Counterfactual Temporal Point Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Machine learning models based on temporal point processes are the state of the art in a wide variety of applications involving discrete events in continuous time. However, these models lack the ability to answer counterfactual questions, which are increasingly relevant as these models are being used to inform targeted interventions. In this work, our goal is to fill this gap. To this end, we first develop a causal model of thinning for temporal point processes that builds upon the Gumbel-Max structural causal model. This model satisfies a desirable counterfactual monotonicity condition, which is sufficient to identify counterfactual dynamics in the process of thinning. Then, given an observed realization of a temporal point process with a given intensity function, we develop a sampling algorithm that uses the above causal model of thinning and the superposition theorem to simulate counterfactual realizations of the temporal point process under a given alternative intensity function. Simulation experiments using synthetic and real epidemiological data show that the counterfactual realizations provided by our algorithm may give valuable insights to enhance targeted interventions.

1 Introduction

In recent years, machine learning models based on temporal point processes have become increasingly popular for modeling discrete event data in continuous time [1, 2]. This type of data is ubiquitous in a wide range of application domains, from social and information networks to finance or epidemiology. For example, in social and information networks, events may represent users' posts, clicks or likes [3]; in finance, they may represent buying and selling orders [4]; or, in epidemiology, they may represent when an individual gets infected or recovers [5]. In many of these domains, these models have become state of the art at predicting future events given a sequence of past events [6].

Building upon the above models, a recent line of work [7–9] has developed machine learning methods to automate online, adaptive targeted interventions using reinforcement learning and stochastic optimal control. While this line of work has shown early promise, particularly in personalized teaching and viral marketing, there are many high-stakes application in which targeted interventions are unlikely to be automated. For example, in epidemiology, fine-grained interventions that are targeted at particular sites or individuals (*e.g.*, hygienic measures at work places, closures of schools, or contact tracing) are likely to be decided by governments, policy makers and health authorities, at least in the foreseeable future. In this work, our goal is to develop machine learning methods that, given the outcome of an intervention implemented in the past, are able to assist decision makers at implementing better interventions in these high-stakes applications.

More specifically, we focus on facilitating counterfactual thinking, a type of thinking that has been argued to help humans correct and improve behavior that has been unsuccessful in the past [10, 11]. In counterfactual thinking, given a history of past events that have already occurred, one asks what past events would have instead occurred if certain intervention had been in place. For example, in epidemiology, assume that, during a pandemic, a government decides to implement business

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

restrictions every time the weekly incidence—the (relative) number of new cases—is larger than certain threshold but unfortunately the incidence nevertheless spirals out of control. In this case, counterfactual thinking would help the government understand retrospectively to what extent the incidence would have grown had a lower threshold been implemented[1].

Our contributions. Our starting point is Lewis' thinning algorithm [18], one of the most popular techniques for sampling in the temporal point process literature. Lewis' thinning algorithm first samples a sequence of potential events from a temporal point process with a constant intensity that upper bounds the intensity of the temporal point process of interest. Then, it accepts each of these events with probability proportional to the ratio between the intensity of the temporal point process of interest at the time of the sampled event and the constant intensity. In our work, the key idea is to augment the above thinning process using a particular class of structural equation models (SCMs)—the Gumbel-Max structural causal model [19]. This causal model satisfies a desirable monotonicity condition and, given a sequence of accepted and rejected events, it allows us to reliably estimate what events would have been accepted and rejected under an alternative intensity. Put differently, it can be used to answer counterfactual questions about a set of previously accepted and rejected events by the thinning algorithm.

Unfortunately, the above causal model on its own is not sufficient to answer counterfactual questions about observed sequence of real events. This is because, in general, real event data is not *generated* by a thinning process and, as a consequence, it does not *include* rejected events. However, we are able to overcome this limitation by using the superposition theorem [20] to sample *plausible* sequences of events that the thinning process would have rejected if it had accepted the observed sequence of real events. Then, these generated sequences of rejected events, together with the observed events, can be fed into the above causal model of thinning to sample counterfactual events given an alternative intensity. Importantly, while our causal model of thinning only allows for inhomogeneous Poisson processes (*i.e.*, it requires the intensities of interest to be deterministic), we can still use it to sample counterfactual events from linear Hawkes processes [21], a popular type of temporal point processes with stochastic self-exciting intensities, by carefully exploiting their branching process interpretation [22].

Finally, we evaluate our sampling algorithm using both synthetic and real epidemiological data and show that the counterfactual events provided by our algorithm can give valuable insights to enhance targeted interventions. To facilitate research in this area, we will release an open-source implementation of our algorithms and data with the final version of the paper.

Further related work. The literature on temporal point processes related to causal inference has mostly focused on measuring causal influence by means of, *e.g.*, Granger causality [23–26], integrated cumulants [27, 28] or Wold processes [29, 30], and on predicting quantities related to an interventional distribution of interest [31–33] (*e.g.*, conditional average treatment effect (CATE)). However, there exist a few notable exceptions, which have focused on counterfactual reasoning [34–36]. The work by Schulam and Saria [34] focuses on reasoning about the counterfactual distributions of the marks in a marked temporal point process, rather than reasoning about the counterfactual intensities of the events as we do. The works by Ryalen et al. [35] and Roysland [36] focus on survival analysis, *i.e.*, temporal point processes that terminate after one event, and are very different to ours at a technical level. In contrast, we focus on temporal point processes with multiple events.

More broadly, the literature on causal inference has a long and rich history [37]. However, most of this literature has used counterfactual reasoning to predict quantities related to the interventional distribution of interest such as, *e.g.*, the conditional average treatment effect (CATE). A few recent notable exceptions are by Oberst and Sontag [19] and Tsirtsis et al. [38], which have used the Gumbel-Max structural causal model to reason about counterfactual distributions in Markov decision processes (MDPs), and by Buesing et al. [39] and Pitis et al. [40], which have used counterfactuals to improve the training of reinforcement learning agents. However, to the best of our knowledge, the Gumbel-Max structural causal model has not been used previously to reason about counterfactual events in temporal point processes.

[1]Note that existing epidemiological models [12], including those developed in the context of COVID-19 [13–17], are unable to answer such counterfactual questions—they can only predict what the (average) future would look like under certain interventions given the past.

2 Preliminaries

In this section, we first briefly revisit the frameworks of temporal point processes [41] and structural causal models [42].

Temporal point processes. A temporal point process is a stochastic process whose realization consists of a sequence of discrete events localized in continuous time, $\mathcal{H} = \{t_i \in \mathbb{R}^+ \mid i \in \mathbb{N}^+, t_i < t_{i+1}\}$. A temporal point process can be equivalently represented as a counting process, $N(t)$, which records the number of events before time t. Moreover, in an infinitesimally small time window dt around time t, it is assumed that only one event can happen, *i.e.*, $dN(t) \in \{0, 1\}$.

A temporal point process is typically characterized via its intensity function $\lambda(t) \geq 0$, which determines the probability of observing an event in $[t, t + dt)$, *i.e.*,

$$\lambda(t)dt = \mathbb{P}\{dN(t) = 1\} = \mathbb{E}[dN(t)]. \tag{1}$$

In general, the intensity $\lambda(t)$ may depend on the history $\mathcal{H}(t) = \{t_i \in \mathcal{H} \mid t_i < t\}$ up to time t and its functional form is often chosen to capture the phenomena of interest. Throughout the paper, we will consider inhomogeneous Poisson processes, *i.e.*, $\lambda(t) = g(t)$, where $g(t) \geq 0$ is a time-varying function, and linear Hawkes processes, *i.e.*,

$$\lambda(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} g(t - t_i), \tag{2}$$

where $\mu \geq 0$ and the second term, with $\alpha \geq 0$, $g(t) \geq 0$ and $g(t) = 0$ for all $t < 0$, denotes the influence of previous events on the current intensity[2].

Structural causal models. Given a set of random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, a structural causal model (SCM) $\mathcal{C}$ defines a complete data-generating process via a collection of assignments

$$X_i := f_i(\mathbf{PA}_i, U_i), \tag{3}$$

where $\mathbf{PA}_i \subseteq \boldsymbol{X} \backslash X_i$ are the direct causes of X_i, $\boldsymbol{U} = \{U_1, \ldots, U_n\}$ are noise variables, and $P(\boldsymbol{U})$ denotes the (prior) distribution of the noise variables. Here, note that, given an observational distribution $P(X_1, \ldots, X_n)$, there always exists a distribution for the noise variables and functions f_i so that $P = P^{\mathcal{C}}$, where $P^{\mathcal{C}}$ is the distribution entailed by $\mathcal{C}$. In general, the noise variables $\boldsymbol{U}$ may not be jointly independent if there are unmeasured confounders[3]. However, in our work, we will assume there are no unmeasured confounders and $\boldsymbol{U}$ are jointly independent.

