Inference for Mark-Censored Temporal Point Processes

Alex Boyd[1] **Yuxin Chang**[2] **Stephan Mandt**[1,2] **Padhraic Smyth**[1,2]

[1]Department of Statistics, University of California, Irvine
[2]Department of Computer Science, University of California, Irvine

Abstract

Marked temporal point processes (MTPPs) are a general class of stochastic models for modeling the evolution of events of different types ("marks") in continuous time. These models have broad applications in areas such as medical data monitoring, financial prediction, user modeling, and communication networks. Of significant practical interest in such problems is the issue of missing or censored data over time. In this paper, we focus on the specific problem of inference for a trained MTPP model when events of certain types are not observed over a period of time during prediction. We introduce the concept of mark-censored subprocesses and use this framework to develop a novel marginalization technique for inference in the presence of censored marks. The approach is model-agnostic and applicable to any MTPP model with a well-defined intensity function. We illustrate the flexibility and utility of the method in the context of both parametric and neural MTPP models, with results across a range of datasets including data from simulated Hawkes processes, self-correcting processes, and multiple real-world event datasets.

1 INTRODUCTION

Stochastic models for event data evolving in continuous time are typically referred to as temporal point processes. An important class within this general family is *marked temporal point processes* (MTPPs), where each event in time is associated with a random outcome known as a mark. In general, the mark can either be discrete or continuous; in this work we focus on discrete marks. The flexibility of MTPPs has allowed them to be applied to a broad range of applications, including medical diagnosis [Islam et al., 2017], epidemic spread models [Marmarelis et al., 2022], environmental data analysis [Brillinger, 2000], financial data prediction [Zhu et al., 2021, Shi and Cartlidge, 2022], communication network modeling [Mishra and Venkitasubramaniam, 2013], user behavior analysis [Yang et al., 2021, Hatt and Feuerriegel, 2020], misinformation spread models [Zhang et al., 2021], and activity prediction [Fortino et al., 2020].

The foundations for MTPP models have their origins in the statistical literature (e.g., Cox and Lewis [1972], Daley and Vere-Jones [2003], Andersen et al. [2012]), with subsequent development of specific classes of MTPPs such as multivariate self-exciting Hawkes processes [Hawkes, 1971] and multivariate self-correcting processes [Zheng and Vere-Jones, 1991]. More recently, there has been significant activity in the development of machine learning methods for MTPPs, with a significant emphasis on approaches that take advantage of neural representation learning, such as recurrent MTPPs [Du et al., 2016], neural Hawkes processes [Mei and Eisner, 2017], stochastic variants of deep MTPPs [Hong and Shelton, 2022], scalable deep MTPPs Türkmen et al. [2020], as well as general approaches to forecasting with deep MTPP models [Deshpande et al., 2021].

An important practical aspect of working with real-world event data is that censoring of observations can occur in a number of different ways. For example, a common example of right-censoring often occurs in survival analysis (a subfield of temporal point processes) in which a patient's event of interest is unobserved due to the end of a data collection period. This particular type of censoring is well-studied and there are well-known methods for accommodating this during training and inference. More recently, there has been work on handling broader categories of censoring for neural MTPP models, for example, censoring where each event has a type-specific probability of being missing [Mei et al., 2019].

In this paper we focus on a different problem, the problem of making predictions when some, or all, marks are censored over (potentially open-ended) intervals of time, i.e., there

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

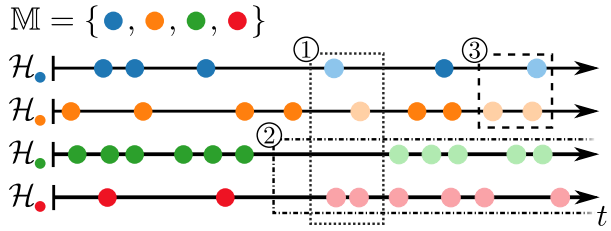

Figure 1: Visualization of an example sequence with four possible marks. $\mathbb{M}$ is the vocabulary of possible event types, $\mathcal{H}_k$ is the history of events with types equal to k. Boxes over sequences represent different modes of censoring that could occur during generation: (1) mark-agnostic censoring for a particular interval, (2) censoring of green and red marks over an open interval, and (3) censoring of blue and orange marks over a finite interval. The occurrence of an event or the total count of events during an interval are not known, differentiating our scenarios from the typical "interval censoring" in survival analysis or MTPPs.

is partial censoring of a specific subset of marks. We will refer to this type of censoring as *mark-censoring*. To our knowledge, there has been no prior work that addresses this problem of adapting MTPPs to mark-censored sequences at inference time. The problem is motivated by the real-world scenario where an MTPP model has been trained on a known set of marks with fully-observed data, but where at prediction time some of the marks (and their associated timing) are no longer observable. For example, in medical data analysis, certain types of events that were measured in the training dataset at a particular hospital might no longer be recorded when the model is deployed at a different hospital. Or, in system monitoring, all events of a certain type could be censored over a window of time due to events such as network and power outages, and accommodating such gaps is important for modeling future dynamics once outages are resolved.

Previous work such as Linderman et al. [2017] focuses on special cases of missingness patterns and/or only applies to specific model architectures (as will be discussed in more detail in Section 2). In contrast, our work is able to handle all of the scenarios shown in Figure 1. The basis of our approach is a novel marginalization technique that can correct the intensity for the censoring of marks. Our proposed method is *model-agnostic* in that it can be applied to any MTPP with a well-defined intensity function. We demonstrate this by employing our method on different types of MTPP models and evaluating predictive performance and simulation behavior under a censored-mark regime.

2 RELATED WORK

A broad range of temporal censoring scenarios have been studied in the literature, such as asynchronous event times [Upadhyay et al., 2018, Trouleau et al., 2019] and interval-censored point process data [Fan, 2009, Rizoiu et al., 2022]. Here we focus the discussion of related work to MTPPs where the marks are from a fixed vocabulary. Existing work on missingness in this context can broadly be divided into three categories.

The first category considers various incomplete intervals, regardless of event types, and focuses on novel tasks such as imputing missing events and sequential representation learning. For example, Shchur et al. [2019] proposed a flow-based mixture model that enables closed-form sampling and handles missing data through imputation. Xu et al. [2017] assumes that a proportion of each short doubly-censored event sequence is observed, and in turn proposes a sampling-stitching data synthesis method based on parametric Hawkes processes to sample long training sequences that improve predictions.

The second category considers the scenario in which each individual event, regardless of mark or time of occurrence, has a chance of being censored. For the Hawkes process, for example, sampling methods were developed to identify latent structure in the data [Shelton et al., 2018] or to correct for biased marks that are underrepresented [Zhou and Sun, 2021]. In neural settings, Gupta et al. [2021, 2022] proposed the use of two MTPPs to model missing events in order to make better predictions. Mei et al. [2019] proposed bidirectional-LSTM models that are conditioned on future observations to apply particle smoothing to impute unobserved events.

