k)}$ when $X_s^{(m)}$ is intervened on [15]. However, this quantity disregards the fact that the entire history $\mathbf{X}_{1:T}$ is observed. Moreover, ACE also does not take into account how (marginally) rare the target event $\{X_t^{(k)} = 1\}$ is. In this light, we define our first measure of causal influence on a different quantity, the *direct effect* [25, Sec 4.5], which refers to the isolated effect of changing only a single parent $X_s^{(m)}$ having observed all other parents of $X_t^{(k)}$. We will denote this quantity $\text{DE}(X_s^{(m)} \to X_t^{(k)})$, defined

$$\mathbb{E}\left[X_t^{(k)} \mid \text{do}(X_s^{(m)} = 1, \mathbf{X}_{\neg(m,s)} = \mathbf{x}_{\neg(m,s)})\right] - \mathbb{E}\left[X_t^{(k)} \mid \text{do}(X_s^{(m)} = 0, \mathbf{X}_{\neg(m,s)} = \mathbf{x}_{\neg(m,s)})\right],$$

where the notation $\mathbf{X}_{\neg(m,s)}$ is used to refer to all variables in the history except $X_s^{(m)}$.

We can now show that under the DTHP SCM and in the rare event regime, the direct effect yields a convenient approximation. Namely,

Proposition 3. *DE*$(X_s^{(m)} \to X_t^{(k)}) = \mathbf{A}_{km} g(t-s) + O(\bar{p}^2)$.

This result links the proposed contribution measure to a well known quantity in the analysis of Hawkes processes, namely the incremental intensity due to a previous event in the Hawkes process, *i.e.*, the summand in $\lambda^{(k)}(\mathbf{X}_{1:t-1})$ due to $X_s^{(m)}$.

The direct effect is based on a "total" intervention on all of the parents of $X_t^{(k)}$, comparing the intervention where there is a source event at $X_s^{(m)}$ to one where there is not. In this sense, it already takes into account the full information available ($\mathbf{x}_{1:t}$). However, it is still scaled in terms of the marginal probability of $\{X_t^{(k)} = 1\}$. In order to quantify the proportion of influence of each past event on a given target event, we can define a normalized quantity.

Definition 4 (Normalized Direct Effect)**.** *The normalized effect is defined*

$$\widetilde{DE}(X_s^{(m)} \to X_t^{(k)}) = \frac{DE(X_s^{(m)} \to X_t^{(k)})}{\lambda_k^t}.$$

Note that, $\widetilde{\text{DE}}(X_s^{(m)} \to X_t^{(k)})$ is exactly equivalent to the posterior "parent" distribution in the immigration-birth representation of Hawkes processes [14, 3]. Indeed, this representation of Hawkes processes captures an intuitive notion of a causal chain of events.As previously indicated, we expect that in an unconfounded system, the causes of events can only be (a combination of) other events, but not the lack thereof. Similarly, we are more rarely interested in causal questions such as "what previous event caused the lack of an event at time t?" In this sense, the immigration-birth process

naturally captures an intuitive notion of causality among events. Our results show that the influence of a direct cause, in the direct parenthood sense of a Hawkes process, is analogous to the direct effect in causal inference. We describe this link in detail in Appendix B. Finally, we observe that for the normalized direct effects to add to one, a summand $\mu^{(k)}/\lambda_t^{(k)}$ is also required. This quantity can be thought of as the probability that an event has no observed causal parent. In the immigration-birth interpretation, the same quantity can be understood as the probability that an event is an "immigrant," and not a descendant of any previous events. Finally, the following result links Granger causality to our approximate contribution measure $\mathrm{DE}(X_s^{(m)} \to X_t^{(k)})$.