Given a SCM $\mathcal{C}$, we can express (atomic) interventions $\mathcal{I}$ using the *do-operator*, *e.g.*, $\mathcal{I} = \text{do}(X_i = x)$ corresponds to replacing the causal mechanism $f_i(\mathbf{PA}_i, U_i)$ with x. The intervened SCM is typically denoted as $\mathcal{C}^{\mathcal{I}}$ and the interventional distribution entailed by the intervened SCM as $P^{\mathcal{C}\,;\,\mathcal{I}}$. Moreover, given a SCM $\mathcal{C}$ and an observed realization of assignments $\boldsymbol{X} = \boldsymbol{x}$, we can define a counterfactual SCM $\mathcal{C}_{\boldsymbol{X}=\boldsymbol{x}}$ where the noise $\boldsymbol{U}$ variables are distributed according to the posterior distribution $P(\boldsymbol{U} \mid \boldsymbol{X} = \boldsymbol{x})$ and not necessarily jointly independent anymore. Counterfactual statements can now be seen as interventions in a counterfactual SCM $\mathcal{C}_{\boldsymbol{X}=\boldsymbol{x}}$ and, given an intervention $\mathcal{I}$, we denote the interventional counterfactual distribution entailed by $\mathcal{C}^{\mathcal{I}}_{\boldsymbol{X}=\boldsymbol{x}}$ as $P^{\mathcal{C} \mid \boldsymbol{X}=\boldsymbol{x}\,;\,\mathcal{I}}$. However, the posterior distribution of the noise variables may be non-identifiable without further assumptions. This is because there may be several noise distributions and functions g_i consistent with the observational distribution but result in different counterfactual distributions.

In the context of binary random variables, monotonicity is an assumption that avoids the above mentioned non-identifiability issues—it restricts the class of possible SCMs to those which all yield equivalent counterfactual distributions over a binary variable of interest [43, 19]. More specifically, a SCM $\mathcal{C}$ of a binary variable Y is monotonic with respect to a binary variable T if and only if the condition

$$P^{\mathcal{C}\,;\,\text{do}(T=t)}(Y = y) \geq P^{\mathcal{C}\,;\,\text{do}(T=t')}(Y = y)$$

implies that $P^{\mathcal{C} \mid Y=y,\, T=t'\,;\,\text{do}(T=t)}(Y = y') = 0$, where $y' \neq y$.

[2]The function $g(t)$ is often called triggering kernel.

[3]An unmeasured confounder is an unobserved variable that is a direct cause of two (observed) variables X_i and X_j.

Algorithm 1 It samples a counterfactual sequence of accepted events given a sequence of accepted and rejected events provided by Lewis' thinning algorithm

Input: $\lambda_m(t)$, $\lambda_{m'}(t)$, $\mathcal{H}_m$, $\mathcal{H}_{\max}$, $\lambda_{\max}$.
Initialize: $\mathcal{H}_{m'} = \emptyset$.

function ACC($\lambda_m(t), \lambda_{m'}(t), \mathcal{H}_m, \mathcal{H}_{\max}, \lambda_{\max}$)
  $\mathcal{H}_{m'} \leftarrow \emptyset$
  for $t_i \in \mathcal{H}_{\max}$ **do**
    $x_i \leftarrow \mathbf{1}[t_i \in \mathcal{H}_m]$
    $x'_i \sim P^{\mathcal{C} \mid X_i = x_i, \Lambda_i = \lambda_m(t_i)\,;\, \text{do}(\Lambda_i = \lambda_{m'}(t_i))}(X)$
    if $x'_i = 1$ **then**
      $\mathcal{H}_{m'} \leftarrow \mathcal{H}_{m'} \cup \{t_i\}$
    end if
  end for
  return $\mathcal{H}_{m'}$
end function

3 A Causal Model of Thinning

In this section, we first revisit Lewis' thinning algorithm [18], one of the most popular techniques to simulate event data from inhomogeneous Poisson processes. Then, we augment this classical algorithm using a particular class of SCMs satisfying the monotonicity assumption, the Gumbel-Max SCMs [19]. Finally, we demonstrate that the resulting algorithm can be used to answer counterfactual questions about a set of previously simulated events.

Let $\mathcal{M}$ be a set of inhomogeneous Poisson processes of interest and, for each $m \in \mathcal{M}$, assume its corresponding intensity function $\lambda_m(t) \leq \lambda_{\max}$ for all $t \in \mathbb{R}^+$. To sample a sequence of events $\mathcal{H}_m$ from any process $m \in \mathcal{M}$, Lewis' thinning algorithm first samples a sequence of potential events $\mathcal{H}_{\max}$ from a homogenous Poisson process with intensity $\lambda_{\max}$. Then, for each event $t_i \in \mathcal{H}_{\max}$, it additionally samples a Bernoulli random variable X_i with parameter $p = p(\lambda_m(t_i)) = \lambda_m(t_i)/\lambda_{\max}$. Finally, it accepts all the events t_i such that $X_i = 1$. *i.e.*, $\mathcal{H}_m = \{t_i \in \mathcal{H}_{\max} \mid X_i = 1\}$. Here, note that the specific choice of $\lambda_{\max}$ does not affect the distribution of accepted events as long as $\lambda_{\max} \geq \lambda_m(t)$ for all $m \in \mathcal{M}$ and $t \in \mathbb{R}^+$.[4] Algorithm 4 in Appendix C.1 summarizes the overall procedure, where the parameter p is often called the thinning probability.

Given a process of interest with intensity $\lambda_m(t)$, Lewis' thinning algorithm is helpful to make predictions about future events. For example, it can be used to compute Monte Carlo estimates of the average number of events $\mathbb{E}[N(t)]$ at a time t in the future. However, it is not sufficient to make counterfactual predictions, *e.g.*, given the sequences of events $\mathcal{H}_{\max}$ and $\mathcal{H}_m$, we cannot know what would have happened if, at time t_i, the intensity had been $\lambda_{m'}(t_i)$, with $m \neq m'$, instead of $\lambda_m(t_i)$. To overcome this limitation, we will now augment the above thinning algorithm using a Gumbel-Max SCM. More specifically, let $\mathcal{C}$ be a SCM defined by the assignments:

$$X_i = \underset{x \in \{0,1\}}{\text{argmax}}\; g(x, \Lambda_i, \boldsymbol{U}_i) \quad \text{and} \quad \Lambda_i = \lambda(t_i), \tag{4}$$

where

$$g(x, \Lambda_i, \boldsymbol{U}_i) = \log p(X_i = x \mid \Lambda_i) + U_{i,x},$$

with $p(X_i = x \mid \Lambda_i) = x\, p(\Lambda_i) + (1-x)\,(1-p(\Lambda_i))$, $p(\Lambda_i) = \Lambda_i/\lambda_{\max}$, $U_{i,x} \sim \text{Gumbel}(0,1)$, and $t_i \sim \lambda_{max}$. Then, the thinning probabilities in Lewis' thinning algorithm are given by the following interventional distributions over $\mathcal{C}$[5]:

$$P^{\mathcal{C}\,;\,\text{do}(\Lambda_i = \lambda(t_i))}(X_i = 1) = p(\lambda(t_i)) = \frac{\lambda(t_i)}{\lambda_{\max}}. \tag{5}$$

Under this view, given a sequence of accepted events $\mathcal{H}_m$ and rejected events $\mathcal{H}_{\max} \backslash \mathcal{H}_m$ under an intensity $\lambda_m(t)$, as determined by the binary samples $\{x_i\}$, we can estimate the posterior distribution

[4] In this context, note that, rather than using an homogeneous Poisson process with intensity $\lambda_{\max}$, one could use any process with (time-varying) intensity $\lambda'(t) \geq \lambda_m(t)$ for all $m \in \mathcal{M}$ and $t \in \mathbb{R}^+$, as shown in Theorem 1 in Lewis and Shedler [18].

[5] This equality follows immediately from the Gumbel-Max *trick* [44]

$P^{\mathcal{C} \mid X_i = x_i, \Lambda_i = \lambda_m(t_i) \,;\, \text{do}(\Lambda_i = \lambda_{m'}(t_i))}(U_{i,x})$ of each Gumbel noise variable $U_{i,x}$ using an efficient procedure [45, 19]. Then, we can use these noise posterior distributions to compute an unbiased finite sample Monte-Carlo estimate of the counterfactual thinning probability, *i.e.*,

$$P^{\mathcal{C} \mid X_i = x_i, \Lambda_i = \lambda_m(t_i) \,;\, \text{do}(\Lambda_i = \lambda_{m'}(t_i))}(X_i = x) = \mathbb{E}_{\boldsymbol{U}_i \mid X_i = x_i, \Lambda_i = \lambda_m(t_i)}[\mathbf{1}[x = \operatorname*{argmax}_{x' \in \{0,1\}} g(x', \lambda_{m'}(t_i), \boldsymbol{U}_i)]],$$

where we drop $\text{do}(\cdot)$ because $\boldsymbol{U}_i$ and $\lambda_m(t_i)$ are independent in the counterfactual SCM. Importantly, the above counterfactual thinning probability allows us to make counterfactual predictions, *e.g.*, given a sequence of accepted events $\mathcal{H}_m$ and rejected events $\mathcal{H} \backslash \mathcal{H}_m$ under intensity $\lambda_m(t)$, we can use the counterfactual thinning probability to predict which events, among those in $\mathcal{H}$, would have been accepted if the intensity had been $\lambda_{m'}(t)$ instead of $\lambda_m(t)$. Algorithm 1 summarizes the algorithm. Here, it is important to note that the specific choice of $\lambda_{\max}$ does not affect the distribution of counterfactual events as long as $\lambda_{\max} \geq \lambda_m(t)$ for all $m \in \mathcal{M}$ and $t \in \mathbb{R}^+$, similarly as in standard thinning (refer to Appendix A for a proof).

