UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Modeling the Online Spread of Ideas in a Finite Attention Environment

by Pio Gabrielle Battad Calderon

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr. Marian-Andrei Rizoiu

University of Technology Sydney
Faculty of Engineering and Information Technology

September 2024

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Pio Gabrielle Battad Calderon*, declare that this thesis is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *School of Computer Science, Faculty of Engineering and Information Technology* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

SIGNATURE: Production Note: Signature removed prior to publication.

Pio Gabrielle Battad Calderon

DATE: 5th September, 2024

PLACE: Sydney, Australia

ABSTRACT

Online social systems are challenging to model due to the heterogeneity of human behavior influenced by diverse cultural, economic, historical, and political factors. Additionally, measurements from these systems tend to be incomplete and noisy due to platform constraints and unpredictable human actions. Despite these challenges, online social systems are governed by foundational mechanisms that can be modeled to gain insights into collective behaviors. This thesis explores models of the spread of ideas in online social systems with three primary objectives: learn the latent mechanisms that can explain the observed noisy data, predict future online diffusions, and evaluate the impact of external interventions. The first contribution is the Opinion Market Model (OMM), a two-tier system of the online opinion ecosystem that jointly captures inter-opinion interactions and the impact of positive interventions in a finite attention environment. The OMM outperforms state-of-the-art models in understanding opinion dynamics and can be leveraged as a testbed to evaluate media as an intervention to redirect attention from extremist to moderate opinions. The second contribution is the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of separable Hawkes process that can jointly capture the influence of source, content, and cascade-level factors on the spread dynamics of online items. The BMH model excels in predicting content popularity in the cold-start setup and can differentiate the impact of different headline styles across publishers. The third contribution is the development of the Partially Censored Multivariate Hawkes Process (PCMHP), which addresses the challenge of fitting the self- and cross-exciting multivariate Hawkes process in the partially interval-censored setting. The PCMHP can model cross-platform data with limited availability, such as mixed event-timestamp data and daily-aggregated counts, and outperforms existing models in predicting YouTube popularity. This thesis advances our understanding of the spread of ideas in online social systems, providing robust models for explaining, predicting and influencing online behavior.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to everyone who has supported me throughout this journey.

I am deeply indebted to my supervisor, Dr. Marian-Andrei Rizoiu, for his constant guidance, support, and patience throughout my PhD. I am grateful for the opportunities he has opened for me, which have significantly enriched my academic and professional growth.

I am grateful to UTS for providing the support and technical resources essential for my research.

My thanks also go to my colleagues at the Behavioral Data Science Lab - Elaine, Frankie, Rohit, Jooyoung, Quyu, Philipp, and Lin - for the conversations and support.

Finally, I would like to extend my deepest gratitude to my family and friends: to my parents and siblings for their unwavering love and encouragement, to my friends here in Sydney, and to my support system back in Manila. Thank you all.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

1. **Pio Calderon**, Rohit Ram, and Marian-Andrei Rizoiu. “Opinion Market Model: Stemming Far-Right Opinion Spread using Positive Interventions." *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18, 177-190. 2024. (Chapter 2)

2. **Pio Calderon** and Marian-Andrei Rizoiu. “What Drives Online Popularity: Author, Content or Sharers? Estimating Spread Dynamics with Bayesian Mixture Hawkes." *ECML-PKDD 2024*. (Chapter 3)

3. **Pio Calderon**, Alexander Soen, and Marian-Andrei Rizoiu. “Linking Across Data Granularity: Fitting Multivariate Hawkes Processes to Partially Interval-Censored Data." *IEEE Transactions on Computational Social Systems*, Vol. 12, 1, 25-37, 2025. (Chapter 4)

4. Quyu Kong, **Pio Calderon**, Rohit Ram, Olga Boichak, and Marian-Andrei Rizoiu. “Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems." *Proceedings of the ACM Web Conference 2023*, pp. 1813-1821. 2023. (summary presented in Chapter 4)

OTHERS :

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

6. **Pio Calderon**, Lean Palma, Franz Kappel, Aurelio De Los Reyes. “Control, sensitivity and identification of a cardiovascular-respiratory system model." *Modelling, Simulation and Applications of Complex Systems*, 359, p.151. 2021.