Proposition 4. *Assume $\forall \tau, g(\tau) > 0$. Then, DE$(X_s^{(m)} \to X_t^{(k)}) > 0$ if and only if $X^{(m)}$ Granger-causes $X^{(k)}$.*

We can build on these observations to define a *total effect* of a single event at $X_s^{(m)}$ on an event at $X_t^{(k)}$, by summing over all indirect paths of influence, weighted by their normalized direct effects. In the following, let $\mathcal{B}_{s,t}$ define the set of all points $\{X_{s'}^{(k')} = 1 | s < s' < t, k' \in \{1, \cdots, d\}\}$, and $\mathcal{P}_o(\mathcal{B}_{s,t})$ all ordered sets in the power set of $\mathcal{B}_{s,t}$ such that temporal ordering is preserved. In other words,

$$\mathcal{P}_o(\mathcal{B}_{s,t}) = \{(X_{s_1}^{k_1}, \cdots, X_{s_n}^{k_n}) | \forall n \in [|\mathcal{B}_{s,t}|], s_i < s_{i+1} \forall i\}.$$

Note that the empty ordered set $\varnothing \in \mathcal{P}_o(\mathcal{B}_{s,t})$. For brevity, let us also define the path effect

$$\widetilde{\mathrm{DE}}((X_{s_1}^{k_1}, \cdots, X_{s_n}^{k_n})) := \prod_{i=1}^{n-1} \widetilde{\mathrm{DE}}(X_{s_i}^{k_i} \to X_{s_{i+1}}^{k_{i+1}}).$$

We heuristically define the total effect as,

Definition 5 (Total Effect)**.** *The total effect TE$(X_s^{(m)} \to X_t^{(k)})$ is defined*

$$\sum_{\mathbf{Z} \in \mathcal{P}_o(\mathcal{B}_{s,t})} \widetilde{DE}((X_s^{(m)}, \mathbf{Z}, X_t^{(k)})), \tag{2}$$

where we use the notation $(X_s^{(m)}, \mathbf{Z}, X_t^{(k)})$ to denote the sequence generated by prepending (resp. appending) $X_s^{(m)}$ (resp. $X_t^{(k)}$) to the sequence $\mathbf{Z}$.

The simple intuition behind our definition is hidden away by the cumbersome notation required. In other terms, the total effect captures the total influence an event has on a descendant, summing over all paths of descendance—direct or indirect. While (2) seemingly requires summing over exponentially many paths, its computation can be greatly accelerated via simple heuristics such as dropping connections below a certain NDE.

Proposition 4 highlights that one event can be the cause of another in our sense of DE only if there is a Granger-causality relationship between their marks. Note, however, that the same is not true for our definition of total effects, where one mark can indirectly cause events in another mark. To understand this relationship, and to contrast the two measures, assume a multivariate Hawkes process of three marks is used to represent "delay" events of three consecutive trains, where the delay of the first train directly causes a delay in the second, and a delay in the second causes a delay of the third train. Using our measures of influence, and with perfect information, we will always attribute direct causation to the previous train only. However, through total effects, we can attribute the third train's delay to that of the first. Finally note that, in the sense of Granger causality, the first train cannot be said to cause the delay of the third train as, given knowledge of the second train, we cannot better predict the delay of the third train. Although we will not make a rigorous argument in this work, the DE measure, when viewed as an attribution method, readily satisfies the axioms of [33].

Finally, let us highlight that methods proposed in this section can be viewed as the parts of a single framework. Given sparse event data that are sampled in discrete time and can be identified as one of finitely many types, our framework only makes the additional assumption that past events will have linear and additive (self-exciting) effects on future events. Under these assumptions, to identify the causal effects among individual events we (i) fit a DTHP model to the observed sequence and (ii) use DE, NDE and TE as measures of causal contribution to trace individual events to their causal parents.

4 Related Work

The Hawkes process has been studied commonly to establish Granger causality—i.e., causal links among different types of events as opposed to individual events. Eichler *et al.* explore the link between Hawkes process infectivity kernels and Granger causality [8]. Achab *et al.* use this link and previous results on moment-matching methods for Hawkes process estimation to introduce a fast algorithm for uncovering Granger causality [1]. Xu *et al.* consider group sparsity regularization for a more precise recovery of the Granger causal graph [39]. Notably, Prabhakar *et al.* introduced an algorithm for Granger-causal discovery directly from a cross-spectral estimate of multivariate TPPs, without making any parametric assumptions on the form of the conditional intensity [28]. We also refer the reader to [34] for a discussion of causal discovery in multitype event sequences. Many other works, which focus on more effective methods for recovering the infectivity matrix of a Hawkes process, can be seen as causal discovery algorithms in the context of multitype event sequences. Among these we can cite [21] who use an EM algorithm for better stability, [10] who work with more general transmission models, [35] who employ Bayesian inference for more accurate recovery of the graph, and [37] who employ low rank factorizations for improved scalability.