Finally, it is important to note that, by using Gumbel-Max SCMs, our causal model of thinning satisfies the monotonicity assumption [43, 19], discussed in Section 2, and thus the counterfactual thinning probability does not suffer from non-identifiability issues. More formally, we have the following proposition (proven in Appendix B):

Proposition 1 *Let $\mathcal{M} = \{\lambda_m(t), \lambda_{m'}(t)\}$ and $\mathcal{C}$ be the corresponding causal model of thinning. Then, if $\lambda_m(t_i) \geq \lambda_{m'}(t_i)$, it holds that $P^{\mathcal{C} \mid X_i = 0, \Lambda_i = \lambda_m(t_i) \,;\, do(\Lambda_i = \lambda_{m'}(t_i))}(X_i = 1) = 0$. Conversely, if $\lambda_m(t_i) \leq \lambda_{m'}(t_i)$, it holds that $P^{\mathcal{C} \mid X_i = 1, \Lambda_i = \lambda_m(t_i) \,;\, do(\Lambda_i = \lambda_{m'}(t_i))}(X_i = 0) = 0$.*

The above result directly implies that, if a potential event $t_i \in \mathcal{H}_{\max}$ was rejected under $\lambda_m(t)$, *i.e.*, $t_i \notin \mathcal{H}_m$, then in a counterfactual scenario, if $\lambda_{m'}(t) \leq \lambda_m(t)$, it is *impossible* that the event is accepted under $\lambda_{m'}(t)$, *i.e.*, $t_i \in \mathcal{H}_{m'}$. Conversely, if a potential event $t_i \in \mathcal{H}_{\max}$ was accepted under $\lambda_m(t)$, *i.e.*, $t_i \in \mathcal{H}_m$, then in a counterfactual scenario, if $\lambda_{m'}(t) \geq \lambda_m(t)$, it is *impossible* that the event is rejected under $\lambda_{m'}(t)$, *i.e.*, $t_i \notin \mathcal{H}_{m'}$.

4 Sampling Counterfactual Events

In this section, we develop a sampling algorithm that, given an observed realization from a temporal point process with a given intensity function, it uses Algorithm 1 and the superposition theorem [20] to generate counterfactual realizations of the temporal point process under a given alternative intensity function. To ease the exposition, we first focus on inhomogeneous Poisson processes and then generalize our algorithm to linear Hawkes processes.

Inhomogenous Poisson processes. Assume we have observed a sequence of real events $\mathcal{H}_m$ and this sequence can be accurately characterized, observationally, using an inhomogeneous Poisson process with intensity $\lambda_m(t) = g(t)$, where $g(t) \geq 0$ is a time-varying function. In reality, this sequence of events has not been *generated* by Lewis' thinning algorithm but by a natural phenomena of interest. As a result, we cannot directly apply Algorithm 1 since it requires both a sequence of accepted and rejected events. However, we can find plausible sequences of events that Lewis' thinning algorithm would have rejected if it had accepted the observed sequence of events.

By construction, the intensity of accepted and rejected events $\lambda_{\text{accepted}}(t)$ and $\lambda_{\text{rejected}}(t)$ provided by Lewis' thinning algorithm should satisfy that $\lambda_{\max} = \lambda_{\text{accepted}}(t) + \lambda_{\text{rejected}}(t)$. Then, if $\lambda_{\text{accepted}}(t) = \lambda_m(t)$, $\lambda_{\max} \geq \max_t \lambda_m(t)$ and $\mathcal{H}_{\text{accepted}} = \mathcal{H}_m$, by the superposition theorem, we can find plausible sequences of events that Lewis' algorithm would have rejected just by sampling from the intensity $\lambda_{\text{rejected}}(t) = \lambda_{\max} - \lambda_m(t)$. Then, these generated sequences of rejected events, together with the sequence of observed events, can be fed into Algorithm 1 to sample sequences of counterfactual events given an alternative intensity $\lambda_{m'}(t)$. Algorithm 2 summarizes the algorithm, where $\textsc{Lewis}(\cdot)$ samples a sequence of events using Algorithm 4 in Appendix C.1 and $\textsc{Acc}(\cdot)$ samples a counterfactual sequence of accepted events using Algorithm 1.

Linear Hawkes processes. Assume we have observed a sequence of real events $\mathcal{H}_m$ and it can be accurately characterized, observationally, using a linear Hawkes process with an intensity $\lambda_m(t)$ parameterized by μ_m, α_m and $g_m(\cdot)$, as defined in Eq. 2. If our goal is to sample sequences of counterfactual events $\mathcal{H}_{m'}$ given an alternative Hawkes intensity $\lambda_{m'}(t)$ parameterized by $\mu_{m'}$, $\alpha_{m'}$ and $g_{m'}(\cdot)$, we cannot proceed similarly as in the case of inhomogeneous Poisson processes. This is because, to sample plausible rejected events within Algorithm 2, we need to pick a value of $\lambda_{\max}$

Algorithm 2 It samples a counterfactual sequence of events given a sequence of observed events from an inhomogeneous Poisson process.

Input: $\lambda_m(t)$, $\lambda_{m'}(t)$, $\mathcal{H}_m$, $\lambda_{\max}$, T.
Initialize: $\mathcal{H}_{m'} = \emptyset$.

function CF($\lambda_m(t), \lambda_{m'}(t), \mathcal{H}_m, \lambda_{\max}, T$)
  $\mathcal{H}_{\max}, _ \leftarrow$ LEWIS($\lambda_{\max} - \lambda_m(t), \lambda_{\max}, T$)
  $\mathcal{H}_{\max} \leftarrow \mathcal{H}_{\max} \cup \mathcal{H}_m$
  $\mathcal{H}_{m'} \leftarrow$ ACC($\lambda_m(t), \lambda_{m'}(t), \mathcal{H}_m, \mathcal{H}_{\max}, \lambda_{\max}$)
  return $\mathcal{H}_{m'}$
end function

Algorithm 3 It samples a counterfactual sequence of events given a sequence of observed events from a Hawkes process.

Input: μ_m, α_m, $g_m(t)$, $\mu_{m'}$, $\alpha_{m'}$, $g_{m'}(t)$, $\mathcal{H}_m$, $\lambda_{\max}$, T.
Initialize: $\mathcal{H}_{m'} = \emptyset$.

$\{\mathcal{H}_{m,j}\} \leftarrow$ ASSIGN($\mathcal{H}_m, \lambda_m(t)$)
$\mathcal{H}_{m',0} \leftarrow$ CF($\gamma_{m,0}(t), \gamma_{m',0}(t), \mathcal{H}_{m,0}, \lambda_{\max}, T$)
$\mathcal{H}_{m'} \leftarrow \mathcal{H}_{m'} \cup \mathcal{H}_{m',0}$

for $t_j \in \mathcal{H}_m$ **do**
  if $t_j \in \mathcal{H}_{m'}$ **then**
    $\mathcal{H}_{m',j} \leftarrow$ CF($\gamma_{m,j}(t),\ \gamma_{m',j}(t),\ \mathcal{H}_{m,j},\ \lambda_{\max},\ T$)
    $\mathcal{H}_{m'} \leftarrow \mathcal{H}_{m'} \cup \mathcal{H}_{m',j}$
  end if
end for

$\mathcal{H} \leftarrow \mathcal{H}_{m'} \backslash \mathcal{H}_m$
while $|\mathcal{H}| > 0$ **do**
  $t_k \leftarrow \min_{t \in \mathcal{H}} t$
  $\mathcal{H}_{m,k}, _ \leftarrow$ LEWIS($\gamma_{m',k}(t), \lambda_{\max}, T$)
  $\mathcal{H} \leftarrow \mathcal{H}_{m,k} \cup \mathcal{H} \backslash \{t_k\}$
  $\mathcal{H}_{m'} \leftarrow \mathcal{H}_{m'} \cup \mathcal{H}_{m,k}$
end while
return $\mathcal{H}_{m'}$

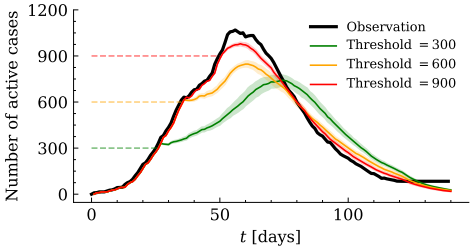
that upper bounds both $\lambda_m(t)$ and $\lambda_{m'}(t)$, otherwise, Algorithm 1 might break because the thinning probabilities p used by the causal model of thinning could be greater than 1. Unfortunately, since $\lambda_{m'}(t)$ depends on the counterfactual history $\mathcal{H}_{m'}$ we aim to sample, we cannot know its maximum value at the time we sample the rejected events. However, we can overcome this challenge by resorting to the branching process interpretation of Hawkes processes [22].