The third category of prior work assumes that events are observed but the mark and/or the exact event time is unknown. For instance, Deutsch and Ross [2020] developed an approximate Bayesian algorithm to fit Hawkes processes in the presence of noisy event times, and Calderon et al. [2021] addressed partially interval-censored Hawkes processes, where the total event counts on the censored intervals are available. For the case of Hawkes models, Linderman et al. [2017] imputed latent marks and developed a sequential Monte Carlo approach for latent Hawkes processes that can also be applied to multiple types of censoring.

In summary, previous approaches to censoring in MTPPs either focus on specific types of missingness mechanisms during training time or focus on one specific type of model such as parametric or neural Hawkes process models. In contrast, our approach considers a broad range of interval- and mark-censoring mechanisms (see Fig. 1) and is model-agnostic in that it can work with any MTPP model with a marked intensity function at prediction time. Furthermore, the results of our method yield a well-defined intensity function of a MTPP that can be used just the same as any other MTPP, meaning various statistics can be computed such as expected next event (time and mark), log likelihood of partially observed sequences, etc.

3 MARK-CENSORED TEMPORAL POINT PROCESSES

3.1 PRELIMINARIES

Notation Let $\tau_1, \tau_2, \dots \in \mathbb{R}_{\geq 0}$ be a sequence of continuous random variables that are ordered, or more formally $\forall i : \tau_i < \tau_{i+1}$. These variables represent the time of occurrence for events of interest. Alongside each time of an event is an accompanying piece of information, such as a label or location, that is commonly referred to as a *mark*. We will represent each mark as a random variable drawn that takes on discrete values from a fixed set of M values: $\kappa_i \in \mathbb{M} \equiv \{1, \dots, M\}$.

Let the *history* of events up until, but not including, time t be denoted as $\mathcal{H}(t) = \{(\tau_i, \kappa_i) \,|\, \tau_i < t \text{ for } i = 1, 2, \dots\}$. This implies that $\mathcal{H}(\tau_i) = \{(\tau_1, \kappa_1), \dots, (\tau_{i-1}, \kappa_{i-1})\}$. For our purposes, we will often refer to histories over specific ranges of time such as $\mathcal{H}[a, b)$ for all events with times occurring in the interval $[a, b)$. Additionally, it is often convenient to consider *mark-specific* histories (i.e., sequences that only contain events of specified marks). These will be denoted as either $\mathcal{H}_A := \{(\tau, \kappa) \in \mathcal{H} \,|\, \kappa \in A\}$ or $\mathcal{H}_k := \{(\tau, \kappa) \in \mathcal{H} \,|\, \kappa = k\}$ for $A \subset \mathbb{M}$ and $k \in \mathbb{M}$.

Marked Temporal Point Processes The generative mechanisms for these event sequences are generally referred to as *marked temporal point processes* (MTPPs). MTPP models define a probability distribution over a given sequence of N events, $p(\mathcal{H}[0, \tau_N])$.[1] These models are typically constructed in an autoregressive fashion,

$$p(\mathcal{H}[0, \tau_N]) = \prod_{i=1}^{N} p(\tau_i, \kappa_i \,|\, \mathcal{H}[0, \tau_{i-1}]),$$

where the joint distribution for the next event (τ_i, κ_i) conditioned on all prior events is modeled by the expected, instantaneous rate of change for each mark. This is referred to as the *marked intensity function* and is defined formally as

$$\lambda_k(t \,|\, \mathcal{H}(t))dt := \mathbb{E}_p \left[\mathbb{1}(|\mathcal{H}_k[t, t + dt)| = 1) \,|\, \mathcal{H}(t) \right].$$

For brevity, we typically use the following $*$ convention to suppress the conditional: $\lambda_k^*(t) := \lambda_k(t \,|\, \mathcal{H}(t))$. Additionally, the following notation will be used to represent the sum of different marked intensities: $\lambda_A^*(t) := \sum_{k \in A} \lambda_k^*(t)$ for $A \subset \mathbb{M}$. Note that these functions not only condition on the preceding events, but also on the fact that no events have occurred since the last event up until time t, i.e., $p(\cdot \,|\, \mathcal{H}[0, t)) \neq p(\cdot \,|\, \mathcal{H}[0, \tau_{i-1}])$.

The *total intensity function* $\lambda^*(t) := \lambda_{\mathbb{M}}^*(t)$, also referred to as the *ground intensity*, is sufficient to describe the timing of the next event τ_i. The distribution of the mark conditioned on the timing of the next event is naturally described as $p(\kappa_i = k \,|\, \tau_i = t, \mathcal{H}(t)) \equiv \frac{\lambda_k^*(t)}{\lambda^*(t)}$. We will be assuming that the native output of any model we are working with will produce a vector of marked intensity functions over the mark space $\mathbb{M}$ evaluated at time t.

Lastly, the likelihood of a given sequence $\mathcal{H}$ of length N over an observation window $[0, T]$ can be computed in terms of intensity values:

$$p(\mathcal{H}[0, T]) = \left(\prod_{i=1}^{N} \lambda_{\kappa_i}^*(\tau_i) \right) \exp\left(- \int_0^T \lambda^*(s) ds \right). \quad (1)$$

Sampling Any well-behaved MTPP can be easily sampled by using a thinning procedure [Ogata, 1981], if not directly. This procedure relies on the fact that the superposition of two point processes can be characterized as another point process whose total intensity is the sum of individual total intensities. As such, one can sample candidate event times from a homogenous Poisson process with rate D that dominates the total intensity of the MTPP of interest. These times will be accepted iteratively with probability $\lambda^*(t)/D$, and subsequent marks are sampled from $p(\kappa_i = k \,|\, \tau_i = t, \mathcal{H}(t))$.

3.2 MARK-CENSORED SUB-PROCESS

Problem Statement Assume that we have access to a trained MTPP with intensity functions $\lambda_k^*(t)$ for $k \in \mathbb{M}$. We are interested in performing inference on such a model in the presence of censoring. In particular, we are interested in a type of censoring we term *mark-censoring* in which only events of types $k \in \mathbb{O} \subset \mathbb{M}$ are observed, while all events of types $k \in \mathbb{C} := \mathbb{M} \setminus \mathbb{O}$ are censored and unobserved. In particular, we assume in mark-censoring that we know (a) the time-interval where censoring occurs and (b) which kinds of marks are missing (e.g., knowing the time intervals and colors of marks in the censoring boxes displayed in Fig. 1). Below we develop the framework for the case when censoring takes place over all of time (i.e., $t \in [0, \infty)$); however, as we will discuss later in this section, the general approach can be directly applied to a range of more complicated censoring schemes (such as those illustrated in Fig. 1).

