Linking Across Data Granularity: Fitting Multivariate Hawkes Processes to Partially Interval-Censored Data

Pio Calderon, Alexander Soen, and Marian-Andrei Rizoiu

Abstract—The multivariate Hawkes process (MHP) is widely used for analyzing data streams that interact with each other, where events generate new events within their own dimension (via self-excitation) or across different dimensions (via cross excitation). However, in certain applications, the timestamps of individual events in some dimensions are unobservable, and only event counts within intervals are known, referred to as partially interval-censored data. The MHP is unsuitable for handling such data since its estimation requires event timestamps. In this study, we introduce the partially censored multivariate Hawkes process (PCMHP), a novel point process that shares parameter equivalence with the MHP and can effectively model both timestamped and interval-censored data. We demonstrate the capabilities of the PCMHP using synthetic and real-world datasets. First, we illustrate that the PCMHP can approximate MHP parameters and recover the spectral radius using synthetic event histories. Next, we assess the performance of the PCMHP in predicting YouTube popularity and find that the PCMHP outperforms the Hawkes intensity process (HIP) [1]. Comparing with the fully interval-censored HIP, we show that the PCMHP improves prediction performance by accounting for point process dimensions, particularly when there exist significant cross-dimension interactions. Last, we leverage the PCMHP to gain qualitative insights from a dataset comprising daily COVID-19 case counts from multiple countries and COVID-19-related news articles. By clustering the PCMHP-modeled countries, we unveil hidden interaction patterns between occurrences of COVID-19 cases and news reporting.

Index Terms—Partially observed data, popularity prediction, temporal point process.

I. INTRODUCTION

The Hawkes process, introduced by [2], is a temporal point process that exhibits the self-exciting property, i.e., the occurrence of one event increases the likelihood of future events.

Received 5 October 2023; revised 24 May 2024 and 18 September 2024; accepted 17 October 2024. Date of publication 12 November 2024; date of current version 31 January 2025. This work was supported in part by the Australian Department of Home Affairs, the Defence Science and Technology Group, the Defence Innovation Network, the Australian Academy of Science, and the National Science Centre, Poland, under Grant 2021/41/B/HS6/02798. (Corresponding author: Pio Calderon.)

Pio Calderon and Marian-Andrei Rizoiu are with the University of Technology Sydney, Ultimo, NSW 2017, Australia (e-mail: piogabrielle.b.calderon@student.uts.edu.au).

Alexander Soen is with Australian National University, Canberra, ACT 2601, Australia.

This article has supplementary downloadable material available at https://doi.org/10.1109/10.1109/TCSST.2024.3486117, provided by the authors.

Fig. 1. Example of multiplatform interaction between view events on YouTube (red lollipops) and tweets on Twitter (blue lollipops). The data are partially interval-censored, as YouTube does not expose individual views, but only the view counts C_i’s over the predefined intervals [\alpha_i, \alpha_{i+1}] (shown as red rectangles). The dashed lines show the latent branching structure between views and tweets. The red lollipops are also dashed and empty, indicating that YouTube views are not observed.

The Hawkes process is widely applied in both the physical and social sciences. For example, earthquakes are known to be temporally clustered: the mainshock is often the first in a sequence of subsequent aftershocks. In online social media, tweets by influential users typically induce cascades of retweets as the message diffuses over the social network [3]. The multivariate Hawkes process (MHP) [2] extends the univariate process by allowing events to occur in multiple parallel timelines—dubbed as dimensions. These dimensions interact via cross excitation, i.e., events in one dimension can spawn events in the other dimensions. Fig. 1 schematically exemplifies the interaction between two social media platforms: YouTube and Twitter. An initial tweet (denoted as A on the figure) spawns a retweet (B) via self-excitation and a view (C) via cross excitation. The cross excitation goes both ways: the view C generates the tweet D.

Given the event timestamps, we can fit the parameters of the Hawkes process using maximum likelihood estimation (MLE). However, in many practical applications, the event times are not observed, and only counts over predefined time partitions are available. We denote such data as interval-censored. For multivariate data, we denote the case when all dimensions are interval-censored as completely interval-censored. If only a subset of dimensions is interval-censored, we have partially interval-censored data.
One reason for interval-censoring is data availability—for epidemic data [4], we usually observe the aggregated daily counts of reported cases instead of detailed case information. Another reason is space limitations—for network traffic data [5], storing high-resolution event logs is impractical; they are stored as summaries over bins instead. A third reason is data privacy. This is the case for YouTube, as shown in the upper half of Fig. 1, where the individual views are interval-censored, and we only observe aggregated daily counts.

This article tackles three open questions about using the MHP with partially interval-censored data. The first question relates to fitting the process to both event time and interval-censored data. When the data are presented as event times, the MHP can be fitted using MLE [6]. However, if the data are partially or completely interval-censored, MLE cannot fit the MHP process parameters because it lacks the independent increments property [7]. Given interval-censored counts, one could approach fitting the Hawkes process naïvely by sampling event times uniformly over the intervals [8]. However, this quickly hits scalability issues for high-interval-censored counts. For instance, the YouTube videos in our real-world dataset often have millions of views per day. For completely interval-censored univariate data, Rizoiu et al. [7] proposed the mean behavior Poisson (MBP)—an inhomogeneous Poisson process that approximates the mean behavior of the Hawkes process—to estimate the parameters of a corresponding Hawkes process. However, a model and fitting scheme remained elusive for the partially interval-censored data. The question is, can we devise a method to fit the MHP in the partially interval-censored setting?

What are the limits to MHP parameter recovery in the partially interval-censored setting?

The second question relates to modeling and forecasting online popularity across social media platform boundaries. Online popularity has been extensively studied within the realm of a single social media platform—see Twitter [9], [10], [11], [12], YouTube [1, 13], and Reddit [14]—and the self-exciting point processes are the tool of choice for modeling. However, content is often shared across multiple interacting platforms—such as YouTube and Twitter—and we need to account for cross excitation using multivariate processes. However, YouTube only exposes view data as interval-censored, rendering it impossible to use the classical MHP. The Hawkes intensity process (HIP) [1] proposes a workaround and treats the tweet and share counts as external stimuli for views. Its shortcoming is that it cannot model the cross excitation from views to tweets and shares. The question is, can we improve performance in the YouTube popularity prediction task by modeling the views, tweets, and shares through fitting on partially interval-censored data?

The third question concerns analyzing interaction patterns across the online and offline environments, enabling us, for example, to determine whether online activity preempts or reacts to events that happen offline. Previous work has demonstrated the complex link between news and infectious disease outbreaks, notably the 2009 A/H1N1 outbreak in the Shaanxi province in China [15], the 2010 cholera outbreak in Haiti [16], and the early spread of COVID in 2020 in various provinces in China [17]. The association between media and case counts has typically been investigated by examining the cross correlation of the news counts and case counts as paired time series and demonstrating that significant correlations exist when temporal lags are applied. The authors [15], [17] show correlations between news and cases for both positive and negative lags, suggesting that news both had an impact and had been impacted by reported disease counts. Chunara et al. [16] show that news typically lags behind cases; they also showcase how news counts can be used as a proxy for estimating crucial disease measures such as the basic reproduction number $R_0$. This highlights that the connection between news and cases is particularly relevant given that news counts can be retrieved in near real time; in contrast, official case counts reporting is often lagging. In most previous work, uncovering time-series cross correlation is the focus, without building explanatory models to produce nuanced views of the interactions through interpretable parameters. The question is, can we apply MHP on partially interval-censored data to uncover country-level differences in the interplay between recorded daily case counts of COVID-19 and the publication of COVID-19-related news articles?

We address these three questions by introducing the partially censored multivariate Hawkes process (PCMHP), a novel multivariate temporal point process that operates on partially interval-censored data. We answer the first question in Section II-B, where we detail the PCMHP. The event intensity of PCMHP on the interval-censored dimensions is determined by the expected Hawkes intensity, considering the stochastic history of those dimensions conditioned on the event time dimensions. On the event time dimensions, the intensity of PCMHP corresponds to that of the respective Hawkes process. This construction allows us to fit the PCMHP to partially interval-censored data and estimate the parameters of the MHP through parameter equivalence.