More specifically, we can view any linear Hawkes process as a superposition of several temporal point processes, *i.e.*, a process with constant intensity $\gamma_0(t) = \mu$ and, for each event $t_i \in \mathcal{H}$, a process with intensity $\gamma_i(t) = \alpha g(t - t_i)$. Under this view, we can naturally derive the following thinning algorithm to sample from linear Hawkes processes [46]. First, we sample a sequence of events t_i from the process with intensity $\gamma_0(t)$. Then, for each sampled event t_i, we create a process with intensity $\gamma_i(t)$ and sample a sequence of events t_j independently from each of them using Algorithm 4 in Appendix C.1. Further, for each of these sampled events t_j, we again create another set of processes with intensity $\gamma_j(t)$ and sample from them independently, and continue recursively. By the superposition theorem, it readily follows that the overall sequence of sampled events is a valid realization of the original Hawkes process. Algorithm 5 in Appendix C.2 summarizes the algorithm.

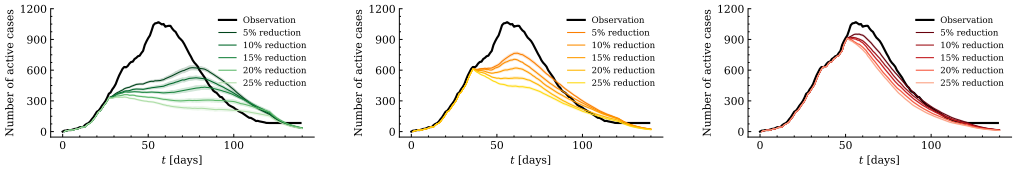
Importantly, all the intensities $\gamma_j(t)$ we sample from are bounded by $\max\{\mu, \max_t \alpha g(t)\}$. As a result, given two intensities of interest $\lambda_m(t)$ and $\lambda_{m'}(t)$, we can sample sequences of counterfactual events by running Algorithm 2 independently for each of the above processes with

$$\lambda_{\max} \geq \max\{\mu_m, \mu_{m'}, \max_t \alpha_m g_m(t), \max_t \alpha_{m'} g_{m'}(t)\}.$$

However, to do so, we also need to assign each observed event $t_i \in \mathcal{H}_m$ to one of the above processes with probability $\gamma_{m,j}(t_i) / \sum_{k<i} \gamma_{m,k}(t_i)$, which is the probability that the process has *caused* the event [47]. Algorithm 3 summarizes the resulting algorithm, where ASSIGN($\cdot$) returns the



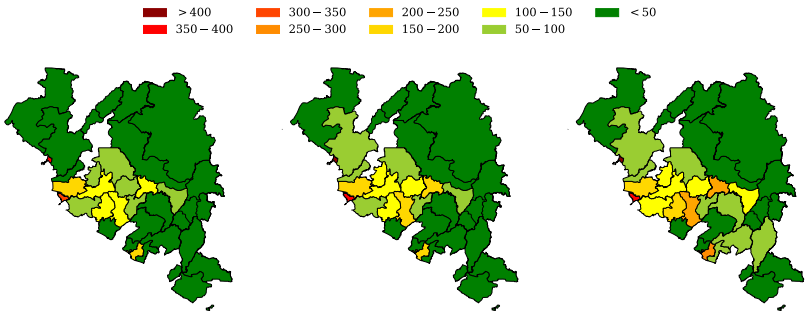
(a) Contacts reduction in highest incidence district



Threshold = 300 Threshold = 600 Threshold = 900

(b) Contacts reduction in all districts

Figure 1: Effect of interventions where individuals reduce their contacts after the overall number of active infections reaches a varying threshold. In all figures, the black line ("Observation") corresponds to an outbreak sampled from the fitted SIR model, the remaining lines correspond to counterfactual realizations for this outbreak under different interventions, and the shaded regions correspond to 95% confidence intervals. For each threshold and reduction level, we repeat the experiment five times.



Threshold = 300 Threshold = 600 Threshold = 900

Figure 2: Geographical distribution of the overall cumulative number of infections per district the counterfactual outbreaks where all individuals reduce their individual contacts by 25% after the overall number of active infections reaches a certain threshold. Darker red (green) corresponds to high (low) number of cumulative infections.

observed events $\mathcal{H}_{m,j}$ assigned to each process j, $\textsc{Cf}(\cdot)$ samples a counterfactual sequence of events using Algorithm 2, and $\textsc{Lewis}(\cdot)$ samples a sequence of events using Algorithm 4 in Appendix C.1. Within the algorithm, it is also worth noting that lines 6-11 need to go through the observed events t_j in chronological order and, for each t_j, the algorithm only accepts counterfactual events from the corresponding process with intensity $\gamma_{m',j}(t)$ if the event t_j has been previously accepted in the counterfactual realization. Moreover, lines 12-18 sample a sequence of events for each of the processes triggered by the counterfactual events that did not exist in the observed sequence of events.

5 Experiments on Epidemiological Data

In this section, we run a (simple) variation of Algorithm 3 (refer to Appendix C.3) to quantify the effect of interventions on a networked Susceptible-Infectious-Recovered (SIR) epidemiological model [48] fitted using real event data from an Ebola outbreak in West Africa in 2013-2016 [49]. Appendix D contains additional experiments on synthetic data sampled from inhomogeneous Poisson processes and linear Hawkes processes[6].

[6] All experiments were performed on a machine with 48 Intel(R) Xeon(R) 3.00GHz CPU cores and 1.5TB.

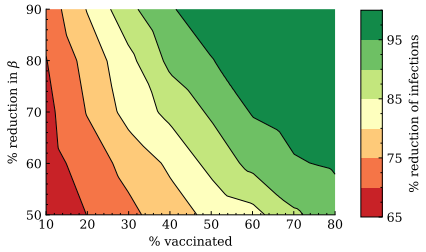


Figure 3: Effect of interventions where a percentage of the overall population receives a vaccine with a certain level of efficacy. The figure shows the average reduction in the cumulative number of infections under each intervention with respect to an outbreak sampled from the fitted SIR model.

Experimental setup. We build upon the networked Susceptible-Infectious-Recovered (SIR) epidemiological model introduced by Lorch et al. [48], which is based on temporal point processes. Given a contact network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we represent the times when each node gets infected and recovered using a collection of binary counting processes $\boldsymbol{Y}(t)$ and $\boldsymbol{W}(t)$ and we track the current state of each node using a collection of state variables $\boldsymbol{X}(t) = \boldsymbol{Y}(t) - \boldsymbol{W}(t)$, where $X_i(t) = 1$ indicates node $i \in \mathcal{V}$ is infected at time t and $X_i(t) = 0$ indicates it is susceptible or recovered. For each node $i \in \mathcal{V}$, we characterize the above counting processes using the following intensities

$$\mathbb{E}[dY_i(t) \,|\, \mathcal{H}(t)] = (1 - X_i(t)) \sum_{j \,|\, (i,j) \in \mathcal{E}} \beta \, X_j(t) dt \quad \text{and} \quad \mathbb{E}[dW_i(t) \,|\, \mathcal{H}(t)] = \delta \, X_i(t) dt, \tag{6}$$

where note that the counting process $\boldsymbol{Y}(t)$ can be viewed as a (networked) multidimensional Hawkes process with stochastic triggering kernels defined by step functions.

To set the β and δ, we resort to well-known epidemiological quantities whose values have been estimated by the World Health Organization (WHO) for the Ebola outbreak we study [50]. More specifically, the average generation time, *i.e.*, the time between the infection of a primary case and one of its secondary cases [51], is $1/\beta \approx 15.3$ days and the mean time from the onset of symptoms to death or discharge from the hospital is $1/\delta \approx 11.4$ days. To generate the contact network $\mathcal{G}$ which underpins our model, we use a stochastic block model network whose parameters are informed by estimates of the reproduction numbers for each of the countries (Guinea, Liberia, and Sierra Leone) affected by the Ebola outbreak. For computational reasons, for each of the 55 districts where infection cases were identified, we create a set of nodes $\mathcal{V}_i$ proportional to the population in the district, as estimated by the WorldPop project [52], and assume $\cup_i \mathcal{V}_i = \mathcal{V}$ with $|\mathcal{V}| = \sum_i |\mathcal{V}_i| = 8{,}000$ individuals. Moreover, we add an edge between each pair of nodes (u, v) nodes independently at random with probability $p(u, v) = 10^{-2}$ if (u, v) are in the same district, $p(u, v) = 2.15 \cdot 10^{-3}$ if they are in two contiguous districts in Guinea, $p(u, v) = 3 \cdot 10^{-3}$ if they are in two continuous districts in Liberia, $p(u, v) = 3.15 \cdot 10^{-3}$ if they are in two continuous districts in Sierra Leona, and $p(u, v) = 1.9 \cdot 10^{-3}$ if they are in two continuous districts in different countries. Here, we found contiguous districts using publicly available district-level shapefiles[7] and set the values of the between-district probabilities using grid-search so that, at a country level, the basic reproduction number of the simulated outbreaks matches that estimated from real data (refer to Table 1 in Appendix E).

In our experiments, we simulate a *realistic* outbreak by sampling a realization from the above fitted model. To this end, for each real recorded case up to January 1, 2014 [49], we sample a seed node at random from the same district as the observed case[8]. Figure 8 in Appendix E shows the geographical distribution of seed infections and overall cumulative number of infections in the outbreak. Then, given this sampled realization, we quantify the effect of two types of interventions by sampling counterfactual realizations using a variation of Algorithm 3 (refer to Appendix C.3):

— *Reduction of number of contacts*: individuals reduce their individual contacts after the overall number of active infections reaches a certain threshold. In one scenario, only individuals from the district with the highest incidence reduce their contacts within the district and get isolated from all other districts and, in another scenario, everyone reduces their individual contacts.