On Censoring The term "censoring" can be quite a loaded concept with regards to statistical models. In our work we assume the absence of certain marks over a time interval to correspond to *missing completely at random* (MCAR) [Heitjan and Basu, 1996], i.e., we assume that the realized sequence $\mathcal{H}$ (both observed and unobserved portions) are independent of *why* it is censored in the first place. We leave handling of more informative censoring to future work.

[1] For brevity and consistent notation, we will be using $p(\cdot)$ in reference to both probability density and mass when appropriate.

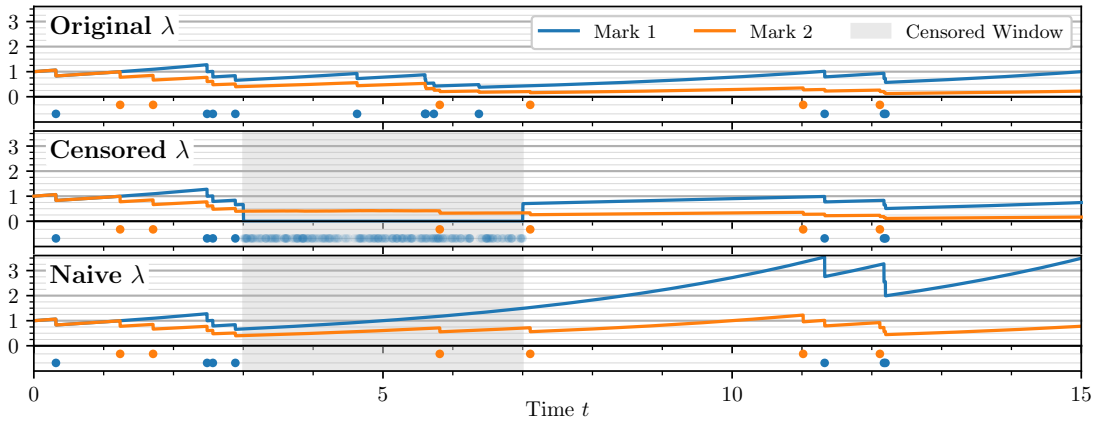

Figure 2: Intensity visualizations (lines) alongside conditioned sequences (dots) for a sequence sampled from a self-correcting point process (top), the same process with blue marks censored from time 3 to 7 (middle), and the naive intensity results for the censored sequence (bottom). The middle sequence displays both the observed sequence as opaque dots and the various censored continuations sampled from the importance distribution as transparent dots. Note that the intensity of the censored mark (blue) after the censoring interval (at time 7) does not necessarily equal the intensity before censoring (at time 3).

Censored Intensity Function Since we have access to the original MTPP, which models the entire distribution for event sequences as a whole, embedded within this model is a well-defined sub-process that represents an MTPP that only observes events of types $k \in \mathbb{O}$. We refer to this embedded model as a *mark-censored sub-process*. This sub-process can be thought of as the original model with the censored information marginalized out of it. Had this sub-process been our intended model from the beginning, we could have achieved comparable results by censoring the original training data and training a model on what remains. There is one key difference, however, which is that the mark-censored sub-process still allows for conditioning on events of types $\mathbb{C}$ even if they are censored moving forward in time (e.g., in the case that the censoring interval only started at time $t > 0$ instead of at $t = 0$—see case 3 in Fig. 1).

The censored sub-process is a fully-fledged MTPP, and as such it has its own set of marked intensity functions. We will denote these as $\underline{\lambda}_k^*(t)$ for $k \in \mathbb{O}$ (should $k \in \mathbb{C}$ then $\underline{\lambda}_k^*(t) = 0$). Likewise, the total intensity for a censored sub-process is defined as $\underline{\lambda}^*(t) := \underline{\lambda}_{\mathbb{O}}^*(t)$. These will be referred to as the *censored intensity* from here forward. Note that for any MTPP with well-defined intensity functions λ_k^*, by the point process superposition property it is justified for the censored intensity $\underline{\lambda}_k^*$ to exist for any arbitrary censoring [Daley and Vere-Jones, 2003].

High Level Intuition for Censored Intensity Later in this section we will present a formal definition of the censored intensity, as well as a tractable estimator for it that solely relies on the original underlying MTPP with likelihood p and intensity $\lambda_k^*(t)$ functions for $k \in \mathbb{M}$. However, prior to presenting these, we will first give an informal overview to help understand the arguments at a high level.

We start by recognizing that we are interested in obtaining the intensity at time t for a censored point process where we only observe events of types $k \in \mathbb{O}$ and no events of types $k \in \mathbb{C}$. To accomplish this, we would prefer to directly marginalize out all possible sequences of $\mathcal{H}_{\mathbb{C}}(t)$; however, for most MTPPs this is unobtainable analytically. Instead, we can approximate the censored intensity $\underline{\lambda}_k^*(t)$ for $k \in \mathbb{O}$ with the original intensity by simply sampling a possible sequence $\tilde{\mathcal{H}}_{\mathbb{C}}(t)$ from the original point process:

$$\underline{\lambda}_k^*(t) \approx \lambda_k(t \,|\, \mathcal{H}_{\mathbb{O}}(t), \tilde{\mathcal{H}}_{\mathbb{C}}(t)),$$

where $\tilde{\mathcal{H}}_{\mathbb{C}}(t) \sim p(\cdot \,|\, \mathcal{H}_{\mathbb{O}}(t))$. Naturally, we cannot directly perform this sampling, so we will do the next best thing and simply sample from the model as usual except that we will prevent any new event with types $k \in \mathbb{O}$ from occurring (i.e., set $\lambda_k^*(t) = 0$ when sampling).

To get a better approximation, this should be done many times with different sampled trajectories: $\tilde{\mathcal{H}}_{\mathbb{C}}^{(i)}(t)$ for $i = 1, \ldots, n$. One could simply compute a standard average where $\underline{\lambda}_k^*(t) \approx 1/n \sum_{i=1}^n \lambda_k(t \,|\, \mathcal{H}_{\mathbb{O}}(t), \tilde{\mathcal{H}}_{\mathbb{C}}^{(i)}(t))$; however, since we didn't sample $\tilde{\mathcal{H}}_{\mathbb{C}}^{(i)}(t)$ perfectly from the model without adjustments we must account for the fact that some samples will be more likely under the original model than others.