We address the second question in Section IV by using PCMHP to predict the popularity of YouTube videos on both YouTube and Twitter. We demonstrate that PCMHP consistently outperforms the related HIP method [1], provides quantification of prediction uncertainty, and extends predictions to all dimensions—unlike HIP, which can only predict the views’ dimension.

We address the third question in Section V by utilizing PCMHP to investigate the relationship between COVID-19 case incidence and news coverage. We fit a country-specific PCMHP for each of the 11 countries using a dataset consisting of reported COVID-19 cases (with interval-censored data) and the publication dates of COVID-19-related news articles during the early stage of the outbreak. We identify three distinct groupings by clustering countries using the fitted PCMHP parameters. In the first group (U.K., Spain, Germany, and Brazil), we observe preemptive news coverage, where an increase in news leads to a rise in cases. The second group (China and France) exhibits reactionary news coverage, with news lagging behind the cases. No significant interaction between news and cases is found in the third group (USA, Italy, Sweden, India, and Philippines).

Implementation available at https://github.com/behavioral-ds/pmbp_implementation.
A. Related Work

A significant portion of the recent literature on the Hawkes process, and on point processes in general, deals with estimation from partially observed data. This problem is nontrivial as standard MLE techniques require the complete dataset.

It was shown in [18] that a sequence of (integer-valued autoregressive time series) INAR(∞)-based family of point processes converges to the Hawkes process. Under this convergence, they concluded that the INAR(∞) is the discrete-time version of the Hawkes process. In a follow-up, Kirchner [19] presented an alternative procedure to MLE, which fits the associated bin-count sequences to the INAR(p) process. As the bin size goes to zero and the order p of the process goes to ∞, the INAR sequence converges to the Hawkes process and the parameter estimates converge to the Hawkes parameters. However, though fitting is performed on count data, convergence only actually occurs for small bin size.

A spectral approach to fitting the Hawkes process given interval-censored data for arbitrary bin size is presented in [20], solving the issue in [19]. Their proposed method is based on minimizing the log-spectral likelihood of the bin-count sequence instead of the usual log-likelihood of the Hawkes process. They showed that optimization converges to the Hawkes parameters under certain assumptions on the kernel.

The sample-based Monte Carlo expectation maximization (MC-EM) algorithm was introduced in [5] and [21] for the univariate and multivariate cases, respectively, which uses sampling to obtain proposals for the hidden event times. They showed that their approach recovers parameters more reliably than the INAR(p) estimates from [19] in synthetic experiments. Another sample-based approach, the recursive identification with sample correction (RISC) algorithm, was introduced in [22], where synthetic sample paths are iteratively generated and corrected to match the observed bin counts. Reliance on sampling makes these approaches more computationally expensive than the others.

Several modifications have been proposed to estimate the Hawkes process from daily count data in the context of modeling the spread of COVID [23]. A Hawkes process incorporating spatiotemporal covariates was estimated using an EM algorithm in [24], while the least-squares approach was utilized in [25] to model state-level differences in transmission rates in the U.S. A discrete-time Hawkes process for country-level COVID transmission was introduced in [4] and fit using Bayesian inference.

The notion of interval-censored data is also used in other fields, most prominently in the study of time-to-event (failure time) data [26], [27], [28]. In these works, “interval-censored” refers to situations where the precise time of an event of interest in an observational study is unknown; instead, we only know that it occurred within a certain window or follow-up period [29]. Meanwhile “partly interval-censored” refers to situations where some failure times are exactly observed while others are only known to lie within certain intervals [30]. This data type is prevalent in health and clinical research [31], where exact event times may not be directly observable due to the nature of study designs. Contrary to this, our work adopts the definition of “interval-censored” as outlined in [7] and [32], where event times are inaccessible and we instead observe event counts over predefined time intervals, as exemplified by YouTube views in Fig. 1. Furthermore, in this work, we define partial censoring of a multivariate process as the censoring of specific dimensions, with the rest being observed as point processes.

II. MODEL

For convenience, the list of notation that we use in this work is provided in Table I.

A. Background

A temporal point process can be specified by its conditional intensity function. In this work, we consider simple point processes, where no two events can occur simultaneously. Let \( N(t) \) represents the number of events that occurred up until time \( t \) and \( \mathcal{H}_t \) be the set of all events that occur in dimension \( j \) up until \( t \), for \( j \in \{1,\ldots,d\} \). We further denote the union of all history dimensions as \( \mathcal{H}_t := \bigcup_{j=1}^{d} \mathcal{H}_t^{(j)} \). The \( d \)-dimensional conditional intensity function \( \lambda(t | \mathcal{H}_t) \) is defined as
\[
\lambda(t | \mathcal{H}_t) = \lim_{h \to 0^+} \frac{1}{h} \mathbb{P} \left\{ N^j(t+h) - N^j(t) = 1 | \mathcal{H}_t \right\}
\]
which gives the instantaneous probability of a dimension \( j \) event occurring in the increment \( [t, t+dt) \), conditioned on all events that happened before \( t \). For brevity and whenever it is clear from context, we drop the explicit conditioning on the history \( \mathcal{H}_t \) and write \( \lambda(t) := \lambda(t | \mathcal{H}_t) \).

1) Hawkes Process: The univariate Hawkes process is a type of temporal point process that models a sequence of events on a single dimension exhibiting a self-exciting behavior. Given \( d \) types of events, the corresponding \( d \)-dimensional MHP is a point process where each dimension tracks the dynamics of

**TABLE I IMPORTANT NOTATION USED IN THIS ARTICLE**

| Symbol | Meaning |
|--------|---------|
| \( \alpha \) | Hawkes branching matrix |
| \( \rho(\alpha) \) | Spectral radius of \( \alpha \) |
| \( f_j^{(i)}(t) \) | Exponential kernel from dimension \( j \) to \( i \) |
| \( \theta \) | Exponential kernel decay parameter matrix |
| \( \mu(t) \) | Deterministic background intensity |
| \( \phi(t) \) | Hawkes kernel, where \( \phi^{(i)}(t) = \alpha^{(i)} f_j^{(i)}(t) \) |
| \( \lambda(t) \) | MHP conditional intensity |
| \( \xi(t) \) | MBP conditional intensity |
| \( D \) | Overall set of dimensions for PCMHP \((d,e)\) |
| \( E \) | Set of MBP dimensions for PCMHP \((d,e)\) |
| \( E^i \) | Set of Hawkes dimensions for PCMHP \((d,e)\) |
| \( \mathcal{H}_t^{(j)} \) | Event sequence history on dimension \( j \in D \) |
| \( \mathcal{H}_t \) | Union of event sequence histories on \( A \subset D \) |
| \( \mathcal{L}(\Theta, T) \) | Log-likelihood function for PCMHP \((d,e)\) |
| \( \mathcal{L}_\Theta(\Theta, T) \) | Gradient of log-likelihood function for PCMHP \((d,e)\) |
| \( \Delta^h \) | Time axis partition length for numerical convolution |
| \( \gamma^h \) | Convergence threshold for infinite sum truncation |
each event type. In addition to being self-excitation, the MHP is cross exciting among event types, i.e., an event occurring in one type of event increases the probability of any type of event occurring in the near future. The conditional intensity of the \(d\)-dimensional Hawkes process is given by

\[
\lambda(t) := \mathbf{\mu}(t) + \sum_{j=1}^{d} \sum_{i,t_k' \in \mathcal{H}_{t-}} \varphi^j(t - t_k')
\]

where \(\mathbf{\mu}(t)\) is the (deterministic) background intensity, a \(d\)-dimensional nonnegative vector for each \(t\) controlling the arrival of external events into the system. The matrix \(\varphi(t)\) is called the Hawkes kernel, a \(d \times d\) matrix of functions that characterize the self- and cross excitation across the event types representing the \(d\) dimensions. Let \(\varphi^j(t)\) represents the \(j\)th column of the Hawkes kernel. The diagonal entries \(\varphi^{jj}(t)\) and off-diagonal entries \(\varphi^{ij}(t), i \neq j\), represent the self- and cross exciting components of the Hawkes kernel, respectively. Note that the Hawkes intensity \(\lambda(t)\) defined in (1) is a stochastic function dependent on the history of any particular realization \(\mathcal{H}_{t-}\). Given a fixed \(\mathcal{H}_{t-}\), the intensity before or equal to \(t\) is deterministically calculated using (1). On the other hand, the intensity is a random variable for any time greater than \(t\) or if the history \(\mathcal{H}_{t-}\) itself is not observable, such as in the case of interval censoring.