[7] https://data.humdata.org/

[8] The first real case was recorded on Dec 26, 2013. By January 1, 2014, there were six recorded cases in four different districts.

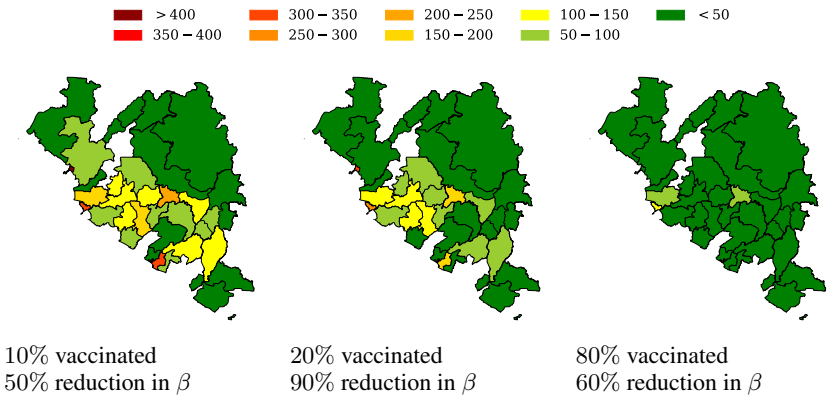


Figure 4: Geographical distribution of the overall cumulative number of infections per district the counterfactual outbreaks where percentage of the population receives a vaccine with a certain level of efficacy. Darker red (green) corresponds to high (low) number of cumulative infections.

— *Vaccination*: a percentage of the overall population receives a vaccine with a certain level of efficacy. Within our model, we measure vaccine efficacy in terms of reduction of the value of the parameter β, which controls the infection rate between individuals.

In each experiment, to estimate the average and confidence intervals of the outcome of interest (*e.g.*, number of cases), we sample 20 counterfactual realizations.

Results. Figures 1–2 summarize the results for the interventions where individuals reduce their individual contacts after the overall number of active infections reaches a certain threshold. Our results suggest that, at all threshold levels, reducing the individual contacts across all districts, even by just 5%, would have been more effective than isolating and reducing the contacts by 50% in the district with higher incidence. Moreover, we also find that, for lower threshold values, the counterfactual outbreaks would have spread to fewer districts and the overall number of infections would have been significantly lower. Figures 3–4 summarize the results for the interventions where a percentage of the overall population receives a vaccine with a certain level of efficacy. Our results suggest that a high level of vaccine effectivity would have not been sufficient to reduce the number of infections if the percentage of the population who had received the vaccine was *low*. For example, if less than 20% of the population had received the vaccine, even a vaccine with a 90% effectivity would have been unable to reduce the infections by >70%. If >80% of the population had received the vaccine, a vaccine with just a 60% effectivity would have reduced the infections by >95%. Finally, we find that, similarly as in the scenario where individuals reduce their individual contacts, the reductions in the overall number of infections would have also led to lower geographical dispersion.

6 Conclusions, Limitations and Future Work

Since counterfactual reasoning lies within level three in the "ladder of causation" [42], we have been unable to validate our counterfactual predictions using observational nor interventional experiments. That being said, we have made an intuitive assumption—monotonicity—about the causal mechanism of the world—our causal model of thinning—which specifies how changes on the intensity function of a temporal point process may have lead to particular outcomes while holding "every-thing else" fixed, similarly as previous work [19, 38]. In this context, it would be worth to understand the sensitivity of counterfactual realizations to the specific choice of SCM.

In our work, we have focused on temporal point processes, however, some of our ideas can be readily extended to spatial point processes. Moreover, it would be interesting to extend our methodology to support nonlinear Hawkes processes and neural Hawkes processes [53, 54]. To this end, the main challenge is that there is no branching process interpretation for nonlinear or neural Hawkes and thus we would have to find alternative ways to bound the value of the relevant intensity functions.

Finally, it would be important to carry out a user study in which the counterfactual realizations provided by our algorithm are shared with domain experts (*e.g.*, epidemiologists) and evaluate our sampling algorithm using other real datasets from other applications such as recommendation systems and voting data.

References

[1] M Gomez Rodriguez and Isabel Valera. Learning with temporal point processes. *Tutorial at ICML*, 2018.

[2] Junchi Yan, Hongteng Xu, and Liangda Li. Modeling and applications for temporal point processes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3227–3228, 2019.

[3] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez-Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. Coevolve: A joint point process model for information diffusion and network evolution. *The Journal of Machine Learning Research*, 18(1):1305–1353, 2017.

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

[5] Minkyoung Kim, Dean Paini, and Raja Jurdak. Modeling stochastic processes in disease spread across a heterogeneous social system. *Proceedings of the National Academy of Sciences*, 116(2):401–406, 2019.

[6] Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*, 2021.

[7] Ali Zarezade, Abir De, Utkarsh Upadhyay, Hamid R Rabiee, and Manuel Gomez-Rodriguez. Steering social activity: A stochastic optimal control point of view. *J. Mach. Learn. Res.*, 18:205–1, 2017.

[8] Utkarsh Upadhyay, Abir De, and Manuel Gomez-Rodriguez. Deep reinforcement learning of marked temporal point processes. In *Advances in Neural Information Processing Systems*, 2018.

[9] Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993, 2019.

[10] Ruth MJ Byrne. Mental models and counterfactual thoughts about what might have been. *Trends in cognitive sciences*, 6(10):426–431, 2002.

[11] Kai Epstude and Neal J Roese. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192, 2008.

[12] Herbert W Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4):599–653, 2000.

[13] Andrea L Bertozzi, Elisa Franco, George Mohler, Martin B Short, and Daniel Sledge. The challenges of modeling and forecasting the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(29):16732–16738, 2020.

[14] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[15] Adam J Kucharski, Petra Klepac, Andrew JK Conlan, Stephen M Kissler, Maria L Tang, Hannah Fry, Julia R Gog, W John Edmunds, Jon C Emery, Graham Medley, et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of sars-cov-2 in different settings: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(10):1151–1160, 2020.

[16] Lars Lorch, Heiner Kremer, William Trouleau, Stratis Tsirtsis, Aron Szanto, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Quantifying the effects of contact tracing, testing, and containment measures in the presence of infection hotspots. *arXiv preprint arXiv:2004.07641*, 2020.

[17] Chad R Wells, Pratha Sah, Seyed M Moghadas, Abhishek Pandey, Affan Shoukat, Yaning Wang, Zheng Wang, Lauren A Meyers, Burton H Singer, and Alison P Galvani. Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. *Proceedings of the National Academy of Sciences*, 117(13):7504–7509, 2020.

[18] PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

[19] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

[20] John Frank Charles Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

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

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

[23] Vanessa Didelez. Causal reasoning for events in continuous time: A decision-theoretic approach. In *ACI at UAI*, pages 40–45, 2015.

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

[25] 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.

[26] Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. Thp: Topological hawkes processes for learning granger causality on event sequences. *arXiv preprint arXiv:2105.10884*, 2021.

[27] Massil Achab, Emmanuel Bacry, Stéphane Gaıffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. In *International Conference on Machine Learning*, pages 1–10. PMLR, 2017.

[28] William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Cumulants of hawkes processes are robust to observation noise. In *International Conference on Machine Learning*, pages 10444–10454. PMLR, 2021.

[29] Flavio Figueiredo, Guilherme Borges, Pedro OS de Melo, and Renato M Assunção. Fast estimation of causal interactions using wold processes. *arXiv preprint arXiv:1807.04595*, 2018.

[30] Jalal Etesami, William Trouleau, Negar Kiyavash, Matthias Grossglauser, and Patrick Thiran. A variational inference approach to learning multivariate wold processes. In *International Conference on Artificial Intelligence and Statistics*, pages 2044–2052. PMLR, 2021.

[31] Tian Gao, Dharmashankar Subramanian, Debarun Bhattacharjya, Xiao Shou, Nicholas Mattei, and Kristin P Bennett. Causal inference for event pairs in multivariate point processes. *Advances in Neural Information Processing Systems*, 34:17311–17324, 2021.

[32] Odd O Aalen, Mats J Stensrud, Vanessa Didelez, Rhian Daniel, Kjetil Røysland, and Susanne Strohmaier. Time-dependent mediators in survival analysis: Modeling direct and indirect effects with the additive hazards model. *Biometrical Journal*, 62(3):532–549, 2020.

[33] Judith J Lok. Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, 36(3):1464–1507, 2008.

[34] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems*, 30:1697–1708, 2017.

[35] Pål Christie Ryalen, Mats Julius Stensrud, Sophie Fosså, and Kjetil Røysland. Causal inference in continuous time: an example on prostate cancer therapy. *Biostatistics*, 21(1):172–185, 2020.

[36] Kjetil Røysland. Counterfactual analyses with graphical models based on local independence. *The Annals of Statistics*, 40(4):2162–2194, 2012.

[37] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[38] Stratis Tsirtsis, Abir De, and Manuel Gomez-Rodriguez. Counterfactual explanations in sequential decision making under uncertainty. 2021.

[39] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019.