To our knowledge, a "discretized" Hawkes process appears only in [22], who allow each time step to have more than one points—*i.e.*, work with time series of positive integers instead of binary sequences. Other discrete time point processes, towards recovering Granger-causal structure, have also been introduced in the context of neural structure learning [17].

Sun and Janzing study a similar form for causal discovery in arbitrary causal graphs of binary variables, although their setting is more general and their methods do not address temporal data [32]. Similar to our framework, their probabilities of occurrence also admit linear approximations around 0, although the authors do not explore this direction. In [5], Budhatoki *et al.* discuss methods for root cause analysis of outlier values, which could be regarded as rare events.

Recently, Tran *et al.* introduced QTree [36], a method that draws from extreme value theory and causal inference to infer graphs (more specifically, root-directed trees) of causal influence among nodes where *simultaneous* outlier events occur jointly. The max-linear Bayesian network model used is able to handle missing values as well as infer graphs of influence among network nodes in a robust fashion. Moreover, the authors employ the Chu-Liu-Edmonds algorithm for minimum cost arborescence to heuristically recover root-directed trees, as required by their application in uncovering hidden river networks. CAUSE, by Zhang *et al.* , is the closest to our work [40]. Here, the authors consider an axiomatic causal attribution method that obeys the axioms of [33]. Notably, the method considered attributes causal influence among events, using an "explainable" recurrent point process—a neural TPP model. The authors then show that an aggregation of these influence scores can be interpreted as a measure for Granger causality among event marks. However, the neural network-based model used and the attribution methods make computation under this method prohibitively costly. Finally, in "counterfactual" TPPs [23], Noorbakhsh and Gomez-Rodriguez, describe a structural causal model analogue of Lewis' thinning algorithm which they then use to answer counterfactual queries in observed point sequences.

5 EXPERIMENTS

Model Performance We start by validating the performance of DTHP on three data sets for the task of inferring the latent network of influence among event types—the first step of the framework we propose. We compare the performance of DTHP with two baselines: QTree [36] and CAUSE [40]. To contrast these baselines with ours, QTree is able to handle missing values and can work with general real-valued variables to infer both general graphs and trees of influence. However, QTree only works with instantaneous effects, *i.e.*, assumes that each time step is i.i.d. CAUSE [40] is based on neural TPPs and does not consider instantaneous effects. Both algorithms are developed for causal discovery in sequences of rare events. For both baselines, we use repositories made available by the authors and keep the original hyperparameters included in the libraries.[3]

[3] see `https://github.com/razhangwei/CAUSE`, `https://github.com/princengoc/QTree`.

Metric	**AUC**			**F1**		
Model	QTree	CAUSE	DTHP (ours)	QTree	CAUSE	DTHP (ours)
hawkes-1	0.248	0.431	**0.830**	0.114	0.286	**0.471**
hawkes-2	0.563	0.610	**0.765**	0.265	0.254	**0.467**
hawkes-3	0.563	0.536	**0.736**	0.255	0.242	**0.424**
danube	**0.897**	0.628	0.841	**0.800**	0.118	0.308
lower-colorado	**0.712**	0.639	0.701	**0.450**	0.200	0.214
middle-colorado	**0.951**	0.734	0.563	**0.909**	0.286	0.235
upper-colorado	**0.931**	0.570	0.660	**0.875**	0.267	0.333
Connectomics-1	0.499	0.525	**0.623**	0.185	0.186	**0.234**
Connectomics-2	0.519	0.514	**0.639**	0.179	0.178	**0.243**
Connectomics-3	0.590	0.508	**0.670**	0.206	0.175	**0.267**
Connectomics-4	0.702	0.527	**0.738**	0.301	0.185	**0.348**
Connectomics-5	0.730	0.515	**0.745**	0.320	0.186	**0.357**
Connectomics-6	0.859	0.715	**0.880**	0.487	0.307	**0.545**

Table 1: Experiment results comparing QTree, CAUSE, and DTHP algorithms given in AUC and maximum F1 scores (higher better). Top scores in each row are given in bold.