The Hawkes kernel is often specified in a parametric form to facilitate simple interpretability. Let \(D\) denotes the index set \(\{1, \ldots, d\}\). If we assume \(\varphi^{ij}(t) = \alpha^{ij}f^{ij}(t)\), \(\alpha^{ij} \geq 0\), \(f^{ij}(t) \geq 0\), and \(\int_0^\infty f^{ij}(t)dt = 1\) for \((i,j) \in D \times D\). We call \(\alpha^{ij}\) the branching factor from \(j\) to \(i\) and the matrix \(\alpha = (\alpha^{ij}) \in (\mathbb{R}^+)^{d \times d}\) the branching matrix. The branching factor \(\alpha^{ij}\) gives the expected number of offspring events in dimension \(i\) that are triggered by an event in dimension \(j\). The function \(f^{ij}(t)\) is typically selected to be monotonically decreasing to model the empirically observed decay in the attention that online content receives over time [13]. This is explained by viewing human attention as a limited resource that online content competes for, resulting in content being forgotten over time. In this work, we consider the widely used exponential kernel [7], [11], [33], which takes the form \(\varphi^{ij}(t) = \alpha^{ij}\theta^{ij} \exp(-\theta^{ij}t)\), where \(\theta^{ij}\) controls the rate of influence decay from \(j\) to \(i\). We choose the exponential kernel due to its simplicity, enabling efficient computation of the intensity due to its Markovian property [34], efficient sampling [35], and strong theoretical properties, as evidenced by its link to the stochastic SIR model [36]. An alternative choice of kernel is the power law, which, while yielding better model likelihood, requires significantly more time to fit and presents a more challenging estimation due to the need to fit two parameters. More prerequisite details on the MHP are provided in Sec. 1.1 of the supplementary material [37].

2) MBP Process [7]: Consider a univariate Hawkes process with conditional intensity \(\lambda(t)\). The MBP process introduced by [7] is the inhomogeneous Poisson process with conditional intensity

\[
\xi(t) := \mathbb{E}_{\mathcal{H}_{t-}} \left[ \lambda(t) \right].
\]

In contrast to the stochastic Hawkes intensity \(\lambda(t|\mathcal{H}_{t-})\), the MBP intensity \(\xi(t)\) is a deterministic function obtained by taking the expectation of the Hawkes intensity over all possible realizations \(\{\mathcal{H}_{t-}\}\). It was shown in [7] that \(\xi(t)\) follows the self-consistent equation

\[
\xi(t) = \mu(t) + (\varphi \ast \xi(t))
\]

where \(\ast\) denotes convolution. Furthermore, the mapping \(\mu(t) \mapsto \xi(t)\) in (3) defines a linear time-invariant (LTI) system [38], meaning that it obeys linearity \((\mu_1(t) \mapsto \xi_1(t)\) and \(\mu_2(t) \mapsto \xi_2(t)\) imply that \(a\mu_1(t) + b\mu_2(t) \mapsto a\xi_1(t) + b\xi_2(t)\) for \(a, b \in \mathbb{R}\)) and time invariance \((\mu(t) \mapsto \xi(t))\) implies that \(\mu(t - t_0) \mapsto \xi(t - t_0)\) for \(t_0 > 0\). As an LTI system, the response \(\xi(t)\) to the input \(\mu(t)\) can be obtained by solving for the response of the system to the Dirac impulse \(\delta(t)\), derived in [7] to be

\[
\xi(t) = \left( \delta(t) + \sum_{n=1}^{\infty} \varphi^{\otimes n}(t) \right) \ast \mu(t)
\]

where \(\otimes\) corresponds to \(n\)-time self-convolution.

Since the MBP process is a Poisson process, its increments are independent, which allows the likelihood function to be expressed as a sum of the likelihood of disjoint Poisson distributions. This enables the MBP process to be fitted in interval-censored settings via MLE. More prerequisite details are provided in Sec. 1.2 of the supplementary material [37].

3) HIP [1]: The HIP, introduced in [1], is a temporal point process that can be fit to interval-censored data. It was used primarily for YouTube popularity prediction, where YouTube video views are daily-censored, and external shares and tweets that mention the video act as the exogenous intensity \(\mu(t)\).

Given a partition \(\mathcal{P} [0, T) = \bigsqcup_{k=1}^{m} \{0, t_{k-1}, t_k\},\) where \(t_0 = 0\) and \(t_m = T\), and the associated view counts \(\{C_k\}_{k=1}^{m}\), the HIP model \(\xi [\cdot; \Theta]\) is fitted by finding the parameter set \(\Theta\) that minimizes the sum of squares error \(\sum_{k=1}^{m} (C_k - \xi [0; \Theta])^2\) of the following recursive formula for \(\xi [0; \Theta]\):

\[
\xi [0; \Theta] = \mu [0] + \sum_{k=1}^{m} \sum_{s=0}^{k-1} \varphi [0, \theta; \Theta] \cdot \xi [0, \theta; \Theta].
\]

The use of brackets emphasizes that the quantities are discretized over a partition of time. It was shown in [7, Theorem 10] that HIP is a discretized approximation of the MBP process, where an implicit assumption that the observation intervals being unit length is reflected in the sum of squares error, i.e., \(\xi [0, \theta; \Theta] \cdot (0, \theta)\) is approximated as \(\xi [0; \Theta]\).

B. PCMHHP

We define the PCMHHP \((d, e)\) with intensity \(\xi_E [t]\) as follows. A key idea of the PCMHHP \((d, e)\) is to fix the history of different dimensions. As such we denote the history union over a subset of dimensions \(A \subset \{1 \ldots d\}\) as \(\mathcal{H}_A^{-} := \bigsqcup_{k \in A} \mathcal{H}_k^{-}\).

Definition 1: Consider a \(d\)-dimensional Hawkes process with conditional intensity \(\lambda(t)\) as defined in (1). Given a nonnegative integer \(e \leq d\) and the index sets \(D := \{1, \ldots, d\}, E := \{1, \ldots, e\}\) and \(E^c := \{e + 1, \ldots, d\}\), the PCMHHP \((d, e)\) is the
temporal point process whose conditional intensity \( \xi_E(t) \) is the expectation of \( \lambda(t) \) conditioned on the set of event histories \( \mathcal{H}_E \) in the \( E \) dimensions and averaged over the set of event histories \( \mathcal{H}_E \) in the \( E \) dimensions. That is,
\[
\xi_E(t) := \mathbb{E}_{\mathcal{H}_E} \left[ \lambda(t) \mid \mathcal{H}_E \right].
\] 

The PCMHPP \((d, e)\) intensity is a stochastic function due to its dependence on the current realization of \( \mathcal{H}_E \). On the dimensions we take the expectation over all possible realizations of \( \mathcal{H}_E \), similar to the MBP intensity in \((2)\).

In practice, \( E \) would be chosen to be the set of dimensions where event times are inaccessible and only interval-censored event counts can be obtained, while \( E^c \) would be the dimensions with event time information. In Fig. 1, for instance \( E \) would be YouTube views and \( E^c \) the set of tweets.

1) Is the PCMHPP \((d, e)\) Poisson? Due to its dependence on the history of the \( E \) dimensions, the PCMHPP \((d, e)\) is not a Poisson process. From \((5)\), the PCMHPP \((d, e)\) can be interpreted as a collection of \( d \) processes, where \( \xi_E(t) \) follows a Hawkes process for \( j \in E \) and follows an inhomogeneous Poisson process (conditional on the event history of the \( E^c \) dimensions) for \( j \in E \). In fact, the PCMHPP \((d, e)\) generalizes both the MHP (by setting \( e = 0 \)) and the MBP process (by setting \( e = d \)).

2) Convolutional Formula: Consider the kernel \( \varphi(t) = \left[ \varphi^1(t) \ldots \varphi^e(t) \varphi^{e+1}(t) \ldots \varphi^d(t) \right] \), setting \( \varphi^i(t) \) to be the \( i \)th column of \( \varphi(t) \).