[40] Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

[41] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

[42] Judea Pearl. *Causality*. Cambridge university press, 2009.

[43] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

[44] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Wiley, New York, 1959.

[45] Chris J. Maddison and Danny Tarlow. Gumbel machinery, Jan 2017.

[46] Michael George Moore. *Maximum likelihood estimation of Poisson and Hawkes processes and extensions to Hawkes process analysis*. PhD thesis, Georgia Institute of Technology, 2018.

[47] Michael G Moore and Mark A Davenport. A hawkes' eye view of network information flow. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, 2016.

[48] Lars Lorch, Abir De, Samir Bhatt, William Trouleau, Utkarsh Upadhyay, and Manuel Gomez-Rodriguez. Stochastic optimal control of epidemic processes in networks. In *Machine Learning for Health (ML4H) NeurIPS Workshop*, 2018.

[49] Tini Garske, Anne Cori, Archchun Ariyarajah, Isobel M Blake, Ilaria Dorigatti, Tim Eckmanns, Christophe Fraser, Wes Hinsley, Thibaut Jombart, Harriet L Mills, et al. Heterogeneities in the case fatality ratio in the west african ebola outbreak 2013–2016. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1721):20160308, 2017.

[50] WHO Ebola Response Team. Ebola virus disease in west africa—the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, 371(16):1481–1495, 2014.

[51] Åke Svensson. A note on generation times in epidemic models. *Mathematical biosciences*, 208(1):300–311, 2007.

[52] Christopher T Lloyd, Alessandro Sorichetta, and Andrew J Tatem. High resolution global gridded data for use in population studies. *Scientific data*, 4(1):1–17, 2017.

[53] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1555–1564, 2016.

[54] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

[55] Iris AM Huijben, Wouter Kool, Max Benedikt Paulus, and Ruud JG Van Sloun. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Refer to "Remarks" in Section 5 and Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We cannot think of any potential negative societal impact for our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix B
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please, refer to the README file in the Supplementary Material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Refer to Section 5 and Appendix D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Refer to footnote 6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] The dataset we used was released by the World Health Organization (WHO) for general distribution. It can be reviewed, reproduced or translated for research or private study. The WHO does not explicitly mention any license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proof of invariance with respect to $\lambda_{\max}$

To prove that Algorithm 1 is invariant with respect to the choice of $\lambda_{\max}$ as long as $\lambda_{\max} \geq \lambda_m(t)$ for all $m \in \mathcal{M}$ and $t \in \mathbb{R}^+$, we only need to prove that

$$\text{(i)} \quad \lambda_m(t) P^{\mathcal{C} \mid X=1, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}(X) \quad \text{whenever} \quad \lambda_m(t) < \lambda_{m'}(t)$$

and

$$\text{(ii)} \quad \lambda_m(t) + (\lambda_{max} - \lambda_m(t))\, P^{\mathcal{C} \mid X=0, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}(X) \quad \text{whenever} \quad \lambda_m(t) \geq \lambda_{m'}(t)$$

are invariant to the specific choice of λ_{max}.

To prove that (i) is invariant, we first rewrite the counterfactual probability $P^{\mathcal{C} \mid X=1, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}$ in terms of uniform random variables, following Section 4.2 in Huijben et al. [55]:

$$\begin{aligned}
&P^{\mathcal{C} \mid X=1, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}(X=1) = \\
&\qquad \mathbb{P}_{U_0,U_1}\Bigg[-\log\left(-\log(U_1)\right) + \log\left(\frac{\lambda_{m'}(t)}{\lambda_{\max}}\right) - \log\left(\frac{\lambda_m(t)}{\lambda_{\max}}\right) > \\
&\qquad -\log\left(-\log(U_1) - \frac{\log(U_0)}{1-\frac{\lambda_m(t)}{\lambda_{\max}}}\right) + \log\left(1-\frac{\lambda_{m'}(t)}{\lambda_{\max}}\right) - \log\left(1-\frac{\lambda_m(t)}{\lambda_{\max}}\right)\Bigg]
\end{aligned}$$

where $U_0, U_1 \sim U[0,1]$. Now, for a fixed U_0 and U_1, the above inequality can be rewritten as follows:

$$\begin{aligned}
-\log\left[-\log(U_1)\frac{\lambda_m(t)}{\lambda_{\max}}\frac{\lambda_{\max}}{\lambda_{m'}(t)}\right] &> -\log\left[\left(-\log(U_1) - \frac{\log(U_0)}{1-\frac{\lambda_m(t)}{\lambda_{\max}}}\right)\frac{1-\frac{\lambda_m(t)}{\lambda_{\max}}}{1-\frac{\lambda_{m'}(t)}{\lambda_{\max}}}\right] \\
-\log(U_1)\frac{\lambda_m(t)}{\lambda_{m'}(t)}\left(1-\frac{\lambda_{m'}(t)}{\lambda_{\max}}\right) &< -\log(U_1)\left(1-\frac{\lambda_m(t)}{\lambda_{\max}}\right) - \log(U_0) \\
-\log(U_1)\left(\frac{\lambda_m(t)}{\lambda_{m'}(t)} - \frac{\lambda_m(t)}{\lambda_{\max}}\right) &< -\log(U_1)\left(1-\frac{\lambda_m(t)}{\lambda_{\max}}\right) - \log(U_0)
\end{aligned}$$

In the last inequality, the terms containing $\lambda_{\max}$ on the left and right hand side are identical and can be canceled. This proves that (i) is invariant to the specific choice of λ_{max}.

To prove that (ii) is also invariant, we proceed similarly as in (i) and first rewrite the counterfactual probability $P^{\mathcal{C} \mid X=0, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}$ in terms of uniform random variables:

$$\begin{aligned}
&P^{\mathcal{C} \mid X=0, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}(X=1) = \\
&\qquad \mathbb{P}_{U_0,U_1}\Bigg[-\log\left(-\log(U_0) - \frac{\log(U_1)}{\frac{\lambda_m(t)}{\lambda_{\max}}}\right) + \log\left(\frac{\lambda_{m'}(t)}{\lambda_{\max}}\right) - \log\left(\frac{\lambda_m(t)}{\lambda_{\max}}\right) > \\
&\qquad\qquad -\log\left(-\log(U_0)\right) + \log\left(1-\frac{\lambda_{m'}(t)}{\lambda_{\max}}\right) - \log\left(1-\frac{\lambda_m(t)}{\lambda_{\max}}\right)\Bigg]
\end{aligned}$$

where $U_0, U_1 \sim U[0,1]$. Now, for a fixed U_0 and U_1, the above inequality can be rewritten as follows:

$$\log(U_0)\left(1-\frac{\lambda_m(t)}{\lambda_{m'}(t)}\right) \leq \log(U_1)\left(\frac{\lambda_{\max}}{\lambda_{m'}(t)} - 1\right)$$

Now, using the fact that $-\log(U_0)$ and $-\log(U_1)$ are distributed as exponential random variables with rate $\lambda = 1$ and the CDF of the ratio $X = \log(U_0)/\log(U_1)$ of two exponential random variables with rate λ_1 is given by $\mathbb{P}[X \leq x] = 1/(1/x+1)$, we have that:

$$P^{\mathcal{C} \mid X=0, \Lambda=\lambda_m(t)\,;\,\text{do}(\Lambda=\lambda_{m'}(t))}(X=1) = \mathbb{P}_{U_0,U_1}\left[\frac{\log(U_0)}{\log(U_1)} > \frac{\frac{\lambda_{\max}}{\lambda_{m'}(t)} - 1}{1-\frac{\lambda_m(t)}{\lambda_{m'}(t)}}\right] = \frac{\lambda_{m'}(t) - \lambda_m(t)}{\lambda_{\max} - \lambda_m(t)}.$$

Then, it readily follows that

$$(\lambda_{\max} - \lambda_m(t))\,\frac{\lambda_{m'}(t) - \lambda_m(t)}{\lambda_{\max} - \lambda_m(t)} = \lambda_{m'}(t) - \lambda_m(t).$$

This proves that (ii) is invariant to the specific choice of λ_{max}.

B Proof of Proposition 1

If $\lambda_m(t_i) \geq \lambda_{m'}(t_i)$, then we have:

$$\lambda_m(t_i) \geq \lambda_{m'}(t_i) \implies \lambda_m(t_i)\left(1 - \frac{\lambda_{m'}(t_i)}{\lambda_{\text{max}}}\right) \geq \lambda_{m'}(t_i)\left(1 - \frac{\lambda_m(t_i)}{\lambda_{\text{max}}}\right) \implies \frac{1 - \frac{\lambda_{m'}(t_i)}{\lambda_{\text{max}}}}{1 - \frac{\lambda_m(t_i)}{\lambda_{\text{max}}}} \geq \frac{\frac{\lambda_{m'}(t_i)}{\lambda_{\text{max}}}}{\frac{\lambda_m(t_i)}{\lambda_{\text{max}}}},$$

Now, by Eq. 5, the last inequality is equivalent to:

$$\frac{P^{\mathcal{C}\,;\,\text{do}(\Lambda=\lambda_{m'}(t_i))}(X=0)}{P^{\mathcal{C}\,;\,\text{do}(\Lambda=\lambda_m(t_i))}(X=0)} \geq \frac{P^{\mathcal{C}\,;\,\text{do}(\Lambda=\lambda_{m'}(t_i))}(X=1)}{P^{\mathcal{C}\,;\,\text{do}(\Lambda=\lambda_m(t_i))}(X=1)}, \tag{7}$$

which is exactly the counterfactual stability property. Finally, by Theorem 2 in [19], we know that the Gumbel-Max SCM satisfies the counterfactual stability property. As a result, Eq. 7 implies that

$$P^{\mathcal{C}\,|\,X=0,\Lambda=\lambda_m(t_i)\,;\,\text{do}(\Lambda=\lambda_{m'}(t_i))}(X=1) = 0,$$

as desired. The proof for the second case is exactly the same.