The objective of all experiments is the recovery of an underlying causal graph from observed time series. We use three groups of data sets, the first of which is simulated, and the others taken from real applications. Further details on synthetic data generation and benchmark data sets are given in Appendix C.

- We simulate data from **continuous-time multivariate Hawkes processes** using tick [2].
- The **River Basin Data Sets** include data collected from two river basins in Europe and the US [36], for the so-called *hidden river* discovery task. We experiment with four data sets, belonging to the Danube, as well as lower, middle and upper sections of the Lower Colorado river basin.
- We use the neural connectome data set from the **Chalearn Connectomics** challenge [4]. The data set includes realistically generated spike trains from neuronal networks [31] Specifically, we perform experiments on the **small** data sets which are numbered in increasing order from the most challenging setting to the least.

All data sets have ground truth causal networks available. For the river basin data sets, we use raw measurements in the QTree algorithm, however threshold the data to convert it into binary time series for use in CAUSE and DTHP models. For the neural connectome data sets, we threshold each data set at the 99th percentile, obtaining binary time series used in all of the algorithms. We use both versions of the QTree algorithm, with and without the minimum cost arborescence step, and report the best results. As CAUSE is designed for continuous time data sets, we "dequantize" binary time series by adding random noise drawn from a uniform distribution between 0 and 1 to each timestamp.

Our results are presented in Table 1. We report the area under the ROC curve (AUC) and maximum attained F1 score for edge classification. As expected, our model is significantly superior in the Hawkes process data sets, and the QTree algorithm dominates in the river basin data sets where it was designed to perform well. Both CAUSE and DTHP perform significantly below the QTree baseline in the river data sets. We believe this is primarily due to two reasons. First, neither model performs Bayesian treatment of missing values which is especially important in the Lower Colorado river basin data sets. Second, these algorithms do not search for the best tree with minimum cost arborescence. Let us note, however, that running the Chu-Liu-Edmonds algorithm alone on graphs recovered by DTHP and CAUSE also did not yield significantly better results. Still, DTHP appears to perform slightly more favorably than CAUSE, which does not address instantaneous effects.

In the Connectomics experiments, we find that our algorithm significantly outperforms baselines. This matches our expectation as the Connectomics data set is both high-dimensional (100 marks), and features both delayed and instantaneous effects. Our model is the only one designed to capture all such patterns simultaneously. Overall, we can conclude that DTHP generally yields favorable performance in modeling sparse binary time series where instantaneous effects occur.

Causal Influence Scores For a demonstration of our causal influence scores, we present a set of experiments on synthetic data. Here, our aim is to first exhibit the general difficulty of attributing

d	p_E	True	Fitted
5	0.05	0.735 ± 0.156	0.604 ± 0.156
	0.1	0.717 ± 0.062	0.641 ± 0.081
	0.2	0.672 ± 0.069	0.607 ± 0.070
	0.5	0.503 ± 0.063	0.473 ± 0.058
10	0.05	0.826 ± 0.097	0.600 ± 0.135
	0.1	0.793 ± 0.046	0.600 ± 0.077
	0.2	0.722 ± 0.045	0.569 ± 0.047
	0.5	0.528 ± 0.038	0.449 ± 0.033

Table 2: Comparison of results when retrieving the Hawkes process parent event using normalized direct effect. Numbers reported are means and standard deviations of recall at top 1—*i.e.*, among events with known parents, the ratio of those with the top NDE score assigned to the correct parent. d denotes the dimensionality of the Hawkes process, and p_E is the prior for sparsity.Higher p_E implies lower sparsity.

causal influence among rare events, even with perfect information. To this end, we draw from multivariate Hawkes processes while keeping record of the parents of each event. We regard these parenthood relationships as the ground truth causes of events, and measure if the direct effects computed as per Definition 4 correctly recover the causes. We consider only those events that have a parent in the branching process, and compute recall (at top 1).