Similarly, let \( \varphi_E(t) = \left[ \varphi^1(t) \ldots \varphi^e(t) 0 \ldots 0 \right] \) and \( \varphi_{E^c}(t) = \left[ 0 \ldots 0 \varphi^{e+1}(t) \ldots \varphi^d(t) \right] \).

Similar to MBP, \( \xi_E(t) \) can be expressed as the response of an LTI system, which allows us to express \( \xi_E(t) \) as a convolution with the Dirac impulse \( \delta(t) \).

**Theorem 2:** Given the Hawkes process with intensity \((1)\) \(E\) and kernel parameters satisfying \( \lim_{n \to \infty} \varphi^n_E(t) = 0 \), the conditional intensity of the PCMHPP \((d, e)\) is as follows:
\[
\xi_E(t) = \left[ \delta(t) + \sum_{n=1}^{\infty} \varphi^n_E(t) \right] \ast \left[ \mu(t) + \sum_{j \in E} \sum_{t_j \leq t} \varphi^j(t - t_j) \right].
\] 

In general, \( \xi_E(t) \) does not admit a closed form solution because of the complexity of the infinite convolution sum of \( \varphi_E(t) \) (an interpretation of which is provided in Sec. 2 of the supplementary material [37]). However, in the special case of PCMHPP \((2, 1)\) with the exponential kernel, a closed-form solution for \( \xi_E(t) \) exists, as proven in Sec. 3 of the supplementary material.

3) Regularity Conditions: Imposing regularity conditions on the model parameters ensure process subcriticality, i.e., the expected number of direct and indirect offspring spawned by a single parent is finite. For instance, an MHP is subcritical if the spectral radius \( \rho \) (i.e., magnitude of the largest eigenvalue) of the branching matrix is less than one, i.e., \( \rho(\alpha) < 1 \) [39]. Here, we introduce the regularity conditions applicable for the PCMHPP \((d, e)\).

Consider the following submatrices of \( \alpha \):
\[
\alpha^{EE} = (\alpha^{ij})_{(i,j) \in E \times E}, \quad \alpha^{E^cE} = (\alpha^{ij})_{(i,j) \in E \times E^c}.
\]

The following are three conditions that ensure subcriticality of PCMHPP \((d, e)\).

**Theorem 3:** The PCMHPP \((d, e)\) with branching matrix \( \alpha \) is subcritical if the following conditions hold:
\[
\rho(\alpha^{EE}) < 1, \quad \rho(\alpha^{EE}) < 1, \quad \rho(\alpha^{EE} (I - \alpha^{EE})^{-1} \alpha^{EE}) < 1.
\]

We note that the regularity conditions for PCMHPP \((d, e)\) in Theorem 3 cover the MHP and the MBP as special cases.

Proofs of Theorem 2 and Theorem 3 and a discussion on the nonlinear extension of the PCMHPP are provided in Sec. 4 of the supplementary material [37].

**C. Inference**

We now consider the problem of estimating the PCMHPP parameter set \( \Theta \) given a partially interval-censored dataset consisting of interval-censored data on a subset of dimensions and exact event sequences on the other dimensions. Assuming a constant exogenous term \( \mu(t) = \nu \), the PCMHPP \((d, e)\) parameter set to be estimated is given by \( \Theta = \{ \nu, \alpha, \theta \} \) with size \( |\Theta| = d + 2 \cdot d^2 \).

Consider a \( d \)-dimensional dataset over the time interval \([0, T]\) such that observations in the first \( q \) dimensions are interval-censored, and in the last \( d - q \) dimensions, we observe event times. Formally, let \( Q := \{ 1, \ldots, q \} \) and \( Q^c := \{ q + 1, \ldots, d \} \). For \( j \in Q \), we associate a set of observation points \( o_j^{(1)} < o_j^{(2)} < \ldots < o_j^{(n_j)} \) such that for \( o_j^{(k)} \) where \( k \geq 1 \), we observe the volume \( C_{o_j^{(k)}}^j \) of dimension \( j \) events that occurred during the interval \([o_j^{(k-1)}, o_j^{(k)}]\). Meanwhile, for \( j \in Q^c \), we observe event sequences \( H_{o_j}^j \). In practice, the observation partition \( \bigcup_{j \in Q} \bigcup_{o_j^{(k)}}^{o_j^{(n_j)}} \) is not a model hyperparameter but is determined by real-world dataset availability constraints. For instance, in Fig. 1 we consider a daily partitioning for YouTube views since our dataset consists of aggregated daily view counts.

We use MLE to fit the parameters of a PCMHPP \((d, e)\) process to the above-defined data using the log-likelihood function derived as follows. The proof is available in Sec. 5 of the supplementary material [37].

**Theorem 4:** Given event times \( \mathcal{H}_E \), event volumes \( \mathcal{V}_E \), and a PCMHPP \((d, e)\) model such that \( E \subseteq Q \), the negative log-likelihood of parameter set \( \Theta \) can be written as
\[
\mathcal{L}(\Theta; T) = \sum_{j \in Q^c} \mathcal{L}^{j}_{\text{IC-LL}}(\Theta; T) + \sum_{j \in Q^c} \mathcal{L}^{j}_{\text{PP-LL}}(\Theta; T)
\] 

where
\[
\mathcal{L}^{j}_{\text{IC-LL}}(\Theta; T) = -\sum_{i=1}^{n_j} \log \mathcal{C}_{o_j^{(i+1)}}^j(\xi_j^{(i)}(t_j^i; \Theta)) + \mathcal{V}_E(\xi_j^{(i)}; \Theta),
\]
\[
\mathcal{L}^{j}_{\text{PP-LL}}(\Theta; T) = -\sum_{t_j^i \in H_{o_j}^j} \log \mathcal{C}_{o_j^{(i)}}^j(t_j^i; \Theta) + \mathcal{V}_E(\xi_j^{(i)}; \Theta),
\]
\[\mathcal{L}^{j}_{\text{IC-LL}}(\Theta; T) = -\sum_{i=1}^{n_j} \log \mathcal{C}_{o_j^{(i+1)}}^j(\xi_j^{(i)}(T; \Theta)) + \mathcal{V}_E(\xi_j^{(i)}; \Theta).
\]
and $\Xi_E(t)$ represents the compensator, i.e., the intensity $\xi_E(t)$ integrated over $0$ to $t$.

1) **Choice of Likelihood**: The choice of likelihood on a given dimension $j$ is solely dependent on the type of data on the said dimension. If $j \in Q$ (dimension $j$ is interval-censored), one should use $L_{\text{IC-LL}}(\Theta; T)$; if $j \in Q^c$ (event-times) then $L_{\text{PP-LL}}(\Theta; T)$ should be used.

An event-time dimension ($j \in Q^c$) can be modeled using either the Hawkes dynamics or the MBP dynamics. However, an interval-censored dimension ($j \in Q$) can only be modeled using MBP dynamics, as an interval-censored log-likelihood for the Hawkes dynamics does not exist. It follows that $E \supseteq Q$. In real-world applications, one would choose $E = Q$ because any other choice $E \supset Q$ leads to information loss due to the mismatch between the data generation model (i.e., Hawkes) and the fitting model (MBP). We study the impact of model mismatch loss in Section III.

2) **Runtime Complexity**: Denote $n^{E^c}$ and $n^{Q^c}$ as the total number of observed event times in the $E^c$ and $Q^c$ dimensions, respectively; and $n^E$ and $n^Q$ as the total number of observation intervals in the $E$ and $Q$ dimensions, respectively. That is, $n^E = \sum_{j \in E} |H^E_j|$, $n^Q = \sum_{j \in Q} |H^Q_j|$, $n^E = \sum_{j \in E} n^j$, and $n^Q = \sum_{j \in Q} n^j$. Let $C$ denote a constant independent of the dimension of the PCMHP and the data. Evaluating $L(\Theta; T)$ has a runtime complexity of $O \left((C + n^{E^c}) \cdot (n^Q + n^{Q^c})\right)$ (see Sec. 5 of the supplementary material [37] for more details). In the case $E = Q = \emptyset$, the runtime complexity reduces to $O((n^{E^c})^2)$, consistent with the MHP. If $E = Q = D$, runtime complexity reduces to $O(n^2)$, consistent with the MBP (i.e., Poisson) process.