7. **Pio Calderon**, Mustafa Habib, Franz Kappel, Aurelio De Los Reyes. “Control aspects of the human cardiovascular-respiratory system under a nonconstant workload." *Mathematical Biosciences*, 289, pp.142-152. 2017.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE **Page**

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Online social systems are challenging to model due to the heterogeneity of humans, who have diverse preferences, behaviors, and goals, and whose actions are contextual and influenced by cultural, economic, historical, and political factors. Compounding this challenge is the fact that measurements gathered from these systems tend to be incomplete, due to platform or data privacy constraints, and noisy [56] due to unpredictable human behaviors, potential manipulation [125], and biases [77]. Despite these challenges, online social systems are still governed by foundational mechanisms, which when incorporated into our models allows us to gain insight on how these online social systems work, and ultimately be able to account for the real unobserved human processes that generate the collective behavior in online social system measurements [26]. For instance, the spread of ideas online demonstrates self-exciting (*rich-gets-richer*) behavior, allowing us to leverage a class of temporal point processes called the *Hawkes process* [47]. Similarly, limited human attention due to cognitive limits (as exemplified by Dunbar's number [31], which suggests that humans can only maintain 150 relationships at once) leads to certain online content becoming popular while others fading into obscurity.

This thesis explores models of the spread of ideas in online social systems to solve three objectives: (1) *learn* the foundational mechanisms that generate the incomplete and noisy observed data, (2) *predict* the future of online diffusions, and (3) *evaluate* external interventions that aim to control these online diffusions. The novelty of this thesis lies in the inclusion of contextual and domain-specific expert opinion to impose structural assumptions into the Hawkes process.

Our approach is grounded in the finite nature of online attention [126], which leads to some ideas (or opinions, in the context of Chapter 2) dying out while only a few persist. We aim to uncover the mechanism of how online opinions interact with one another by exploring an analogy between an opinion ecosystem of limited attention and an economic market of limited resources. In the same manner that goods can complement each other (bread and butter) or compete for market share (Pepsi and Coke), can opinions reinforce or inhibit one another? Furthermore, we are interested in understanding the effectiveness of positive interventions [46, 50, 113], external signals which redistribute attention from extremist and toward moderate opinions. We require an approach that jointly models inter-opinion interactions and the influence of positive interventions. Our first research question is: **Can we model the online opinion ecosystem as a finite attention environment where opinions cooperate or compete for market share and test the sensitivity of the online opinion ecosystem to positive interventions?**

We also observe that the online spread of ideas is influenced by factors at multiple levels. At the lowest level, the spread dynamics of an online item (e.g. an online news article) hinges on the popularity of the user sharing the content [4] (e.g. user tweeting a link to the article). On the other hand, the nature and category of the content itself are influential, as various topics resonate with different target groups [103, 118]. At the highest level, the source of the online item (e.g. the publisher of the article) also affects the spread dynamics [87], e.g. users would be more willing to share news from reputable sources and be reluctant to share news from lesser-known blogs. Our second research question is: **Can we model the online spread of ideas accounting for the intertwining influence of source, content, and cascade-level factors?**

As previously noted, online data is often imperfect. While some social media platforms provide the timing of events (e.g. Twitter/X retweets), others only offer interval-censored counts (e.g. Facebook likes, Youtube views) due to privacy concerns. This data incongruency, which we call the partially interval-censored setting, poses challenges for modeling cross-platform interaction dynamics with the multivariate Hawkes process, as it requires event timing data for inference. Our last research question is: **Can we devise a method to fit the multivariate Hawkes process in the partially interval-censored setting?**

Thesis Overview. This thesis explores the three research questions in detail, as outlined below.