Results, for varying dimensionality and degrees of sparsity in the infectivity matrix, are presented in Table 2. Here, we observe that direct effects computed with known parameters already fall to an accuracy of around 50% when 50% of the edges in the ground truth graph are active—highlighting a general ambiguity with assigning causes among events when many such causes are possible. Moreover, we find that causal attribution with fitted parameters ("Fitted") performs slightly worse than when ground truth parameters are known ("True"), but also that it is relatively robust. However, as expected, robustness decreases when dimensionality is increased. Further details are available in Appendix C.

6 CONCLUSION

In this paper we introduced a framework for attributing causal influence among individual events observed in time. Assuming only that events are sparse and there is a quasi-linear and monotonic relationship among their probabilities of occurence our method proceeds by fitting a newly introduced discrete-time process model, and performing causal attribution via simple quantities based on the fitted parameters of this model Our analysis was cast in a discrete-time framework, enabling unbiased estimation in many real-world scenarios where data is sampled with finite rates and instantaneous effects are also present. Finally, our numerical experiments validate the efficacy of our model for the unique scenarios it addresses, as well as the intuition behind the causal contribution metrics we proposed in this work. While our method can address many discrete event scenarios, its main limitation is that it only allows excitation relationships among events. Several directions remain as next steps to our work, such as extending the model with real-valued marks and inhibitory effects to address more general sparse discrete event sequences.

References

[1] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. *The Journal of Machine Learning Research*, 18(1):6998–7025, 2017.

[2] Emmanuel Bacry, Martin Bompaire, Philip Deegan, Stéphane Gaïffas, and Søren Poulsen. tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *J. Mach. Learn. Res.*, 18(1):7937–7941, 2017.

[3] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

[4] Demian Battaglia, Isabelle Guyon, Vincent Lemaire, Javier Orlandi, Bisakha Ray, and Jordi Soriano. *Neural connectomics challenge*. Springer, 2017.

[5] Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International Conference on Machine Learning*, pages 2357–2369. PMLR, 2022.

[6] David Roxbee Cox. *Renewal theory*. Methuen, 1962.

[7] Daryl J. Daley and David Vere-Jones. *An introduction to the theory of point processes: Volume I: elementary theory and methods*. Springer Science & Business Media, 2007.

[8] Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

[9] Willliam Feller. *An introduction to probability theory and its applications*. John Wiley & Sons, 1957.

[10] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[11] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[12] Alan G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 438–443, 1971.

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

[14] Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

[15] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[16] Valerie Isham and Mark Westcott. A self-correcting point process. *Stochastic processes and their applications*, 8(3):335–347, 1979.

[17] 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.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

[20] Patrick J. Laub, Thomas Taimre, and Philip K. Pollett. Hawkes Processes. *arXiv:1507.02822 [math, q-fin, stat]*, July 2015. arXiv: 1507.02822.

[21] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421, 2014.

[22] Scott W. Linderman and Ryan P. Adams. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

[23] Kimia Noorbakhsh and Manuel Gomez Rodriguez. Counterfactual temporal point processes. In *Neural Information Processing Systems*, 2019.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[25] Judea Pearl. *Causality: Models, reasoning, and inference.* Cambridge University Press, 2000.

[26] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.

[27] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[28] Karthir Prabhakar, Sangmin Oh, Ping Wang, Gregory D Abowd, and James M Rehg. Temporal causality for the analysis of visual events. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1967–1974. IEEE, 2010.

[29] Aleksandr Simma and Michael I. Jordan. Modeling events with cascades of Poisson processes. *arXiv preprint arXiv:1203.3516*, 2012.

[30] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[31] Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Computational Biology*, 8(8), 2012.

[32] Xiaohai Sun and Dominik Janzing. Exploring the causal order of binary variables via exponential hierarchies of markov kernels. In *15th European Symposium on Artificial Neural Networks (ESANN 2007)*, pages 465–470. D-Side, 2007.