3) **Numerical Considerations**: Due to the complexity of $\sum_{n=1}^\infty \varphi_{E^c}^n(t)$ and its convolutions, a general closed-form expression for $\xi_E(t)$ is not available, requiring us to leverage approximation techniques, i.e., numerical convolution and infinite series truncation, to compute $\sum_{n=1}^\infty \varphi_{E^c}^n(t)$ and $\xi_E(t)$. The approximation error is controlled by two hyperparameters: (1) $\Delta^n$, the partition length of our time axis for the numerical convolution, and (2) $\rho^n$, the max-norm convergence threshold to determine $k^n \in \mathbb{N}$ to truncate the infinite sum, i.e., $\sum_{n=1}^\infty \varphi_{E^c}^n(t) \approx \sum_{n=1}^{k^n} \varphi_{E^c}^n(t)$. The smaller $\Delta^n$ and $\rho^n$ are set, the tighter the approximation, albeit with a longer computation time. Full details and heuristics on hyperparameter choice are discussed in Sec. 7 of the supplementary material [37]. We propose an alternative sampling-based technique to calculate $\xi_E(t)$ that bypasses the calculation of $\sum_{n=1}^\infty \varphi_{E^c}^n(t)$ in Sec. 8 of the supplementary material [37]. Finally, we demonstrate the convergence of the numerical and the sampling-based approximation techniques in Sec. 9 of the supplementary material [37] by showing close agreement between the approximated $\xi_E(t)$ and the closed-form $\xi_E(t)$ of the exponential PCMHP (2.1) derived in Sec. 3 of the supplementary material [37].

Gradient-based optimization tools—including IPOPT [40] that we use in our experiments in Sections IV and V—usually require the gradient. To approximate $L(\Theta; T)$ and its gradient $\nabla_{\Theta} L(\Theta; T)$, we propose a numerical scheme in Secs. 10 and 11 of the supplementary material [37]. We show in the supplementary material that the runtime complexity of the numerical scheme is mostly determined by how many dimensions we model as Hawkes and as MBP. Without any Hawkes dimensions ($E^c = \emptyset$), the scheme scales linearly (similar to the MBP) with the number of observation intervals $n^E$, i.e., $O \left((n^E \cdot |(T/\Delta^c)|) \cdot d \cdot e\right)$. On the other hand, if $E^c \neq \emptyset$, the scheme scales quadratically (similar to the MHP) with the number of observed event times $n^E$, i.e., $O \left((n^E + n^{E^c} + |(T/\Delta^c)|) \cdot n^E \cdot d\right)$. In both cases, the number of partition intervals $|(T/\Delta^c)|$ only appears linearly. For partially interval-censored datasets with high-frequency data, the number of observed event times $n^E$ is the most important determinant of runtime complexity given that it appears quadratically, while the number of observation intervals $n^{E^c}$ and the number of partition intervals $|(T/\Delta^c)|$ only appear linearly.

Last, for the purpose of sampling from the PCMHP $(d, e)$, we propose a modification of the thinning algorithm [34] detailed in Sec. 12 of the supplementary material.

**D. Heuristics for Partially Interval-Censored Data**

The PCMHP is designed for cases where 1) the dataset is multivariate and partially interval-censored; and 2) we hypothesize events are self-exciting and cross exciting across dimensions. To handle partially interval-censored data, our strategy is to adapt the model (i.e., the PCMHP) to the data. However, we can take the reverse approach and apply heuristics to our dataset to be able to leverage preexisting models.

1) **To use count-based time series models**, we transform our partially interval-censored dataset into a fully interval-censored dataset by censoring event times for each $E^c$ dimension.

2) **To use point process models**, we transform our partially interval-censored dataset into a fully time-stamped dataset by sampling event times to match the interval-censored counts, for each dimension in $E$.

There are three arguments against the first heuristic. First, artificially censoring the dataset leads to loss of timing information by hiding self- and cross-exciting interactions between events, particularly if the time scale of the interactions is less than the censor window length. Second, commonly used time series models (such as the Poisson autoregressive model [41] or the discrete-time Hawkes process [44]) assume evenly spaced data [42]. If the censor intervals within or across dimensions do not line up, we would need to perform further data alteration, such as interpolation [43], to attain evenly spaced data. The PCMHP does not require evenly spaced intervals. Third, using time series models on the artificially obtained interval-censored dataset requires additional model choices. For instance, we would have to set the censor window length for each dimension in $E$ when transforming to a fully interval-censored dataset, and for autoregressive models decide up to what lag $p$ to include. The PCMHP requires no adaptation as it was designed for partially interval-censored data.

The main deterrent against the second heuristic, artificial event sampling, is the significant addition to computation time, since evaluating the Hawkes likelihood is $O((n^E)^2)$. This is
particularly infeasible in applications involving high-event volumes, such as YouTube views on a viral video, which typically have view counts of the order $10^6$ or more. If we use the PCMHP, these dimensions with high-event volumes can be modeled as event counts and placed in $E$ instead of $E^c$, significantly reducing computation time. Second, artificially sampling points—when only aggregated counts have been given—have the potential to produce spurious event interactions across dimensions, particularly for wide censor intervals.

III. SYNTHETIC PARAMETER RECOVERY

In this section, we test on synthetic data the MHP parameter recovery by PCMHP$(d,e)$. We use the setting of partial interval-censoring with a constant exogenous term $\mu(t) = \nu$. We sample realizations from a $d$-dimensional MHP, interval-censor $e$ dimensions using increasingly wide observation window lengths, and fit the PCMHP$(d,e)$ model on the obtained partially interval-censored data. We inspect the recovery of parameters when varying $d$ and $e$. We perform convergence analysis on the PCMHP$(d,e)$ parameter estimates for various hyperparameter configurations in Sec. 14.2 of the supplementary material [37].

Throughout this section, we refer to $\{\alpha, \theta, \nu\}$ and $\{\hat{\alpha}, \hat{\theta}, \hat{\nu}\}$ as the true (MHP) and estimated parameter sets, respectively. We first discuss the two types of information loss, then we introduce the synthetic datasets and the likelihood functions. Last, we present the recovery results for the individual parameters $\{\alpha, \theta, \nu\}$.

A. Sources of Information Loss in Fitting

We identify two sources of information loss when fitting in the partially interval-censored setup: (1) the mismatch between the data-generation model (i.e., the $d$-dimensional MHP) and the fitting model [$\text{the PCMHP}(d,e)$]; and (2) the interval-censoring of the timestamped MHP data. Since the intensity $\frac{d}{E}(t)$ of the PCMHP $(d,e)$ has to be estimated numerically (see Section II-C), numerical approximation error also contributes to type (1) information loss. The numerical approximation error is minimal for sufficiently small $\Delta^p$ and $\gamma^h$ (see Sec. 9 of the supplementary material [37]) and vanishes if $\Delta^p, \gamma^h \rightarrow 0$.

When we estimate MHP parameters using PCMHP fit on partially interval-censored data, information losses of both types (1) and (2) occur. We disentangle between the two types of error by also fitting PCMHP$(d,e)$ on the timestamp dataset (i.e., the actual realizations sampled from the MHP, see as follows). Any information loss in this setup is only due to the model mismatch, the information loss of type (1). Note that the likelihood function used for fitting PCMHP depends on the employed version of the dataset (see later in this section).

We can quantify the individual effects of model mismatch and interval-censoring by comparing the parameter estimates on the two dataset versions.

B. Synthetic Dataset

Given $(d,e)$, we construct two synthetic datasets: the timestamp dataset and the partially interval-censored dataset. The timestamp dataset consists of samples from a $d$-dimensional MHP. The partially interval-censored dataset is obtained from the timestamp dataset by interval-censoring the $E$ dimensions while leaving the $E^c$ dimensions unchanged. In this experiment, we focus on the case $d=2$ and $e=1$.

We consider a 2-D MHP with $\rho(\alpha) = 0.75$ and parameters $\alpha^{11} = 0.32, \alpha^{12} = 0.5, \alpha^{21} = 0.3, \alpha^{22} = 0.4, \theta^1 = 0.5, \theta^2 = 1.0, \delta^1 = 0.5, \delta^2 = 1.25,$ and $\nu^1 = \nu^2 = 0.1$ We set $\rho(\alpha) \in \{0.5, 0.75, 0.9\}$. We also test another parameter combination with $\rho(\alpha) = 0.5$ (i.e., subcritical) and $\rho(\alpha) = 0.9$ (i.e., approaching the critical regime) in Sec. 14.1 of the supplementary material [37].