In Chapter 2, we solve the first research question by introducing the Opinion Market Model (OMM), a two-tier online opinion ecosystem that jointly models inter-opinion interactions and the role of positive interventions. In the first tier of the OMM, the size of the

opinion attention market is modeled using a multivariate discrete-time Hawkes process. In the second tier, opinions are allowed to cooperate and compete for limited attention using the market share attraction model. We fit the OMM to a synthetic dataset and show that our proposed estimation algorithm attains convergence. Next, we apply the OMM to two datasets: the first comprising Facebook and Twitter discussions on moderate and far-right opinions about bushfires and climate change, the second capturing popular VEVO artists' Youtube and Twitter attention volumes. On both datasets we show the OMM outperforms the state-of-the-art models and is able to capture latent opinion interactions. We then apply the OMM in a counterfactual analysis to show the effectiveness of media coverage as a positive intervention to mitigate far-right opinion spread.

In Chapter 3, we tackle the second research question by developing the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of Hawkes process that allows us to jointly learn the influence of source-, content- and spread-level feature sets on how widely and rapidly online items spread. We train the BMH model on two real-world retweet cascade datasets referencing articles from controversial and traditional media publishers and evaluate on two learning tasks (cold start popularity prediction and temporal profile generalization performance). We show that the BMH model outperforms the state-of-the-art models and predictive baselines on both tasks. Lastly, we run a counterfactual analysis on the trained BMH models to show how the effectiveness of headline writing styles (neutral, clickbait, inflammatory) varies across publishers.

In Chapter 4, we address the third research question by proposing the Partially Censored Multivariate Hawkes Process (PCMHP), a novel multivariate temporal point process that has a parameter equivalence with the multivariate Hawkes process but unlike the latter can be fit in the partially interval-censored setting. We test the PCMHP on three case studies. First, using a synthetic dataset we demonstrate that the PCMHP can approximate MHP parameters and recover the spectral radius of the process. Second, we show that the PCMHP outperforms the fully interval-censored popularity estimation algorithm Hawkes Intensity Process (HIP) in predicting Youtube popularity, highlighting that modeling time-stamped data using point process dimensions indeed improves prediction performance. Third, we demonstrate qualitative insights from PCHMP parameter fits from a dataset of daily COVID-19 case counts and COVID-19-related news articles, revealing hidden interaction patterns between cases and news reporting.

Lastly, we summarize the thesis and discuss future directions in Chapter 5.

AUTHOR DECLARATION

The following chapter contains content from the following publication.

Pio Calderon, Rohit Ram, and Marian-Andrei Rizoiu. “Opinion Market Model: Stemming Far-Right Opinion Spread using Positive Interventions." *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18, 177-190. 2024.

Author Contributions: P.C. led the research for this study, managed data collection, and conducted the experiments. R.R. supplied stance classifier output for the data. M.A.R. provided supervision throughout the project. P.C. and M.A.R. collaboratively developed the model, interpreted the resuls, and contributed to writing of the manuscript.

Production Note:
Signature removed prior to publication.

Pio Calderon

Production Note:
Signature removed prior to publication.