C Algorithms

C.1 Lewis' thinning algorithm

Within Algorithm 4, lines 4–6 sample an event from a Poisson process with constant intensity λ_{max} using inversion sampling and lines 11–14 accept/reject the event according to the ratio between the intensity of interest and λ_{max} at the time of the event.

Algorithm 4 Lewis' thinning algorithm

1: **Input**: $\lambda(t)$, λ_{max}, T.
2: **Initialize**: $s = 0$, $\mathcal{H} = \emptyset$.

3: **function** LEWIS($\lambda(t), \lambda_{\text{max}}, T$)
4: **while** true **do**
5: $u_1 \sim \text{Uniform}(0, 1)$
6: $w \leftarrow -\ln u / \lambda_{\text{max}}$
7: $s \leftarrow s + w$
8: **if** $s > T$ **then**
9: break
10: **end if**
11: $\mathcal{H}_{\max} \leftarrow \mathcal{H}_{\max} \cup \{s\}$
12: $u_2 \sim \text{Uniform}(0, 1)$
13: **if** $u_2 \leq \lambda(s)/\lambda_{\text{max}}$ **then**
14: $\mathcal{H} \leftarrow \mathcal{H} \cup \{s\}$
15: **end if**
16: **end while**

17: **return** $\mathcal{H}, \mathcal{H}_{\max} \backslash \mathcal{H}$
18: **end function**

C.2 Thinning algorithm for Hawkes processes

Algorithm 5 samples a sequence of events from a linear Hawkes process using its branching process interpretation [46], where LEWIS$(\cdot)$ samples a sequence of events using Algorithm 4, $\gamma_0(t) = \mu$ and $\gamma_i(t) = \alpha g(t - t_i)$.

Algorithm 5 It samples a sequence of events from a linear Hawkes process using its branching process interpretation.

Input: μ, α, $g(t)$, $\lambda_{\max}$, T.
Initialize: $\mathcal{H} = \emptyset$.

function SAMPLEHAWKES($\lambda(t), \lambda_{\max}, T$)
 $\mathcal{H} \leftarrow$ LEWIS$(\gamma_0(t), \lambda_{\max}, T)$

 $\mathcal{H}' \leftarrow \mathcal{H}$
 while $|\mathcal{H}'| > 0$ **do**
 $t_i \leftarrow \min_{t \in \mathcal{H}'} t$
 $\mathcal{H}_i, _ \leftarrow$ LEWIS$(\gamma_i(t), \lambda_{\max}, T)$
 $\mathcal{H}' \leftarrow \mathcal{H}_i \cup \mathcal{H}' \backslash \{t_i\}$
 $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_i$
 end while

 return $\mathcal{H}$
end function

C.3 Sampling Counterfactual Events in a SIR Epidemic Model

As discussed in Section 5, the SIR model defined by Eq. 6 can be viewed as a (networked) multidimensional Hawkes process with stochastic triggering kernels defined by step functions. More specifically, let t_i and $\tau_i = t_i + \Delta_i$ be the infection and recovery times of each node $i \in \mathcal{G}$. Then, the

Algorithm 6 It samples a counterfactual sequence of infections given a sequence of observed infections from the SIR process defined by Eq. 6.

Input: $\beta_m, \delta, \beta_{m'}, \mathcal{G} = (\mathcal{V}, \mathcal{E}), \mathcal{G}' = (\mathcal{V}', \mathcal{E}'), T$.
Initialize: queue = PriorityQueue($\{i \mid i \in \mathcal{V}$ and is_seed(i)$\}$), $\beta_{max} = max(\beta_m, \beta_{m'})$, processed = $\{\}$, infector = $\{\}$.

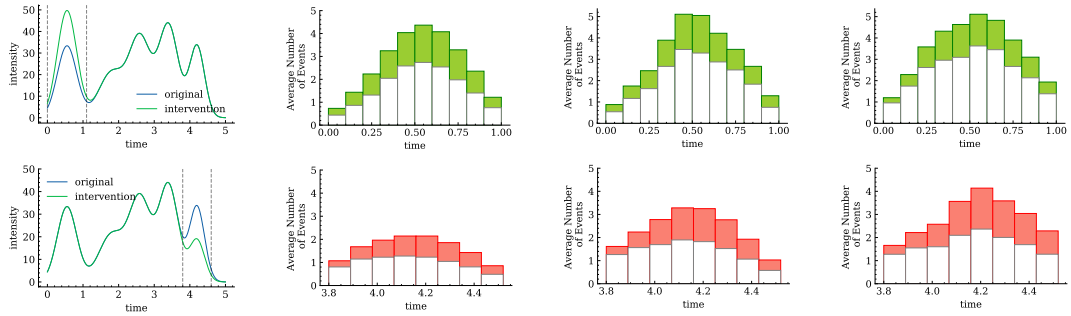
for $i \in \mathcal{G}$ **do**
  processed$[i]$ = False
  if is_seed(i) **then**
    $t_i = 0$
    $\tau_i = t_i + \exp(\delta)$
  else
    $t_i = \infty$
  end if
end for

while $\neg$ queue.isEmpty() **do**
  i = queue.*pop*()
  if $\neg$ processed$[i]$ **then**
    processed$[i]$ = True
    for $j \in \mathcal{G}'$(i) **do**
      $\gamma_{m'}(t) = \beta_{m'} u(t - t_i) - \beta_{m'} u(t - \tau_i)$
      if infector$[j]$ == i **then**
        $t = t_j$
        $\gamma_m(t) = \beta_m u(t - t_i) - \beta_m u(t - min(\tau_i, t))$
        $\mathcal{H}_{m'} = \textsc{Cf}(\gamma_m(t), \gamma_{m'}(t), \{t\}, \beta_{max}, T)$
        if $\mathcal{H}_{m'} \neq \emptyset$ **then**
          $t = \min_{t' \in \mathcal{H}_{m'}} t'$
        else
          $t = \infty$
        end if
      else
        $\mathcal{H}, _ = \textsc{Lewis}(\gamma_{m'}(t), \beta_{max}, T)$
        $t = \min_{t' \in \mathcal{H}} t'$
      end if
      if $t < t_j$ **then**
        $t_j = t$
        $\tau_j = t_j + \exp(\delta)$
        infector$[j] = i$
        queue.*add*(j, priority $= t_j$)
      end if
    end for
  end if
end while

intensity $\mathbb{E}[dY_i(t) \mid \mathcal{H}(t)]$ can be (re-)written as:

$$\mathbb{E}[dY_i(t) \mid \mathcal{H}(t)] = \beta \sum_{j \in \mathcal{G}(i)} g_i(t - t_j)$$

where $\mathcal{G}(i)$ denotes the set of neighborhood of node i, $g_i(t) = [1 - Y_i(t)][u(t) - u(t - \Delta)]$ with $\Delta \sim \text{Exp}(\delta)$ can be viewed as a stochastic triggering kernel, and $u(\cdot)$ is the step function. As a result, we can adapt Algorithm 3, originally developed for unidimensional linear Hawkes processes, to sample counterfactual realizations of the SIR model. Algorithm 6 summarizes the resulting algorithm.



(a) Intensities (b) Low number of events (c) Medium number of events (d) High number of events

Figure 5: Effect of interventions in two inhomogeneous Poisson processes. Panel (a) shows the intensities of the original and the intervened processes (in blue and green, respectively) and a window of interest (dashed vertical lines). Panels (b-d) show the difference in average number of events over time between the counterfactual and the original realizations, where each row corresponds to one process and we group the original realizations in three quantiles according to their overall number of events in the time window of interest. Positive (negative) differences are shown in green (red). In each experiment, we sample $1{,}000$ realizations from the original process and, for each of these realizations, we sample 100 counterfactual realizations from the intervened process.

D Experiments on Synthetic Data

In this section, we feed Algorithms 2 and 3 with realizations of synthetic inhomogeneous Poisson processes and linear Hawkes processes and investigate to what extent the counterfactual realizations returned by the algorithms under alternative intensity functions differ from the original realizations fed to them.