As such, we can instead perform a weighted average:

$$\underline{\lambda}_k^*(t) \approx \frac{\sum_{i=1}^n \lambda_k\left(t \,|\, \mathcal{H}_{\mathbb{O}}(t), \tilde{\mathcal{H}}_{\mathbb{C}}^{(i)}(t)\right) \omega\left(\tilde{\mathcal{H}}_{\mathbb{C}}^{(i)}(t)\right)}{\sum_{i=1}^n \omega\left(\tilde{\mathcal{H}}_{\mathbb{C}}^{(i)}(t)\right)}$$

where $\omega(\cdot)$ determines the weight of a sampled trajectory. We define this weight to be the probability of the imposed

sampling restriction (i.e., no *new* events of types $k \in \mathbb{O}$ allowed) being satisfied under the original model. This can be computed for a given sample and is equal to

$$\omega(\tilde{\mathcal{H}}_{\mathbb{C}}(t)) = \exp\left(-\int_0^t \lambda_{\mathbb{O}}(s \,|\, \mathcal{H}_{\mathbb{O}}(s), \tilde{\mathcal{H}}_{\mathbb{C}}(s))ds\right).$$

As an illustration of this censored intensity $\underline{\lambda}_k^*(t)$, Fig. 2 shows the original, censored, and naive intensities for an example sequence sampled from a self-correcting process. After the censoring interval (in gray) ends at $t = 7$, the censored intensity tracks the original true intensity (top) much more closely than the naive intensity (bottom) does. In this context, naive intensity is referring to the original intensity being computed while treating the partially observed sequence $\mathcal{H}_{\mathbb{O}}$ as if it were the fully observed sequence $\mathcal{H}$.

The approximation of $\underline{\lambda}_k^*(t)$ is for finite samples and is a ratio estimator [Tin, 1965]. Taking the limit as $n \to \infty$ converts each summation into an expected value with respect to the proposal distribution, as ratio estimators are consistent. This description matches what will formally be derived below in Eq. (4). Please refer to the Appendix for an in depth analysis on the bias and variance of this estimator when using finite samples.

Formal Definition of $\underline{\lambda}$ Without loss of generality, we will assume that any prior events being conditioned on have been shifted to end at $t = 0$ such that $\mathcal{H}(0)$ contains all of the previous events. It can be shown that the censored intensity function for the sub-process is just a specific marginalization of the original intensity function:

$$\begin{aligned}
&\underline{\lambda}_k^*(t) := \underline{\lambda}_k(t \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset) \text{ for } k \in \mathbb{O} \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} p(\text{hit}(k) \in [t, t+\Delta) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset) \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathbb{E}_{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)} \big[\\
&\qquad p(\text{hit}(k) \in [t, t+\Delta) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset, \mathcal{H}_{\mathbb{C}}[0,t)) \big] \\
&= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \mathbb{E}_{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)} \big[\\
&\qquad p(\tau_i \in [t, t+\Delta), \kappa_i = k \,|\, \mathcal{H}(t)) \big], \quad |\mathcal{H}(t)| = i-1 \\
&= \mathbb{E}_{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)} \big[\\
&\qquad \lim_{\Delta \downarrow 0} \frac{1}{\Delta} p(\tau_i \in [t, t+\Delta), \kappa_i = k \,|\, \mathcal{H}(t)) \big] \text{ by DCT} \\
&= \mathbb{E}_{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)} \left[\lambda_k^*(t)\right]
\end{aligned}$$

where in this context, hit(k) refers to the first occurrence time of event k, and $\mathcal{H}(t) := \mathcal{H}(0) \cup \mathcal{H}_{\mathbb{O}}[0,t) \cup \mathcal{H}_{\mathbb{C}}[0,t)$. The Dominated Convergence Theorem (DCT) holds true because we assume that there exists some value D that is greater than $\lambda_k^*(t)$ for any given t. Note that this assumption is typically made to sample from arbitrary MTPPs.

Tractable Estimation of Censored Intensity To approximate the censored intensity function $\underline{\lambda}_k^*(t)$, we need to perform a Monte Carlo estimation on the above derived expected value, $\mathbb{E}_{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)} \left[\lambda_k^*(t)\right]$. The only issue is that we cannot directly sample from $p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)$ due to the autoregressive nature of MTPPs.

Consider the proposal distribution q which is a MTPP with intensity function

$$\mu_k^*(t) = \begin{cases} 0 & \text{if } k \in \mathbb{O} \text{ and } t \geq 0 \\ \lambda_k^*(t) & \text{otherwise.} \end{cases} \tag{2}$$

This can essentially be thought of as the original MTPP prior to censoring, and then during sampling it only produces sequences of events that cannot be observed. The likelihood for a sequence under this distribution is computed as follows:

$$\begin{aligned}
&q(\mathcal{H}_{\mathbb{C}}[0,t)) := q(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0)) \\
&= \left[\prod_{i=1}^{N} \mu_{\kappa_i}^*(\tau_i)\right] \exp\left(-\int_0^t \mu^*(s)ds\right) \\
&= \left[\prod_{i=1}^{N} \lambda_{\kappa_i}^*(\tau_i) \mathbb{1}(\kappa_i \in \mathbb{C})\right] \exp\left(-\int_0^t \lambda_{\mathbb{C}}^*(s)ds\right)
\end{aligned} \tag{3}$$

where $|\mathcal{H}_{\mathbb{C}}[0,t)| = N$. Note that the proposal distribution has the same support as $p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)$.[2]

Using importance sampling with this proposal distribution, we can see that the censored intensity becomes tractable:

$$\begin{aligned}
&\underline{\lambda}_k^*(t) = \mathbb{E}_{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)} \left[\lambda_k^*(t)\right] \\
&= \mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\lambda_k^*(t) \frac{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0), \mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)}{q(\mathcal{H}_{\mathbb{C}}[0,t))}\right] \\
&= \mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\lambda_k^*(t) \frac{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0)) \mathbb{1}(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)}{p(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset \,|\, \mathcal{H}(0)) q(\mathcal{H}_{\mathbb{C}}[0,t))}\right] \\
&= \frac{\mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\lambda_k^*(t) \frac{\left[\prod_{i=1}^{N} \lambda_{\kappa_i}^*(\tau_i)\right] \exp\left(-\int_0^t \lambda^*(s)ds\right)}{\left[\prod_{i=1}^{N} \lambda_{\kappa_i}^*(\tau_i) \mathbb{1}(\kappa_i \in \mathbb{C})\right] \exp\left(-\int_0^t \lambda_{\mathbb{C}}^*(s)ds\right)}\right]}{p(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset \,|\, \mathcal{H}(0))} \\
&= \frac{\mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\lambda_k^*(t) \exp\left(-\int_0^t \lambda_{\mathbb{O}}^*(s)ds\right)\right]}{p(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset \,|\, \mathcal{H}(0))}.
\end{aligned}$$

Note that in this context $p(\mathcal{H}_{\mathbb{C}})$ is equivalent to the likelihood of $\mathcal{H}_{\mathbb{C}}$ under the original model p, *as if it were a fully observed sequence* $\mathcal{H}$.