Rohit Ram

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

CHAPTER 2

OPINION MARKET MODEL: STEMMING FAR-RIGHT OPINION SPREAD USING POSITIVE INTERVENTIONS

Online extremism has severe societal consequences, e.g. normalizing hate speech, user radicalization, and increased social divisions. Various mitigation strategies have been explored to address these consequences. One strategy uses positive interventions: controlled signals that add attention to the opinion ecosystem to boost certain opinions. To evaluate the effectiveness of positive interventions, we introduce the Opinion Market Model (OMM), a two-tier online opinion ecosystem model that considers both inter-opinion interactions and positive interventions. The size of the opinion attention market is modeled in the first tier using the multivariate discrete-time Hawkes process; in the second tier, opinions cooperate and compete for market share, given limited attention using the market share attraction model. We demonstrate the convergence of our proposed estimation scheme on a synthetic dataset. Next, we test OMM on two learning tasks, applying to two real-world datasets to predict attention market shares and uncover latent relationships between online items. The first dataset comprises Facebook/Twitter discussions containing moderate and far-right opinions about bushfires and climate change. The second dataset captures popular VEVO artists' YouTube/Twitter attention volumes. OMM outperforms the state-of-the-art predictive models on both datasets and captures latent cooperation-competition relations. We uncover (1) self- and cross-reinforcement between far-right and moderate opinions on the bushfires and (2) artist relations that correlate with real-world interactions such as collaborations and long-lasting feuds. Lastly, we use OMM as a testbed for positive interventions and show how media coverage modulates the spread of far-right opinions.

2.1 Introduction

Online social media platforms are fertile grounds for deliberation and opinion formation [45]. Opinions thrive in the *online opinion ecosystem*, interconnected online social platforms where they interact – cooperating or competing for the finite public attention [126].

We delineate two types of interventions to mitigate the spread of extremist views. *Negative interventions* aim to subtract attention from the opinion ecosystem by placing fact-check warnings on postings [80], shadowbanning [132] or outright banning extremist social media groups and accounts [52]. While negative interventions are effective [24], they are available solely to the social media platforms that tend to use them sparingly [93].

Positive interventions [40], such as misinformation debunking [46, 113] and increasing media coverage [50], mitigate extremist views by adding attention to the online opinion ecosystem through informing the public, redistributing attention away from extremist, and toward moderate views. Such interventions are typically in the hands of government and media agencies [95]. Testing the viability of positive interventions requires capturing reactions to interventions and inter-opinion interactions.

This work develops a model for the dynamics of the opinion ecosystem and a testbed for evaluating positive interventions. We focus on two open questions. The first question explores the analogy between opinions and economic goods. In a competitive economic market of limited resources, coexisting goods can interact in one of two ways: either they compete for market share (*substitute* brands, like Pepsi and Coke) or reinforce each other (*complementary* items, like bread and butter). We argue that opinions in the online ecosystem behave similarly, allowing us to leverage market share modeling tools [25]. The first research question is: **Can we model the online opinion ecosystem as an environment where opinions cooperate or compete for market share?** We propose the Opinion Market Model[1] (OMM), a two-tier model to address this question. Fig. 2.1 illustrates a simple opinion ecosystem under a single intervention $X(t)$ (shown in yellow in the top panel of Fig. 2.1), featuring two opinions (denoted as 0 and 1) on a single social media platform, where the intervention could represent the level of media coverage relevant to the opinions. Each opinion has two polarities: far-right supporters (+) and moderate debunkers (-). The exogenous signal $S(t)$ (shown in gray in the top panel of Fig. 2.1) and intervention $X(t)$ modulate the dynamics of the opinions' sizes. Exogenous signals are naturally occurring events like bushfires, floods, or political speeches. Interventions, like increased media coverage, are designed to add attention to the opinion ecosystem, increasing the market share of certain

[1] The code and datasets are available at https://github.com/behavioral-ds/opinion-market-model.