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[34] Nikolaj Theodor Thams. Causal structure learning in multivariate point processes. Master's thesis, University of Copenhagen, 2019.

[35] Long Tran, Mehrdad Farajtabar, Le Song, and Hongyuan Zha. Netcodec: Community detection from individual activities. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 91–99. SIAM, 2015.

[36] Ngoc Mai Tran, Johannes Buck, and Claudia Klüppelberg. Causal discovery of a river network from its extremes. *arXiv preprint arXiv:2102.06197*, 2021.

[37] Ali Caner Türkmen, Gökhan Çapan, and Ali Taylan Cemgil. Clustering event streams with low rank hawkes processes. *IEEE Signal Processing Letters*, 27:1575–1579, 2020.

[38] Ali Caner Türkmen, Tim Januschowski, Yuyang Wang, and Ali Taylan Cemgil. Forecasting intermittent and sparse time series: A unified probabilistic framework via deep renewal processes. *Plos one*, 16(11):e0259764, 2021.

[39] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726. PMLR, 2016.

[40] Wei Zhang, Thomas Panum, Somesh Jha, Prasad Chalasani, and David Page. Cause: Learning granger causality from event sequences using attribution methods. In *International Conference on Machine Learning*, pages 11235–11245. PMLR, 2020.

A Proofs of Propositions

Proposition 1

Proof. Let $\lambda_t := \mu + \sum_{s=1}^{t-1} X_s g(t-s)$. Note the Taylor series approximation of p_t around 0 is,

$$\bar{p} \geq p_t = \lambda_t - \frac{\lambda_t^2}{2} + O(\lambda_t^3).$$

and also note that $\bar{p}^2 \sim \frac{\lambda_t^2}{2}$ as $\bar{p} \to 0$. Therefore $p_t = \lambda_t + O(\bar{p}^2)$. □

Proposition 2

Proof. We will follow the arguments of [27, Theorem 10.3], assuming causal sufficiency (as required by Granger causality in general). By definition, there exists a link between $X^{(m)}$ and $X^{(k)}$ in the *summary graph* only when there exists a link from $X_s^{(m)}$ to $X_t^{(k)}$ for some $s < t$. However, by definition of Hawkes SCM (Definition 2, extended analogously to the multivariate case), such a link only exists if $\mathbf{A}_{km} > 0$. □

Proposition 3

Proof. Let

$$\lambda_t^{k,0} := \mu^{(k)} + \sum_{m', s' < t | (m', s') \neq (m, s)} x_{s'}^{(m')} \mathbf{A}_{km'} g(t - s'). \tag{3}$$

It follows from Proposition 1 that

$$\begin{aligned}
\mathrm{DE}(X_s^{(m)} \to X_t^{(k)}) &= \mathbb{E}\left[X_t^{(k)} \mid \mathrm{do}(X_s^{(m)} = 1, \mathbf{X}_{\neg(m,s)} = \mathbf{x}_{\neg(m,s)})\right] \\
&\quad - \mathbb{E}\left[X_t^{(k)} \mid \mathrm{do}(X_s^{(m)} = 0, \mathbf{X}_{\neg(m,s)} = \mathbf{x}_{\neg(m,s)})\right], \\
&= \lambda_t^{k,0} + \mathbf{A}_{km} g(t-s) + O(\bar{p}^2) - (\lambda_t^{k,0} + O(\bar{p}^2)) \\
&= \mathbf{A}_{km} g(t-s) + O(\bar{p}^2).
\end{aligned}$$