Now the expected value can be approximated with easy-to-access Monte Carlo samples. The only immediate problem is evaluating $p(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset \,|\, \mathcal{H}(0))$ as this does not have a closed form solution; however, as in the recent approach of Boyd et al. [2023], we can estimate this statement using importance sampling. Interestingly, we can actually utilize the exact same proposal distribution q as specified in Eqs. (2)

[2] It follows that $\mathbb{E}_{q(\mathcal{H}_{\mathbb{C}})[0,t)} \left[\mathbb{1}(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset)\right] = 1$, which becomes useful for subsequent derivations.

and (3) to represent $p(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset \,|\, \mathcal{H}(0))$ as a tractable expected value:

$$\begin{aligned} p(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset \,|\, \mathcal{H}(0)) &= \mathbb{E}_{p(\mathcal{H}[0,t) \,|\, \mathcal{H}(0))} \left[\mathbb{1}(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset) \right] \\ &= \mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\mathbb{1}(\mathcal{H}_{\mathbb{O}}[0,t) = \emptyset) \frac{p(\mathcal{H}_{\mathbb{C}}[0,t) \,|\, \mathcal{H}(0))}{q(\mathcal{H}_{\mathbb{C}}[0,t))} \right] \\ &= \mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\exp\left(- \int_0^t \lambda^*_{\mathbb{O}}(s) ds \right) \right]. \end{aligned}$$

Thus, the censored intensity can be ultimately represented as a ratio of two expected values:

$$\implies \underline{\lambda}^*_k(t) = \frac{\mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\lambda^*_k(t) \exp\left(- \int_0^t \lambda^*_{\mathbb{O}}(s) ds \right) \right]}{\mathbb{E}_{q(\mathcal{H}_{\mathbb{C}}[0,t))} \left[\exp\left(- \int_0^t \lambda^*_{\mathbb{O}}(s) ds \right) \right]}. \quad (4)$$

In practice, this censored intensity can be approximated using Monte Carlo (MC) estimates for both the numerator and denominator.

It is worth reiterating that this estimator, which accounts for the censoring of marks $\mathbb{C}$ at inference time, only requires a trained MTPP along with samples from it. No further training, additional models, or specific architectures are required to properly deal with the censoring.

More Complex Censoring Regimes All of the derivations thus far have been focused on having a static set of marks $\mathbb{C}$ being censored for an indefinite amount of time; however, there are many other types of censoring that can occur for a given MTPP. For example, the censoring could occur over a specific window of time for either some or all marks $\mathbb{M}$. This could occur, for instance, in settings where the connection is briefly lost to some or all sensors in a system. Furthermore, censoring could occur multiple times over different windows, and the marks being censored across each window need not be the same from censoring to censoring. See Fig. 1 for example censoring scenarios.

We can easily extend our previous results to cover the most general case allowing for censoring over arbitrarily many time windows and arbitrarily different censored marks. To do so, first we will define the censoring schedule. The observed and censored marks, $\mathbb{O}$ and $\mathbb{C}$, are no longer static and will potentially change over time. This will be represented via $\mathbb{O}(t), \mathbb{C}(t) \subset \mathbb{M}$ for $t \geq 0$. This results in the proposal distribution q now being characterized by the intensity function $\mu^*_k(t) = \lambda^*_k(t) \mathbb{1}(k \in \mathbb{C}(t))$. Lastly, the resulting censored intensity estimate also accommodates this dynamic censoring:

$$\underline{\lambda}^*_k(t) = \frac{\mathbb{E}_{q(\mathcal{H}[0,t))} \left[\lambda^*_k(t) \exp\left(- \int_0^t \lambda^*_{\mathbb{O}(s)}(s) ds \right) \right]}{\mathbb{E}_{q(\mathcal{H}[0,t))} \left[\exp\left(- \int_0^t \lambda^*_{\mathbb{O}(s)}(s) ds \right) \right]}. \quad (5)$$

This result is achieved effectively for free as the censored intensity $\underline{\lambda}^*_k(t)$ in the static setting is technically defined individually for any given moment in time t, making the swap from $\mathbb{O}$ to $\mathbb{O}(t)$ and $\mathbb{C}$ to $\mathbb{C}(t)$ for each t well defined.

More Complex Mark Spaces $\mathbb{M}$ Our setting of interest has the marks being modeled come from some discrete, finite mark space $\mathbb{M} := \{1, \ldots, M\}$; however, that does not have to be the case. We can easily extend our method to apply for more complex mark spaces. Consider an arbitrary mark space $\mathbb{M}$ which could be finite, continuous, high-dimensional, etc. and let ν be a reference measure for $\mathbb{M}$ (e.g., the Lebesgue measure for $\mathbb{M} \equiv \mathbb{R}$). Assume we have a MTPP model with marked intensity function $\lambda^*(t, m)$ for $m \in \mathbb{M}$, and that under our framework we know the observed and censored portions of the mark space at any given time, $\mathbb{O}(t) \subset \mathbb{M}$ and $\mathbb{C}(t) := \mathbb{M} \setminus \mathbb{O}(t)$ respectively. From this, the censored intensity defined in Eq. (4) can be readily used by letting $\lambda^*_{\mathbb{O}(t)}(t) := \int_{\mathbb{O}(t)} \lambda^*(t, m) d\nu(m)$ which can either be computed analytically or estimated with Monte-Carlo samples. The proposal distribution stays the same as previously defined and samples from it can be achieved easily using either rejection sampling on top of the typical thinning procedure.

4 EXPERIMENTS

We investigate experimentally the impact that mark-censoring has on various MTPP models and the ability of our proposed marginalization method to handle such censoring relative to baseline. Our investigations are carried out across both classical parametric models and neural network-based models on both synthetic and real-world data, respectively. We find, as a whole, that in the presence of mark-censoring, the inference ability of a model (i.e., assigning likelihood to observed sequences) suffers significantly in comparison to properly accounting for the missing data via our method. Not surprisingly, we also find that our method yields larger improvements as the information being censored becomes more influential with respect to the information observed.

We also investigate the effect that mark-censoring has on next event (time and mark) prediction. We observe in general systematic differences that our mark-censored model has on these predictions, with positive improvements in real-world settings. Lastly, we also perform a sensitivity analysis on the effect of both the number of sequences sampled as well as the resolution used in estimating integrals has on our method. We find that our method is typically fairly robust to these hyperparameters. More details and exact results for both of these experiments can be found in the Appendix.

Censoring In each of the experiments, we analyze the performance of models using various sequences $\mathcal{H}(T)$ of differing lengths T. For the synthetic setting, we utilize sequences that have been drawn from the given models. For the real-world data, we use held-out sequences from the dataset that a given model was trained on.

For every sequence being used, we filter out events according to a particular censoring scheme that is selected for each

sequence individually to produce $\mathcal{H}_{\mathbb{O}}(T)$. To ensure that the chosen censoring scheme is relevant for a given sequence, we randomly select a non-empty subset $\mathbb{C}(t)$ of the unique marks that actually appear in $\mathcal{H}(T)$ for $t \in [0, T]$. The proportion of marks to censor, relative to the total number of unique marks in each sequence $\mathcal{H}(T)$, which we will refer to as γ, is varied based on the particular sequence for the experiment being conducted.