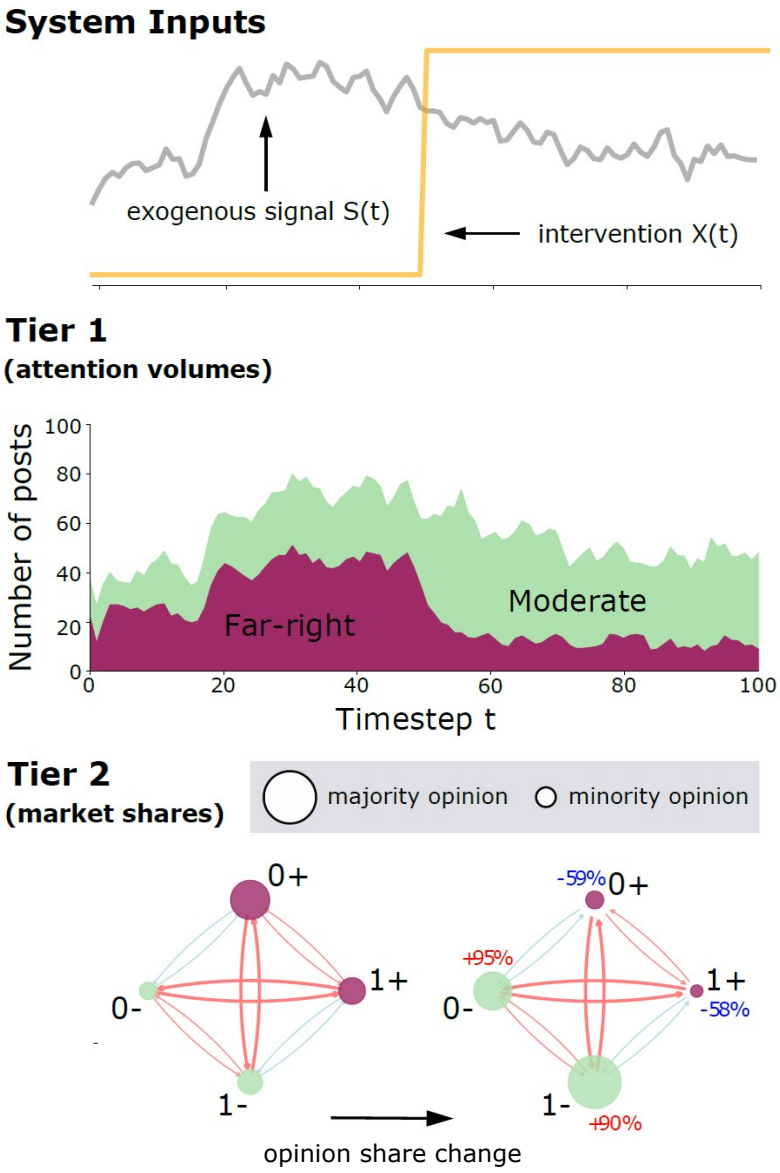


Figure 2.1: We illustrate how the positive intervention $X(t)$ (in Eq. (2.11)) suppresses far-right opinions on a simulated toy opinion ecosystem with two far-right (0+, 1+) and two moderate (0-, 1-) opinions. For instance, 0+ and 1+ can represent the opinions “Greens policies caused the Australian bushfires" and “mainstream media cannot be trusted," respectively; 0- and 1- can be obtained as their negations. Top row: the exogenous signal $S(t)$ (in Eq. (2.5)) and the intervention $X(t)$. Middle row: total daily opinion market size from our model's first tier, split into far-right (+) and moderate (-) opinion volumes. Bottom row: market shares and the interactions between the four opinions from our model's second tier. Nodes are opinions; their sizes indicate market share; edges show exciting (red) and inhibiting (blue) relations. $X(t)$ suppresses far-right opinions for $t > 50$. Shown are average market shares before (left) and after (right) $t = 50$.

opinions while suppressing others. The first tier of OMM (middle row in Fig. 2.1) uses a discrete-time Hawkes process to estimate the size of the opinion attention market – that is, the daily number of postings featuring opinions. The Hawkes process has been widely used to model online attention [103, 134] due to its ability to account for exogenous factors and the endogenous "word-of-mouth" through its self- and cross-exciting property. The second tier of OMM (bottom row in Fig. 2.1) leverages a market share attraction model to capture opinion interactions – we assume that online opinions compete for the users' limited online attention [38, 124]. For the example in Fig. 2.1, opinions 0− and 1+ have a strong reinforcing relation (shown as red arrows), while 1− and 1+ have a weak competing relation (blue arrows).