□

Proposition 4

Proof. From Proposition 3 and 2, this immediately holds for an approximation of direct effects $\mathrm{DE}(X_s^{(m)} \to X_t^{(k)}) \approx \mathbf{A}_{km} g(t-s)$. To see that it also holds exactly, let $\lambda_t^{k,0}$ be defined as in (3) and note that

$$\begin{aligned}
\mathrm{DE}(X_s^{(m)} \to X_t^{(k)}) &= \mathbb{E}\left[X_t^{(k)} \mid \mathrm{do}(X_s^{(m)} = 1, \mathbf{X}_{\neg(m,s)} = \mathbf{x}_{\neg(m,s)})\right] \\
&\quad - \mathbb{E}\left[X_t^{(k)} \mid \mathrm{do}(X_s^{(m)} = 0, \mathbf{X}_{\neg(m,s)} = \mathbf{x}_{\neg(m,s)})\right], \\
&= \exp\left(-\lambda_t^{k,0}\right) - \exp\left(-\lambda_t^{k,0} - \mathbf{A}_{km} g(t-s)\right),
\end{aligned}$$

from where it is apparent that $A_{km} = 0$ implies $\mathrm{DE}(X_s^{(m)} \to X_t^{(k)}) = 0, \forall s, t$. Conversely, assuming $g(t-s) > 0$, $A_{km} = 0$ implies $\mathrm{DE}(X_s^{(m)} \to X_t^{(k)}) = 0$ completing the proof. □

B Equivalence to Hawkes' Branching Process Interpretation

Owing to the convenient additive form of its intensity, the Hawkes process lends itself to interpretation as a Poisson-cluster process, or an infinite cascade of Poisson processes. This description of the process is sometimes intuitively called an *immigration-birth* or *branching* representation[14, 7]. Below, we describe a *new* generative process, one which does not rely on the conditional intensity as in (1). Here, individual points will be denoted as ordered pairs (s_n, z_n) where s_n denotes the timestamp, and z_n the timestamp of the parent event which gave *birth* to the point at s_n.

1. Draw $N_0 \sim \text{Poisson}(\mu \times T)$. Let $\mathcal{D}_0 = \{(s_i, 0)\}_{i=1}^{N_0}$ where s_i are drawn uniformly at random in (0, T]. These points are the so-called *immigrants*.
2. For each generation j, starting from $j = 1$ we draw the children of each point in the previous generation.
   - Letting $\mathcal{D}_{j-1} = \{(s_i^{(j-1)}, z_i^{(j-1)})\}$, draw $N_{s_i^{(j-1)}} \sim \mathcal{PO}(\alpha)$ for each $s_i^{(j-1)}$
   - Let
   $$\mathcal{D}_j^{s_i^{(j-1)}} = \{(\tau_k + s_i^{(j-1)}, s_i^{(j-1)})\}_{k=1}^{N_{s_i^{(j-1)}}},$$
   where we draw $\tau_k \sim g$ i.i.d.
   - Let $\mathcal{D}_j = \bigcup_{s_i^{(j-1)}} \mathcal{D}_j^{s_i^{(j-1)}}$.
   - Stop if there exist no $(s_i^{(j)}, z_i^{(j)}) \in \mathcal{D}_j$ such that $s_i^{(j)} \leq T$.
3. Return $\mathcal{D} = \left\{(s_i, z_i) \in \bigcup_j \mathcal{D}_j | s_i \leq T\right\}$.

Somewhat surprisingly, due to the Poisson superposition property, this process is equivalent to the process determined by the conditional intensity function of (1). Moreover, if one uses this method of generating a Hawkes draw, an auxiliary *parenthood* variable, z_i which refers to the (timestamp of) point which "gave birth" to it, s.t. $z_i < s_i$ always holds. Moreover, if these parenthood variables were known from the beginning, optimal parameters $\{\mu, \alpha, g\}$ could be recovered in a closed-form maximization step since they would just be parameters of iid Poisson process observations.

The discarded parenthood variables z_i define a *forest* of immigrants (root nodes) and their descendants. It is this observation that underlies the EM algorithm for Hawkes processes [14, 3, 29, 20], which proceeds by (E) inferring the parent of each variable (computing $\mathbb{P}\{z_i = s_j\}$ where $s_j < z_i$), and (M) maximizing $\{\mu, \alpha, g\}$ under the expected complete data likelihood. By consulting [29], for example, one can see our approximate normalized direct effect (for the univariate case) $\alpha g(t_i - t_j)/\lambda_t$ appears as the "posterior" probability $\mathbb{P}\{z_i = t_j\}$. While our exposition here is concerned only with the univariate Hawkes process, its extensions to multivariate processes follow easily.