For a given MHP parameter set, we sample 2500 event sequences $H_{100}^{1100} \cup H_{100}^{2100}$ over the time interval $[0, 100)$ using the MHP thinning algorithm [34]. Following a procedure similar to the prior literature [7], we partition the 2500 event sequences into 50 groups of 50 sequences.

We construct the partially interval-censored dataset by interval-censoring $H_{100}^{100}$, the first dimension of each realization in the timestamp dataset. Given a partition of $[0, 100)$, we count the number of events on dimension 1 that fall on each subinterval. We experiment with five observation window lengths to quantify the information loss of type (2)—intuitively, longer intervals lead to more significant information loss. We consider interval lengths of 1, 2, 5, 10, and 20. For instance, with interval length of 2 we tally event counts in the partition $\{[0, 2), [2, 4), \ldots, [98, 100]\}$.

C. PCMHP Log-Likelihood Functions

We fit the parameters of the PCMHP$(d,e)$ model using two different versions of the likelihood function dependent on which dataset we use.

1) Timestamp Dataset: we use the point-process log-likelihood on all dimensions, defined in (9): $\sum_{j=1}^{d} L_{PP-LL}(\Theta, T)$.

2) Partially Interval-Censored Dataset: we use the interval-censored log-likelihood on the $E$ dimensions and the point-process log-likelihood on the $E^c$ dimensions [see (7)]: $\sum_{j=1}^{d} L_{IC-LL}(\Theta, T) + \sum_{c=e+1}^{d} L_{PP-LL}(\Theta, T)$.

In what follows, we specify as PCMHP$(d,e)$-PP and PCMHP$(d,e)$-IC the PCMHP$(d,e)$ model fit on the timestamp dataset and the partially interval-censored dataset, respectively. For brevity and whenever it is clear from context, we drop the dimensionalities $(d,e)$, and refer to the model fits as PCMHP-PP and PCMHP-IC. Also, for the PCMHP-IC fits, we specify $k$—the length of the observation window—as PCMHP-IC[$k]$.

1) Inference: We infer a single $\Theta$ for each group of sequences using a joint fitting procedure. Since there are 50 groups in total, this process yields 50 estimates of $\Theta$. For each group of 50 sequences, we obtain $\Theta$ that minimizes the joint negative log-likelihood, which is defined as $L(\Theta, T) = \sum_{i=1}^{50} L_i(\Theta, T)$, where $L_i(\Theta, T)$ represents the negative log-likelihood for the $i^{th}$ event sequence.

2) Metrics: We evaluate parameter recovery error with four error metrics: the root-mean-squared error (RMSE) of each PCMHP parameter type $\{\alpha, \theta, \nu\}$ concerning the generating
MHP parameters $\{\alpha, \theta, \nu\}$ and the signed deviation $\Delta \rho = \rho(\hat{\alpha}) - \rho(\alpha)$ of the spectral radius.

D. Parameter Recovery Results

Fig. 2 shows RMSE ($\alpha$), RMSE ($\theta$), RMSE ($\nu$), and $\Delta \rho$ across model fits. Each boxplot represents the distribution of errors calculated from 50 estimates of $\Theta$, each obtained from a different group of 50 sequences in the synthetic dataset. Within each subplot, we have seven boxplots. The leftmost boxplot is the MHP fit, followed by the PCMHP-PP fit (i.e., the PCMHP fit on the timestamp dataset). The next five boxplots contain PCMHP-IC fits of increasingly wider observation windows 1, 2, 5, 10, and 20. Note that the MHP fit represents the case where we do not have either model mismatch and interval censoring error.

In each subplot of Fig. 2, the gap between the first two boxplots [i.e., MHP versus PCMHP(2, 1) fitted on timestamp data] indicates model mismatch error; the gap between the second and third boxplots [i.e., PCMHP(2, 1) fitted on timestamp data versus partially interval-censored data] indicates interval censoring error. The gaps between succeeding boxplots indicate the effect of wider observation windows.

Fig. 2 shows three conclusions. First, model mismatch and interval censoring errors contribute to information loss relative to the MHP fit. Second, the approximation quality degrades as the observation window widens, indicating an increasing information loss of type (2). Third, for parameters $\alpha$ and $\nu$, the model mismatch error appears negligible; it is only for higher values of the observation window length ($\geq 5$) that the performance starts degrading due to information loss error. Both error types are present for $\theta$. See Sec. 14.1 of the supplementary material for individual parameter fits.

Though we observe that the generating parameters are not always correctly recovered, we see in the rightmost subplot of Fig. 2 that, interestingly, the spectral radius estimation error $\Delta \rho$ is close to zero regardless of model mismatch and exhibits only slight underestimation for wide observation windows. This is particularly relevant, as $\rho(\alpha)$ is a meaningful quantity relating to information spread virality (for social media diffusions), disease infectiousness (for epidemiology), or local seismicity (in seismology). The result indicates that even when individual parameter fits are inaccurate, the MHP regime is correctly identified.

E. Behavior of $\Delta \rho$ in Higher Dimensions

We further study the behavior of $\Delta \rho$ for varying MBP dimensions $e$ and model dimensionality $d$. We fix $T = 100$ and $\rho(\alpha) = 0.92$. Results for other error metrics are in Sec. 14.2 of the supplementary material.

In the left subplot of Fig. 3, we fix $d = 5$ and observe how the spectral radius error $\Delta \rho$ varies with the number of MBP dimensions $e$. Note that the leftmost boxplot represents the MHP fit (i.e., $e = 0$). Interestingly, we see that all PCMHP($5, e$), $e < 5$ flavors except the fully MBP case (i.e., $e = d = 5$) can estimate the spectral radius as well as the MHP. The gap between the estimated spectral radii and the generating value (blue dashed line) is attributable to the difficulty of recovering MHP parameters in higher dimensions.

In the right subplot of Fig. 3, we fix $e = 1$ and observe how the spectral radius error $\Delta \rho$ varies with the dimensionality $d$ of the PCMHP($d, 1$). The recovery error is generally low (except for $d = 1$). However, we see that the magnitude of the
error $\Delta \rho$ increases with increasing dimensionality starting from $d = 2$, which is not surprising since the number of parameters increases quadratically as we increase the dimensionality of the process. We also see that fitting with a fully MBP model ($d = 1$) does not show good recovery performance due to information loss, implying the necessity of having at least one cross-exciting dimension (i.e., $d - e \neq 0$).

IV. YOUTUBE POPULARITY PREDICTION

In this section, we evaluate PCMHP($d, e$)'s performance in predicting the popularity of YouTube videos. For each video, we capture information about three dimensions—views, external shares, and tweets—linking to the videos—over the time period $[0, T_{\text{train}}]$. We measure time in days relative to the time of posting on YouTube. The first two dimensions (the views and shares) are observed as daily counts, i.e., $E = \{\text{views, shares}\}$. The third dimension (tweets) is provided as event times, i.e., $E' = \{\text{tweets}\}$. Given this data setup, we use PCMHP($3, 2$) to predict the daily counts of views and shares and the timestamp of the tweets posted over the period $[T_{\text{train}}, T_{\text{test}}]$.

A. Interval-Censored Forecasting With PCMHP

Each YouTube video corresponds to a partially interval-censored Hawkes realization. A straightforward approach to predict the unfolding of the realization during $[T_{\text{train}}, T_{\text{test}}]$ is to sample timestamps from PCMHP($3, 2$) on each of the three dimensions, conditioned on data before $T_{\text{train}}$; we then interval-censor the first two dimensions. In practice, sampling individual views takes considerable computational effort due to their high-background rates, sometimes in the order of millions of views per day and usually at least an order of magnitude larger than shares and tweets.

An efficient procedure to calculate expected counts that leverages the compensator $\Xi_E$ and requires sampling only the $E^c$ dimensions (i.e., tweets), as follows. Let $\mathcal{P}(T_{\text{train}}, T_{\text{test}}) = \bigcup_{i=1}^{P-1} [o_i, o_{i+1})$, where $o_1 = T_{\text{train}}$ and $o_P = T_{\text{test}}$, be a partition of $[T_{\text{train}}, T_{\text{test}}]$.