It is important to note that since information in $\mathcal{H}(T)$ is informing the censoring scheme $\mathbb{C}$ that we technically no longer have data that is MCAR. As we will see, in spite of violating this assumption the mark-censored model still yields substantial performance gains.

Methods & Metric of Interest For the main set of experiments, we primarily compared two approaches. Both rely on an existing MTPP and are used to calculate the likelihood of a given observed sequence $\mathcal{H}_{\mathbb{O}}(T)$.[3]

The first approach is our proposed mark-censored model, using $\underline{\lambda}_k^*$ for $k \in \mathbb{O}$. Since this is a well-defined MTPP, we can calculate the likelihood of $\mathcal{H}_{\mathbb{O}}(T)$ using Eq. (1) in conjunction with the censored intensity. Results for this method will be labeled as "Censored." Synthetic and real-world experiments use 128 and 64 MC samples to estimate the censored intensity respectively; both use 1024 integration points for numerically estimating integrals.

The second approach is a baseline method for comparison, based on a slight adaptation to the original model that takes advantage of knowing what marks are being censored for a given sequence. This method uses the original intensity $\lambda_k^*(t)$ for $k \in \mathbb{O}$ and sets the intensity to be 0 when $k \in \mathbb{C}$. Results for this method will be labeled as "Baseline." In general, we expect the two methods to be comparable should $p(\mathcal{H}_{\mathbb{C}}(T) = \emptyset \mid \mathcal{H}_{\mathbb{O}}(T)) \approx 0$ as the two methods would produce similar intensity values.

We do not include results where we evaluate the likelihood of the observed sequence $\mathcal{H}_{\mathbb{O}}(T)$ as if it were a fully observed, uncensored sequence under the original model. Since intensity values are always non-negative, likelihood values using this approach will *never* be better than the baseline. Because of this, we only compare against the baseline as it effectively captures the original model's inference capabilities while still managing to leverage information about the mark-specific censoring scheme to some degree. Note also that none of the methods discussed earlier in Related Work are used as baselines since none are applicable to mark-censoring and model-agnostic.

Results are reported as likelihood ratios between the censored method and the baseline method for individual observed sequences. These ratios directly quantify how much more likely the censored method perceives a sequence to be relative to the baseline. Values above 1 are evidence in favor of the censored method, and below 1 for the baseline. It should be noted that the sequences used in these experiments are censored over the entire observation window $[0, T]$.

[3]Previous works are not compared against in these experiments due to them largely having different goals and setups (such as learning from censored data during training time or imputing missing data), as well as typically not having a proper likelihood.

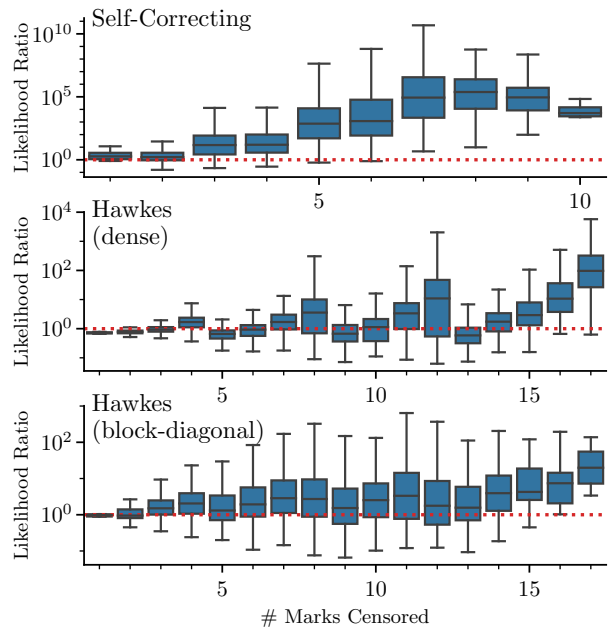

Figure 3: Distributions of likelihood ratios across number of marks censored for the duration of the sequences used for synthetic experiments with self-correcting, Hawkes (dense), and Hawkes (block-diagonal) models. Values greater than 1 indicate higher likelihoods under the mark-censored model.

4.1 EXPERIMENTS ON SYNTHETIC DATA

Models We evaluate our method on randomly instantiated parametric MTPPs including Hawkes processes [Hawkes, 1971] and self-correcting processes [Isham and Westcott, 1979] (also known as stress release model [Zheng and Vere-Jones, 1991]), where the sampled sequences are evaluated on the same model.

For Hawkes processes with exponential kernels, the intensity has the form $\lambda_k^*(t) = \mu_k + \sum_{\tau,\kappa \in \mathcal{H}(t)} \phi_{\kappa,k}(t - \tau)$. The kernel can be expressed as $\phi_{i,j}(z) = \alpha_{ij} \exp(-\beta_{ij} z)$, where parameters $\alpha_{ij}, \beta_{ij} > 0$ for $i, j \in \mathbb{M}$ specify the excitation effects and decay rates respectively that events of type i have on events of type j. We consider two different instantiations of this type of model; both with 20 marks. We refer to the first type as "Hawkes (dense)" with all parameters drawn from the following distributions: $\alpha_{ij} \overset{iid}{\sim} \text{Unif}[0.075, 0.2]$, $\beta_{ij} \overset{iid}{\sim} \text{Unif}[0.4, 1.2]$, and

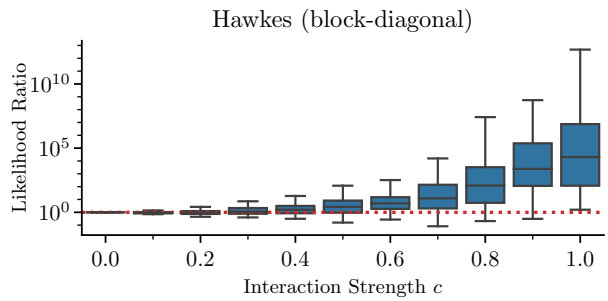

Figure 4: Distributions of likelihood ratios for a block-diagonal Hawkes model with varying interaction strengths applied to off-diagonal α terms. Values greater than 1 indicate higher likelihoods under the mark-censored model compared to the baseline.

$\mu_k \overset{iid}{\sim} \text{Unif}[0.1, 0.5]$. To better emulate realistic settings in which events correlate strongly with other events of certain types, we also consider a sparsely-parameterized version which we refer to as "Hawkes (block-diagonal)" [Wu et al., 2020]. This model is instantiated by drawing $\alpha_{ij} \overset{iid}{\sim} \text{Unif}[0.3, 0.8]$ when $\lfloor \frac{i-1}{5} \rfloor = \lfloor \frac{j-1}{5} \rfloor$ and $\alpha_{ij} = 0$ otherwise.[4] This effectively imposes a block-diagonal structure on the matrix $\{\alpha_{ij}\}$, resulting in four subgroups of correlated marks. Values for μ and β are drawn similarly to the dense model.