Using the same statistical foundation as above we can now argue that our approximated normalized direct effects coherently describe a graph where each node is a point and each edge is weighted by the probability of parenthood. In this formalism, our definition of the total effect also appears as the total path weight where a path weight is defined as the product of the weights of edges it is composed of.

C Further Details on Experiments

C.1 Model Performance

Generated Hawkes processes Data sets are generated with the `SimuHawkesExpKernels` class provided in tick [2]. Namely, we generate infectivity matrices $\mathbf{A} = \mathbf{W} \odot \mathbf{Y}$ where $\mathbf{A} \in \mathbb{R}^{d \times d}$ $\odot$ denotes the Hadamard product, $\mathbf{W}_{km} \overset{iid}{\sim} \text{Exp}(1)$, and $\mathbf{Y} \overset{iid}{\sim} \text{Bernoulli}(0.1)$. We then adjust the spectral radius of the matrix to ρ. We set the baseline intensities $\mu_k = 0.05$, and the maximum number of jumps to 5000. The three data sets hawkes-1, hawkes-2, and hawkes-3 are sampled with parameters $(\rho, d) = (0.5, 10), (0.4, 20), (0.3, 30)$ ranked from least to most challenging respectively. We then binarize these data sets by quantizing time along the unit grid and setting a time interval to 1

if the interval contains a sampled point. The resulting data sets have points in 5.5%, 8.2%, and 6.8% of intervals respectively.

Lower Colorado River Basin Data Sets Except for use in the QTree algorithm, the data sets are preprocessed by binarizing at the 0.99-quantile and filling missing values with 0.

Connectomics Data Set We first preprocess the data set by taking the first difference of the raw action potentials. Except for QTree, we binarize the data by setting a cutoff at at the 99th percentile. In practice, this percentile is also close to the recommended binarization cutoff, 0.12.

Baselines and Hyperparameters We use the `ExplainableRecurrentPointProcess` class from the CAUSE library, and use the default hyperparameters as defined in the training script. By default, the model uses a hidden layer size of 64, embedding dimension of 64, batch size of 64, no dropout or L2 regularization, learning rate of 0.001, 200 epochs and the Adam optimizer. We use the `QTree` class of the QTree library, leaving default hyperparameters `smallR = 0.05, q = 0.8`.

Implementation of DTHP We use our own implementation for the DTHP model, using PyTorch [24]. We implement maximum likelihood optimization for the proposed discrete time model, with added regularization for the graph such that the total loss function is

$$\ell(\mu, \theta, \mathbf{A}) = \log \sum_{k,t} p(X_t^{(k)} | \mathcal{H}_t, \mu, \theta, \mathbf{A}) + \gamma ||\mathbf{A}||_F.$$

In our experiments, we heuristically set $\gamma = 10$. We use the implementation of the Adam optimizer [18] implemented in PyTorch for optimization, setting the learning rate to 0.01. We train for 10K epochs on the Connectomics data set, and 5K epochs on the other data sets. In practice, we truncate the history of each point where influences can flow to a certain maximum history, and set this value to 1 in the river data sets and 5 in the synthetic and connectome data sets.

C.2 Causal Influence

For the causal influence estimation experiments, we generate infectivity matrices $\mathbf{A} = \mathbf{U} \odot \mathbf{Y}$ where $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathbf{U}_{km} \overset{iid}{\sim} \text{Uniform}[0, 1)$, and $\mathbf{Y}_{km} \overset{iid}{\sim} \text{Bernoulli}(p_E)$, set $\mu = (2d)^{-1}$, and $\theta = 0.33$. We use our own implementation of a Hawkes process branching sampler to draw from a Hawkes process while retaining the parent identifiers z_i as explained in Appendix B.

For experiments where the infectivity matrix $\mathbf{A}$ is estimated (denoted "Fitted" in the results), we run DTHP setting maximum lag to 5 and the number of epochs to 3K.