1) Sample only the $E^c$ dimensions on $[T_{\text{train}}, T_{\text{test}}]$.
2) Compute expected counts on $\mathcal{P}(T_{\text{train}}, T_{\text{test}})$ as $\mathbb{E}_E(o_{i+1}) - \mathbb{E}_E(o_i) | i \in 1, \ldots, P - 1 \}.
3) Compute the average of $\mathbb{E}_E(o_{i+1}) - \mathbb{E}_E(o_i) | i \in 1, \ldots, P - 1 \} across samples.

More details of this scheme and a comparison with the standard method of sampling both $E^c$ and $E$ dimensions are provided in Sec. 13 of the supplementary material [37].

B. Dataset, Experimental Setup, and Evaluation

We use two subsets of the ACTIVE dataset [1] for model fitting and evaluation. The first subset—dubbed ACTIVE 20%—contains a 20% random sample of the ACTIVE dataset [1], i.e., 2834 videos published between 29 May 2014 and 26 December 2014. The second subset—dubbed DYNAMIC VIDEOS—contains videos with which users engage significantly during the test period. It is known that users’ attention to YouTube videos decays with time [13, 44]; therefore, the daily views of most ACTIVE videos hover around zero more than 90 days after their upload. We select the 585 dynamic videos with the standard deviations of the views, tweets, and shares counts on days 21–90 higher than the median values on each of the three measures. Technical details of the filtering are in Sec. 15.1 of the supplementary material [37].

For each video, we tune PCMHP hyperparameters and parameters using the first 90 days of daily view counts, share count to external platforms, and the timestamps of tweets that mention each video ($T_{\text{train}} = 90$). It is known that generative models are suboptimal for prediction [11] and have to be adapted to the prediction task for better performance. Similar to HIP, we implement dimension weighting and parameter regularization in the likelihood. Full technical details of the fitting procedure are provided in Sec. 15.2 of the supplementary material.

The days 91–120 are used for evaluation ($T_{\text{test}} = 120$). We measure prediction performance using the absolute percentile error (APE) metric [1], which accounts for the long-tailness of online popularity—e.g., the impact of an error of 10000 views is very different for a video getting 20000 views per day compared to a video getting two million views a day.

We first compute the percentile scale of the number of views accumulated between days 91 and 120. APE is defined as

$$APE = |\text{Per}(\hat{N}_{120}) - \text{Per}(N_{120})|$$

where $\hat{N}_{120}$ and $N_{120}$ are the predicted and observed number of views between days 91 and 120; the function $\text{Per}(\cdot)$ returns the percentile of the argument on the popularity scale.

C. Models and Baseline

We consider two 3-D PCMHP models: PCMHP($3, 2$) and PCMHP($3, 3$). The former treats the tweets as a Hawkes dimension (see Theorem 1) and is thus susceptible to computational explosion for high-tweet counts given the quadratic complexity of computing cross- and self-excitation. The latter is an inhomogeneous Poisson process with no self- or cross-exciting dimension. We, therefore, fit PCMHP($3, 2$) solely on videos that have less than 1000 tweets on days 1–90; we fit PCMHP($3, 3$) on all videos.

We use as a baseline the HIP [1], a parametric popularity prediction model discussed in Section II-A. HIP, however, is designed for use in a forecasting setup. That is, HIP requires the actual counts of tweets and shares in the prediction window $[T_{\text{train}}, T_{\text{test}}]$ to get forecasts for the view counts on $[T_{\text{train}}, T_{\text{test}}]$. To adapt HIP for the prediction setup (i.e., the tweets and shares are not available at test time), we feed HIP for each of the days 91–120 the time-weighted average of the daily tweet and share counts on $1 - 90$, i.e., $(1 / \sum_{t=1}^{90} t) \sum_{t=1}^{90} t \cdot \#tweets(t)$ and $(1 / \sum_{t=1}^{90} t) \sum_{t=1}^{90} t \cdot \#shares(t)$, which assigns a higher weight to more recent counts.
Fig. 4. Comparison of fits and predictions of our proposal PCMHP(3, 2) and the baseline HIP [1] for views (left), shares (center), and tweets (right) for a sample video from ACTIVE: a trailer for the 2014 movie Whiplash (id 7d_jQycedQGo). The first 90 days are used to fit model parameters, while the next 30 days (indicated by the gray shaded area) are unseen by the model and used for evaluation. HIP does not predict the share and tweet counts, as it treats these as exogenous inputs. The blue shaded area shows prediction uncertainty computed for the PCMHP(3, 2) fits.

D. Results

In Table II, we tabulate the mean, median, and standard deviation of the percentile errors for each model.

|                         | ACTIVE 20% (n=2834) | DYNAMIC (n=585) |
|-------------------------|----------------------|-----------------|
|                         | PCMHP (3,3)          | PCMHP (3,2)     | HIP             |
|                         | PCMHP (3,3)          | PCMHP (3,2)     | HIP             |
| Mean                    | 4.82                 | 7.36            | 8.12            |
| Median                  | 2.55                 | 4.69            | 4.96            |
| StdDev                  | 7.13                 | 8.34            | 9.89            |

Note: Best-performing score in bold.

In Table II, we see a reversal of performance ranking: PCMHP(3, 2) performs best, followed by HIP and PCMHP(3, 3). This result corroborates our claim in Section II-D that applying the heuristic of censoring event times leads to information loss. We see that PCMHP(3, 2) (trained on tweet times) can better capture the popularity dynamics of the most complex videos (which are also the most interesting) compared to PCMHP(3, 3) (trained on tweet counts).

V. Interaction Between COVID-19 Cases and News

In the previous section, we have validated the predictive power of the PCMHP. Here, we shift our attention to the interpretability of PCMHP-fitted parameters. We showcase how PCMHP can link online and offline streams of events by learning the interaction between the COVID-19 daily case counts and publication dates of COVID-19-related news articles for 11 different countries during the early stage of the pandemic.

A. Dataset

We curate and align two data sources.

The first dataset contains COVID daily case counts from the Johns Hopkins University [45]. The dataset is a set of date-indexed spreadsheets containing COVID reported case counts split by country and region. We focus on the following 11 countries: U.K., USA, Brazil, China, France, Germany, India, Italy, Spain, Sweden, and Philippines. We select the same countries as [4], to which we add the Philippines.

The second dataset contains timestamps of COVID-19-related news articles provided by the NLP startup Aylien [46]. This dataset is a dump of COVID-related English news articles from 440 major sources from November 2019 to July 2020. We filter the Aylien dataset for news articles that mention the selected 11 countries in the headline. To improve relevancy, for China, we also use several COVID-related keywords (such as coronavirus, covid, and virus). Last, we only select articles from popular news sites with an Alexa rank of less than 150. Such news sources include Google News and Yahoo! News.

For each country, we fit PCMHP(2, 1) with $E = \{cases\}$ and $E^c = \{news\}$. We consider as $t = 0$ the first day on which a minimum of ten cases were recorded. Except for China, which had cases as early as January 2020, the initial time for each country in our sample lies between February and March 2020. We only consider data until $t = 120$, with time measured in days.

B. Incorporating News Information

To demonstrate the utility of news information in modeling COVID case counts, we compare the predictive performance of PCMHP(2, 1) with three variants that leverage different granularities of news information. First, we compare with PCMHP(1, 1), which does not use news information at all. Second, we compare with PCMHP(2, 2) that uses daily aggregated news counts. Last, to test whether the exact timing of news is important, we disaggregate daily news counts by adding a uniform jitter to each time, similar to what is done in
fit the PCMHP\( (2,1) \) to this dataset. We call this baseline PCMHP\( (2,1) \)-jitter.

Similar to Section IV, we split our timeframe into a training period \( [0, T^{\text{train}} = 90] \) and a testing period \( [T^{\text{train}}, T^{\text{test}} = 120] \). In our training period, we fit the models and perform hyperparameter tuning; in our testing period, we sample from the fitted models and evaluate performance. We measure performance using the symmetric mean absolute percentage error (SMAPE), given by

\[
\text{SMAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{|F_t - \hat{A}_t| + |\hat{A}_t - F_t|}{F_t + \hat{A}_t},
\]

where \( F_t \) and \( \hat{A}_t \) are the forecasted and actual values at time \( t \), respectively.