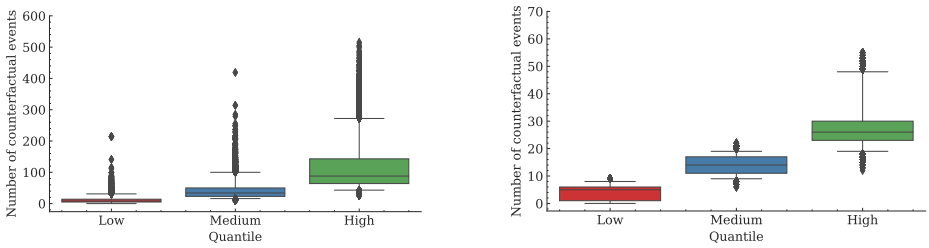
Experimental setup. We consider the family of inhomogeneous Poisson processes $\mathcal{M}(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\tau})$ parameterized by weighted combinations of RBF kernels, *i.e.*,

$$\lambda(t) = \sum_j \phi_j \exp\left(-\alpha_j (t - \tau_j)\right), \; t \geq 0, \tag{8}$$

where $\phi_j, \alpha_j, \tau_j \geq 0$, and the family of linear Hawkes processes $\mathcal{M}(\mu, \alpha, \omega)$ defined in Eq. 2, with exponential triggering kernels $g(t) = \exp(-\omega t)$. Moreover, we experiment with simple interventions under which, for inhomogeneous Poisson processes, one of the RBF kernels, picked at random, change its amplitude, *i.e.*, $\phi_{m',i} = \max(\phi_{m,i} + \epsilon, 0)$, and, for Hawkes processes, the parameter α change its value, *i.e.*, $\alpha_{m'} = \max(\alpha_m + \epsilon, 0)$, where $\epsilon \sim N(0, \sigma)$.

In each experiment, we first sample $1{,}000$ realizations from a process with one set of parameters using Algorithm 4 (or Algorithm 5). Then, we carry out the above mentioned intervention and, for each of the sampled realizations, we use Algorithm 2 (or Algorithm 3) to sample 100 counterfactual realizations under the resulting alternative set of parameters. Finally, we partition the original realizations in three quantiles according to their overall number of events in a time window of interest and, for each quantile, we look at the number of events in the corresponding counterfactual realizations in the same window of interest. In all experiments, Algorithms 1—3 use 100 samples from the posterior distribution $P^{\mathcal{C} \mid X_i = x, \Lambda_i = \lambda(t_i) \,;\, \mathrm{do}(\Lambda_i = \lambda_{m'}(t_i))}(\boldsymbol{U}_i)$ of each Gumbel noise variable $U_{i,x}$ to estimate the counterfactual thinning probabilities $P^{\mathcal{C} \mid X_i = x, \Lambda_i = \lambda(t_i) \,;\, \mathrm{do}(\Lambda_i = \lambda_{m'}(t_i))}(X_i)$.

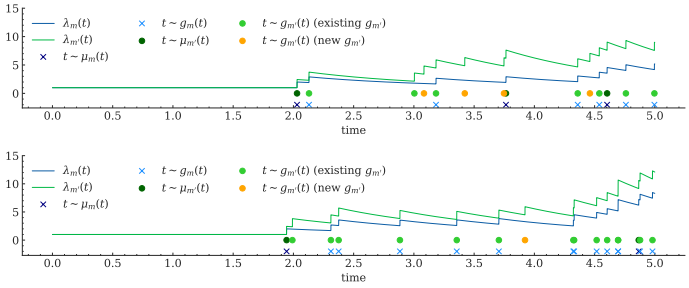
Results. Figure 5 summarizes the results for two specific inhomogeneous Poisson processes undergoing one of the above mentioned interventions—we found qualitatively similar results for other inhomogeneous Poisson processes. In the top row, the results show that, under our model, an intervention that increases the original intensity of the process, by increasing the amplitude of one of the RFB kernels by approximately an additional half does not have the same effect across realizations. In realizations with a low (high) number of events, the average number of counterfactual events in the window of interest increases up to $\sim$60% ($\sim$40%) over time with respect to the average number of events in the original realizations. In the bottom row, we find that this is also true for an intervention



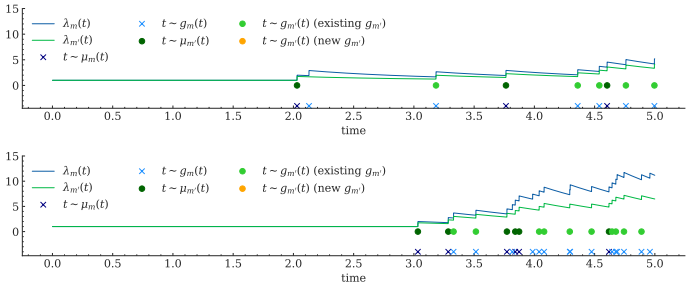
(a) Intervention increases self-excitation (b) Intervention decreases self-excitation

Figure 6: Effect of interventions in Hawkes processes (I). Panels summarize the distribution of the number of events per counterfactual realization corresponding to original realizations with a low (red; $|\mathcal{H}_m| \in [0, 9]$, $\mathbb{E}[|\mathcal{H}_m|] = 5.04$), medium (blue; $|\mathcal{H}_m| \in [10, 22]$, $\mathbb{E}[|\mathcal{H}_m|] = 17.17$) and high (green; $|\mathcal{H}_m| \in [25, 59]$, $\mathbb{E}[|\mathcal{H}_m|] = 36.17$) number of events. The horizontal lines within the boxes indicate average value, the boxes indicate 25% and 75% quantiles, the whiskers indicate 5% and 95% quantiles and the points are outliers. Here, we sample 1,000 realizations from the original process, with parameters $\mu_m = 1$, $\alpha_m = 1$, and $\omega_m = 1$, and, for each of these realizations, we sample 100 counterfactual realizations from each of the intervened process. In panel (a), the intervened process has parameter $\alpha_{m'} = 1.44$, in panel (b), it has parameter $\alpha_{m'} = 0.75$ and, in both panels, the remaining parameters $\mu_{m'} = \mu_m$ and $\alpha_{m'} = \alpha_m$. We set the time horizon $T = 5$.

that decreases the original intensity of the process, by approximately halving the amplitude of another of the RBF kernels. However, the difference is smaller in relative terms.



(a) Intervention increases self-excitation, $\alpha_{m'} > \alpha_m$



(b) Intervention decreases self-excitation, $\alpha_{m'} < \alpha_m$

Figure 7: Effect of interventions in Hawkes processes (II). Panels show the Hawkes intensities and event times corresponding to specific realizations of the original process and the counterfactual processes. Crosses (dots) denote events of the original (counterfactual) processes and the color pattern indicates which type of process generated each event (be it $\mu_{m'}$ or $g_{m'}$ due to a counterfactual event that also existed (did not exist) in the original realization). The parameters of the original and intervened process are the same as in Figure 6.

Figures 6 and 7 summarize the results for a specific Hawkes process undergoing two of the above mentioned interventions—we found qualitative similar results for other Hawkes processes. In figure 6 , we find that, similarly as in inhomogeneous Poisson processes, the interventions do not have the same effect across realizations. However, in this case, the difference among them is more stark—while the average number of counterfactual events (a) increases by 115% and (b) decreases by 11% for realizations with a low number of events, it (a) increases by 216% and (b) decreases by 21% for realizations with a high number of events. Moreover, we also find that, there is a high variability across counterfactual realizations, especially when $\alpha_{m'} > \alpha_m$. For example, while the original realizations with a low number of events never contained more than 9 events, there exists counterfactual realizations with more than 200 events. This is due to the self-exciting property of Hawkes processes by which counterfactual events may trigger the emergence of additional counterfactual events, shown as yellow dots in figure 7.

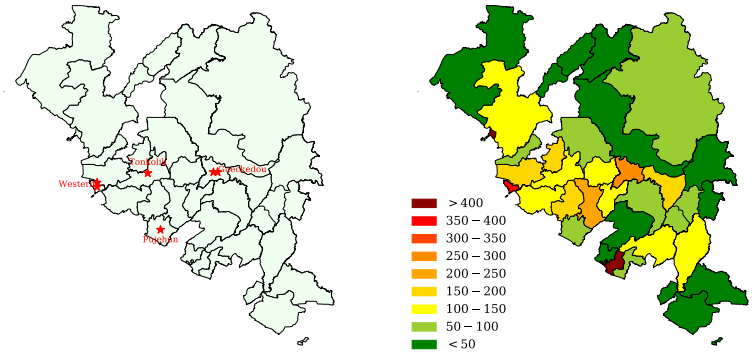
E Additional Details about the Experiments on Real Data

Table 1 shows that, under the values of the between-district probabilities estimated using grid-search, the basic reproduction numbers of the simulated outbreaks match, at a country level, those estimated from real data by WHO [50].

Country	R_0 (WHO)	R_0 (simulated)
Guinea (GN)	1.71 (1.44 – 2.01)	1.71 (1.66 – 1.76)
Liberia (LB)	1.83 (1.72 – 1.94)	1.83 (1.74 – 1.91)
Sierra Leone (SL)	2.02 (1.79 – 2.26)	2.02 (1.95 – 2.10)

Table 1: Reproduction numbers (R_0) for each of the three countries affected by the Ebola outbreak estimated from real cases by WHO [50] and estimated from simulated cases. In each cell, the first number is the average and the numbers in parentheses are the 95% confidence interval.

Figure 8 shows the geographical distribution of seed infections and overall cumulative number of infections in the outbreak.



(a) Seed infections (b) Overall cumulative infections

Figure 8: Geographical distribution of infections in a *realistic* Ebola outbreak in West Africa. Panel (a) shows the seed infections in each district, where each red star represents an infection. Panels (b) shows the overall cumulative number of infections per district. Darker red (green) corresponds to high (low) number of cumulative infections.