In contrast, self-correcting processes use the intensity function $\lambda_k^*(t) = \exp\left(\eta_k t - \sum_{\tau,\kappa \in \mathcal{H}(t)} \delta_{\kappa k}\right)$, where $\delta_{\kappa k} > 0$ determines the inhibition that past events of type κ have on future events of type k. The model used for this class also has 20 marks and is instantiated by drawing weights $\delta_{ij} \overset{iid}{\sim} \text{Unif}[0.3, 0.8]$. Values for η are drawn similarly to μ.

Results We evaluated the likelihood ratio of 1000 censored sequences on all three models with interaction strength fixed at 0.5 (a scalar that controls the interaction between events of different types) for each value $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$. Each sequence prior to censoring was sampled from each model (self-correcting, Hawkes (dense), and Hawkes (block-diagonal)) over the time window $t \in [0, 2]$ and contains at most 200 events. These results are shown in Fig. 3, where the likelihood ratio of the censored method compared to the baseline is visualized with respect to the number of marks censored. We see a systematic improvement in the estimated likelihood when using the mark-censored model. Furthermore, the improvement increases as more information is censored; however, it is clear that the improvement depends on the relationship between events and the underlying model dynamics (i.e., the form of λ) as noted by the difference in results between models.

To further investigate this, for the block-diagonal Hawkes model we artificially modulate the interaction strength between events of different types. To do this, we performed the same likelihood ratio evaluation on 1000 sequences with $\gamma = 0.5$ using the same block-diagonal Hawkes model but with $\alpha'_{ij} := c\alpha_{ij}$ if $i \neq j$ and α_{ij} otherwise for each value of $c \in \{0.1, 0.2, \ldots, 1.0\}$. This results in 10 different models that have the same diagonal values in $\boldsymbol{\alpha}$ but different scales of off-diagonal values. The results in Fig. 4 clearly demonstrate that properly accommodating mark-censored sequences yields the biggest impact when there is high correlation between observed and censored events.

4.2 EXPERIMENTS ON REAL-WORLD DATA

Models Many real-world data involve working with large vocabularies of possible marks, $|\mathbb{M}| = M$. Because of this, it can often be more parameter efficient to train a neural network based MTPP rather than a classical parametric one. The model architecture of choice for our experiments is the neural Hawkes process, a continuous-time RNN that takes inspiration from the parametric Hawkes process [Mei and Eisner, 2017]. Details on model hyperparameters, optimizer, training regime, etc. can be found in the Appendix.

Datasets We evaluate our censoring method on neural Hawkes models that have been trained individually on four different datasets. The **Taobao** user behavior dataset [Zhu et al., 2018] contains page-viewing records of different categories ($M = 1000$) from users on an e-commerce platform. The **Reddit** dataset [Baumgartner et al., 2020] contains comments that users have made on various communities ($M = 1000$) on the social media website `reddit.com`. **MemeTracker** [Leskovec et al., 2009] contains records of what websites ($M = 5000$) a common phrase, or meme, was mentioned on over time. Lastly, the **Email** [Paranjape et al., 2017] dataset contains sequences of sender addresses of incoming emails ($M = 808$) for each recipient within a research organization. More information on various aspects of these datasets and details of data preprocessing can be found in the Appendix. The following results are achieved using models that have been sufficiently trained on their respective datasets.

Results We evaluated the likelihood ratio of 1000 held out, censored sequences for each dataset for each value $\gamma \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$. The results are shown in Fig. 5. Similar to the results in the synthetic experiments, we see a systematic trend towards a large improvement in likelihood over censored sequences across the board. This improvement increases significantly as more marks are censored.

[4] Note that different ranges of values were chosen for α between the dense and block-diagonal Hawkes models to normalize the effective rate of events overall. This is done by, in expectation, having the same values for $\sum_{i \in \mathbb{M}} \alpha_{ij}$.

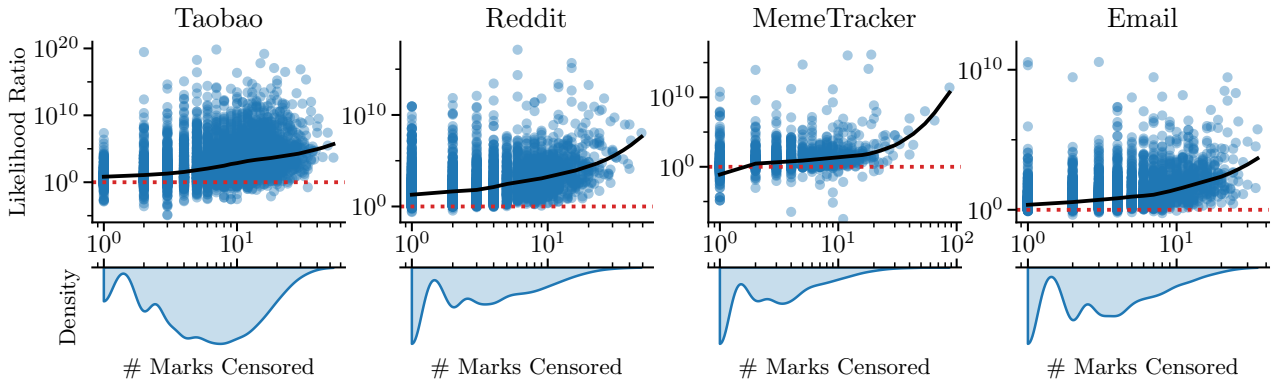

Figure 5: Same setup as Fig. 3 except with results produced on four real-world datasets with trained neural Hawkes models. Note that we display the results with respect to the absolute amount of marks censored rather than the percentage censored as we suspect this has a more direct impact on the likelihood ratios, especially when dealing with sequences that naturally have few unique marks compared to the total mark space $\mathbb{M}$—as is typical in real datasets.

5 CONCLUSIONS

In this work we proposed a novel marginalization technique for inference in the presence of mark-censoring, for any black-box MTPP model trained on complete histories. Our method demonstrates systematic improvements in log-likelihood for both synthetic and real-world data settings.

A limitation of the approach is that it is restricted to prediction time and is not practical for use during training with mark-censored training data. The main hurdle that needs to be overcome to make our method viable in this setting is that current sampling methods for MTPPs are not differentiable. In addition, while our approach is guaranteed to have higher likelihood on average for MTPP models with no misspecification, this guarantee does not hold for misspecified models (e.g., on real data sources)—see Appendix for more details.