We test OMM on two real-world datasets[1]. The first contains Facebook and Twitter discussions expressing moderate and far-right opinions on bushfires and climate change [61]. The second captures the YouTube views and Twitter mentions for the most popular VEVO artists' songs in 2017 [126]. We evaluate OMM on two tasks: predicting attention market share and exposing relationships between online items. For the predictive task, OMM outperforms the current state of the art in market share modeling (Correlated Cascades [134] and Competing Products [120]) and predictive baselines on both datasets. For the second task, we leverage the OMM to expose the relations between opinions on the two platforms. For the bushfire case study, no significant interactions occurred on Facebook, as postings were collected from far-right public groups with limited interaction with the opposing side. On Twitter, we observe self-reinforcement behavior of both far-right and moderate opinions, probably due to the *echo chamber effect* [23] – reinforcing one's beliefs due to repeated interactions with users sharing similar ideologies on social platforms. Surprisingly, we notice that opposing views reinforce each other, probably due to the deliberative nature of Twitter, where far-right sympathizers and opponents oppose each other. For the VEVO artists case study, we uncover nontrivial pairwise interactions of music artists correlating with real-world relationships – such as Ariana Grande's and Calvin Harris' reinforcement relationship due to their collaboration "Heatstroke" and Taylor Swift's and Justin Bieber's inhibiting relationship.

Our second research question is: **Can we test the sensitivity of the opinion ecosystem to positive interventions?** OMM accounts for positive interventions – controlled external signals to boost certain opinions. In Fig. 2.1 an intervention is performed for $t > 50$, which suppresses the far-right opinions (+), leading to the shrinking of their market share. We use OMM for two tasks: first, to estimate whether interventions effectively shape the opinion ecosystem and, second, to construct what-if scenarios as synthetic testbeds for future

interventions. For the bushfire case study, we test whether news coverage from reputable and controversial media outlets suppresses or aids the spread of far-right opinions. We fit OMM twice: with and without media coverage. We find a better fit with the intervention, suggesting that media coverage actively shapes the opinion ecosystem. We perform synthetic what-if experiments: we vary the level of media coverage, simulate the system and observe the effect on opinion market shares. On Facebook, reputable media coverage reduces the prevalence of far-right opinions. On Twitter, both reputable and controversial media coverage suppress far-right opinions. However, for some opinions (like “Mainstream media cannot be trusted"), reputable news backfires increasing far-right opinions share.

The main contributions of the work are as follows:

1. A novel two-tier model of the opinion ecosystem that allows studying opinion interactions through an economics-based cooperation-competition lens. We introduce simulation and estimation algorithms and study the convergence with synthetic tests.

2. A synthetic testbed to uncover interactions across sympathizers and opponents of far-right opinions and likely effects of positive interventions via media coverage.

3. A curated dataset of Twitter and Facebook discussions on bushfires/climate change.

2.2 Related Work

We focus the discussion of related work on models for cooperative-competitive interaction in a set of co-diffusing online items. These models need to be both *predictive* and *interpretable* (usually generative models). We have observed a lack of recent research in this area, with few works dating after 2017. Closely related to our proposal is the Correlated Cascades (CC) model [134], a variant of the multivariate Hawkes process to model product adoption across a set of competing products in a social network. It estimates the interaction parameter β, tuning the market cooperation or competition level. A limitation of CC is that all products share a single β value. This simplifies existing asymmetric relationships and assumes that all brands either cooperate or compete. OMM addresses this issue by fully modeling these asymmetric relationships. Another closely related work is the Competing Products (CP) model [120], a multivariate Hawkes model for product adoption/use where the frequency of use is affected by the usage of other products. Limitations of the work are the absence of the assumption of limited attention and the possibility of negative intensities since competitive interactions are modeled as negative parameters. OMM avoids the weaknesses of CP by using a multiplicative model, thereby avoiding negative intensities and defining opinion

shares as fractions of the total attention volume. The SLANT model [28] and the follow-up SLANT+ [65] extend the CP model to differentiate between a user's latent and expressed opinion and uses a similar Hawkes process to model the intensity. However, SLANT requires fine-grained network information for training, which is prohibitive considering that online platforms are becoming more stringent with fine-grained data access [122]. On the other hand, OMM requires only opinion counts for training.