Across our sample of 11 countries, we see in Fig. 5 that PCMHP\( (2,1) \) has the best performance compared to the three baselines and incorporating more granular news information leads to better predictive performance. We observe that the news-agnostic PCMHP\( (1,1) \) and the day aggregated PCMHP\( (2,2) \) models do not fit the data well and cannot capture the complex COVID case count dynamics. This supports our claim in Section II-D that the application of data-altering heuristics leads to loss of information. However, by incorporating timestamped news information, we see significant performance improvement, and we can match the trend in the case time series. We also see subtle performance improvement by incorporating exact news times [PCMHP\( (2,1) \)-jitter versus PCMHP\( (2,1) \)].

C. Results

Fig. 6 shows the daily COVID-19 case counts and daily news article volume of the PCMHP\( (2,1) \) fits for India and Italy. We show the plots for the other countries, the table of parameter estimates, and the goodness-of-fit analysis in Sec. 16 of the supplementary material [37]. Visible from Fig. 6, PCMHP\( (2,1) \) captures well the dynamics of both countries. Based on the sample-based fit score introduced in the supplementary material, the actual COVID-19 case counts for India and Italy fall within the model’s prediction interval for 97% and 61% of the time, respectively.

D. Cluster Countries Based on Model Fittings

The parameters capture different aspects of the interaction between news and cases. Here, we cluster the fitted parameter sets across countries to identify groups that have similar diffusion profiles. To render the scale of parameters comparable across countries, we rescale the maximum daily number of cases for each country over the considered timeframe to be 100, fit the PCMHP\( (2,1) \) on this scaled data, and perform \( K \)-means clustering on the resulting parameter sets.

The \( k = 5 \) clusters are shown in Table III and visualized in Fig. 7 using t-SNE [47]. The first cluster (U.K., Germany, and Spain) has high \( \alpha^{12} \) and low \( \alpha^{11} \). The second cluster—made solely of Brazil—has both a high \( \alpha^{12} \) and a very high \( \alpha^{11} \). With high \( \alpha^{12} \), the two clusters contain countries where news strongly preempts cases. The third cluster (China and France) has a high \( \alpha^{21} \) indicative of news playing a reactive role in cases. The fourth cluster (USA, Italy, and Sweden) and fifth cluster (India and Philippines) both have low \( \alpha^{12} \) and \( \alpha^{21} \), indicating little interaction between news and cases. We notice that COVID infectiousness is much higher in the fifth cluster (India and Philippines), with \( \alpha^{11} \) greater than one (each case generates more than one case) and \( \theta^{11} \) lowest across all clusters (slow decay, therefore long influence from cases to cases). Our fits indicate that India and Philippines are countries particularly affected by COVID-19 in the early days.

VI. SUMMARY AND FUTURE WORK

This work introduces the partially censored multivariate Hawkes process (PCMHP), a generalization of the MHP where
we take the conditional expectation of a subset of dimensions over the stochastic history of the process. The PCMHP is motivated by the fact that the MHP cannot directly be fit to partially interval-censored data; the PCMHP can be used to approximate MHP parameters via a correspondence of parameters.

In this article, we derive the conditional intensity function of the PCMH by considering the impulse response to the associated LTI system. Additionally, we derive its regularity conditions which leads to a subcritical process; which we find generalizes regularity conditions of the MHP and the previously proposed MBP process. The MLE loss function is also derived for the partially interval-censored setting. To test the practicality of our proposed approach, we consider three empirical experiments.

First, we test the capability of the PCMHP in recovering MHP parameters in the partially interval-censored setting. By using synthetic data, we investigate the information loss from model mismatch and the interval-censoring of the timestamped data. Our results show that the fitted PCMHP can approximate the parameters and recover the spectral radius of the original MHP used to generate the data.

Second, we demonstrate the predictive capability of the PCMHP model by applying it to YouTube popularity prediction and showing that it outperforms the popularity estimation algorithm HIP [1].

Third, to demonstrate the interpretability of the PCMHP parameters, we fit the process to a curated dataset of COVID-19 cases and COVID-19-related news articles during the early stage of the outbreak in a sample of countries. By inspecting the country-level parameters, we show that there is a demonstrable clustering of countries based on how news predominantly played its role: whether it was reactionary, preemptive, or neutral to the rising level of cases.

A. Limitations and Future Work

The number of parameters estimated with the PCMHP($d, e$) increases quadratically with the process dimension $d$, meaning the curse of dimensionality applies. However, since the primary use case for PCMHP($d, e$) is in popularity prediction across social media platforms (i.e., where the dimensions represent platforms), which is a low-dimensional scenario, the curse of dimensionality is not a significant concern. Difficulties would arise, however, if PCMHP($d, e$) were applied at the user level (i.e., where the dimensions correspond to individual users). In such cases, future work would be required to extend PCMH by employing scalable variants of the Hawkes process. These approaches generally reduce the $d$ dimensions into a more compact $r$-dimensional space ($r \ll d$) through factorization of the interaction matrix [48], [49]. This is a direction for future research.

There are three other areas where future work can be explored. First, theoretical analysis on the approximation error of the model mismatch (i.e., fitting Hawkes data to the PCMH model) should be performed, since we only performed an empirical evaluation in this work. Second, methods to approximate the conditional intensity should be investigated, as the current solution relies on the computationally heavy discrete convolution approximation. Last, given that the PCMHP($d, e$), by construction, is not self- and cross-exciting in the $E$ dimensions, an open research question is whether we can construct a process that retains the self- and cross-exciting properties in all dimensions while also being flexible enough to be used in the partially interval-censored setting.

REFERENCES

[1] M. A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, “Expecting to be HIP: Hawkes intensity processes for social media popularity,” in Proc. 26th Int. Conf. World Wide Web (WWW), 2017, pp. 735–744.

[2] A. G. Hawkes, “Spectra of some self-exciting and mutually exciting point processes,” Biometrika, vol. 58, no. 1, pp. 83–90, 1971.

[3] M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, “A tutorial on Hawkes processes for events in social media,” in Frontiers of Multimedia Research. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2017, pp. 191–218.

[4] R. Browning, D. Salem, K. Mengersen, V. Rivoirard, and J. Rousseau, “Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID,” PLoS ONE, vol. 16, pp. 1–28, Apr. 2021.

[5] L. Shlomovich, E. A. Cohen, N. Adams, and L. Patel, “Parameter estimation of binned Hawkes processes,” J. Comput. Graphical Statist., vol. 31, no. 4, pp. 990–1000, 2022.

[6] D. Daley and D. Vere-Jones, An Introduction to the Theory of Point Processes, vol. 1, 2nd ed. (Probability and its Applications). New York, NY, USA: Springer-Verlag, 2003.

[7] M.-A. Rizoiu et al., “Interval-censored Hawkes processes,” J. Mach. Learn. Res., vol. 23, no. 338, pp. 1–84, 2022.

[8] H. J. T. Unwin et al., “Using Hawkes Processes to model imported and local malaria cases in near-elimination settings,” PLOS Comput. Biol., vol. 17, no. 4, pp. 1–18, Apr. 2021.

[9] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, “Seismic: A self-exciting point process model for predicting tweet popularity,” in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (SIGKDD), 2015, pp. 1513–1522.

[10] R. Kobayashi and R. Lambiotte, “TiDeH: Time-dependent Hawkes process for predicting retweet dynamics,” in Proc. 10th Int. AAAI Conf. Web Social Media (ICWSM), pp. 191–200.

[11] S. Mishra, M. A. Rizoiu, and L. Xie, “Feature driven and point process approaches for popularity prediction,” in Proc. 25th ACM Int. Conf. Knowl. Manage. (CIKM), pp. 1069–1078.

[12] A. Zadeh and R. Sharda, “How can our tweets go viral? Point-process modelling of brand content,” Inf. Manage., vol. 59, no. 2, 2022, Art. no. 103594.

[13] R. Crane and D. Sornette, “Robust dynamic classes revealed by measuring the response function of a social system,” PNAS, vol. 105, no. 41, pp. 15649–15653, 2008.