Aside from directly addressing these limitations, potential future directions of this work include applying this approach to applications such as assessing good-of-fit and comparing models with different vocabularies, extending the methodology to the continuous mark setting, incorporating more informative censoring schemes (e.g., assuming data is *not* MCAR), and permanently applying censoring via model distillation with a mark-censored process.

Acknowledgements This work was supported by National Science Foundation Graduate Research Fellowship grant DGE-1839285, by an NSF CAREER Award, by the National Science Foundation under award numbers 1900644, 2003237, and 2007719, by the National Institute of Health under awards R01-AG065330-02S1 and R01-LM013344, by the Department of Energy under grant DE-SC0022331, by the HPI Research Center in Machine Learning and Data Science at UC Irvine, and by Qualcomm Faculty awards.

References

Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical Models based on Counting Processes*. Springer, 2012.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social media*, volume 14, pages 830–839, 2020.

Alex Boyd, Yuxin Chang, Stephan Mandt, and Padhraic Smyth. Probabilistic querying of continuous-time event sequences. In *International Conference on Artificial Intelligence and Statistics*, pages 10235–10251. PMLR, 2023.

David R Brillinger. Some examples of random process environmental data analysis. *Handbook of Statistics*, 18: 33–56, 2000.

Pio Calderon, Alexander Soen, and Marian-Andrei Rizoiu. Linking across data granularity: Fitting multivariate Hawkes processes to partially interval-censored data. *arXiv preprint arXiv:2111.02062*, 2021.

D. R. Cox and P. A. W. Lewis. Multivariate point processes. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, page 401. University of California Press, 1972.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: volume I: Elementary Theory and Methods*. Springer, 2003.

Prathamesh Deshpande, Kamlesh Marathe, Abir De, and Sunita Sarawagi. Long horizon forecasting with temporal point processes. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 571–579, 2021.

Isabella Deutsch and Gordon J Ross. Abc learning of Hawkes processes with missing or noisy event times. *arXiv preprint arXiv:2006.09015*, 2020.

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.

Chun-Po Steve Fan. *Local likelihood for interval-censored and aggregated point process data*. University of Toronto, 2009.

Giancarlo Fortino, Antonella Guzzo, Michele Ianni, Francesco Leotta, and Massimo Mecella. Exploiting marked temporal point processes for predicting activities of daily living. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2020.

Vinayak Gupta, Srikanta Bedathur, Sourangshu Bhattacharya, and Abir De. Learning temporal point processes with intermittent observations. In *International Conference on Artificial Intelligence and Statistics*, pages 3790–3798. PMLR, 2021.

Vinayak Gupta, Srikanta Bedathur, Sourangshu Bhattacharya, and Abir De. Modeling continuous time sequences with intermittent observations using marked temporal point processes. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(6):1–26, 2022.

Tobias Hatt and Stefan Feuerriegel. Early detection of user exits from clickstream data: A Markov modulated marked point process model. In *Proceedings of The Web Conference 2020*, pages 1671–1681, 2020.

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

Daniel F Heitjan and Srabashi Basu. Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213, 1996.

Chengkuan Hong and Christian Shelton. Deep Neyman-Scott processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3627–3646. PMLR, 2022.

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

Kazi T Islam, Christian R Shelton, Juan I Casse, and Randall Wetzel. Marked point process for severity of illness assessment. In *Proceedings of the Machine Learning for Healthcare Conference (MLHC)*, pages 255–270. PMLR, 2017.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–506, 2009.

Scott W Linderman, Yixin Wang, and David M Blei. Bayesian inference for latent Hawkes processes. *Advances in Bayesian Inference Workshop, NeurIPS 31*, 2017.

Myrl G Marmarelis, Greg Ver Steeg, and Aram Galstyan. A metric space for point process excitations. *Journal of Artificial Intelligence Research*, 73:1323–1353, 2022.

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.

Hongyuan Mei, Guanghui Qin, and Jason Eisner. Imputing missing events in continuous-time event streams. In *International Conference on Machine Learning*, pages 4475–4485. PMLR, 2019.

Abhishek Mishra and Parv Venkitasubramaniam. Anonymity of a buffer constrained chaum mix: Optimal strategy and asymptotics. In *2013 IEEE International Symposium on Information Theory*, pages 71–75. IEEE, 2013.

Yosihiko Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27 (1):23–31, 1981.

Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610, 2017.

Marian-Andrei Rizoiu, Alexander Soen, Shidi Li, Pio Calderon, Leanne Dong, Aditya Krishna Menon, and Lexing Xie. Interval-censored Hawkes processes. *Journal of Machine Learning Research*, 23(338):1–84, 2022.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. *arXiv preprint arXiv:1909.12127*, 2019.

Christian Shelton, Zhen Qin, and Chandini Shetty. Hawkes process inference with missing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zijian Shi and John Cartlidge. State dependent parallel neural Hawkes process for limit order book event stream prediction and simulation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1607–1615, 2022.

Myint Tin. Comparison of some ratio estimators. *Journal of the American Statistical Association*, 60(309):294–307, 1965.

William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Learning Hawkes processes under synchronization noise. In *International Conference on Machine Learning*, pages 6325–6334. PMLR, 2019.

Ali Caner Türkmen, Yuyang Wang, and Alexander J Smola. Fastpoint: Scalable deep point processes. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 465–480. Springer, 2020.

Utkarsh Upadhyay, Abir De, and Manuel Gomez Rodriguez. Deep reinforcement learning of marked temporal point processes. *Advances in Neural Information Processing Systems*, 31, 2018.

Jing Wu, Anna L Smith, and Tian Zheng. Diagnostics and visualization of point process models for event times on a social network. *arXiv preprint arXiv:2001.09359*, 2020.

Hongteng Xu, Dixin Luo, and Hongyuan Zha. Learning Hawkes processes from short doubly-censored event sequences. In *International Conference on Machine Learning*, pages 3831–3840. PMLR, 2017.

Kang Yang, Xi Zhao, Jianhua Zou, and Wan Du. ATPP: A mobile app prediction system based on deep marked temporal point processes. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 83–91. IEEE, 2021.

Yizhou Zhang, Karishma Sharma, and Yan Liu. VigDdet: Knowledge informed neural temporal point process for coordination detection on social media. *Advances in Neural Information Processing Systems*, 34:3218–3231, 2021.

Xiao-Gu Zheng and David Vere-Jones. Application of stress release models to historical earthquakes from North China. *Pure and Applied Geophysics*, 135:559–576, 1991.

Zihan Zhou and Mingxuan Sun. Multivariate Hawkes processes for incomplete biased data. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 968–977. IEEE, 2021.

Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1079–1088, 2018.

Yada Zhu, Wenyu Chen, Yang Zhang, Tian Gao, and Jianbo Li. Probabilistic framework for modeling event shocks to financial time series. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–8, 2021.