Ethics of Opinion Moderation and Broader Perspectives. OMM is intended to model interactions between opinions and be used as a testbed for evaluating positive interventions for opinion moderation. As any tool, OMM is unaware of the intention of its user and, in theory, could be used by oppressive regimes to silence or manipulate the liberal opinions of their opponents [95]. In addition, the fundamental value of freedom of speech for democratic societies implies that non-widely accepted opinions also have the right to be expressed. The scientific literature studies this ethical conundrum in the context of Countering Violent Extremism (CVE) initiatives [8, 95]. When viewing OMM as an AI evaluation tool supporting CVE initiatives [33], these ethical issues can be mitigated using online CVE frameworks in liberal democracies [48]. We argue that the implementing body is responsible for OMM's ethical usage, and CVE regulations should be leveraged to mitigate malicious intent.

Causal Impact. OMM measures the effect of media coverage on the opinion market shares using a generative model to disentangle endogenous and exogenous factors from observational data, similar to [36, 37, 104]. Our model works on aggregate observational data (i.e., opinion counts), and it does not prove the causal impact of media coverage on individual opinion formation (i.e., behavior change). We would require a pre-test/post-test control group design to achieve true causal links. Previous work [1, 43, 57] provides evidence of the interventional role of media coverage. In Section 2.9, we explore this further in a what-if experiment to demonstrate how the level of media coverage affects opinion market shares.

2.3 Preliminaries

We introduce two classes of models that form the foundation of our approach: (1) the discrete-time Hawkes process [14], a model of event counts that display self-exciting behavior, and (2) the market share attraction model [25], a marketing model that uncovers the latent competitive structure of brands and interaction with marketing instruments.

2.3.1 Discrete-Time Hawkes Process

The discrete-time Hawkes Process (DTHP) [14] is the discrete-time analogue of the popular self-exciting Hawkes process [47], where instead of modeling the occurrence of events given by $t \in \mathbb{R}^+$, we model the event count $N(t)$ on $[t-1, t)$ for $t \in \mathbb{N}$.

The DTHP is characterized by the conditional intensity function $\lambda(t)$, defined as the expected number of events that occur at time t, conditioned on the history $H_{t-1} = \{N(s) | s < t\}$. For a DTHP, $\lambda(t)$ is given by

$$\lambda(t) = \mathbb{E}[N(t)|H_{t-1}] = \mu + \sum_{s<t} \alpha \cdot f(t-s) \cdot N(s), \tag{2.1}$$

where μ is the baseline count of events, α determines the level of self-excitation and is the expected number of events produced by a single event and $f(t)$ is the triggering kernel, which controls the influence of the past events on the present. We specify $f(t)$ with the geometric probability mass function $f(t) = \theta(1-\theta)^{t-1}$, $t \in \mathbb{N}$, the discrete-time analogue of the exponential distribution [14]. Given $\lambda(t)$, model specification is completed by specifying a probability mass function for the count $N(t)$. Following [14], we set $N(t) \sim \text{Poi}(\lambda(t))$.

2.3.2 Market Share Attraction Model

In marketing literature, *market share attraction models* (MSAMs) [25] model the competitive structure of a set of M brands in a product category, predict their market shares, and evaluate how a set of marketing instruments affects resulting market shares. MSAMs assume that the market share s_i of brand $i \in \{1 \dots M\}$ is proportional to consumers' attraction $\mathscr{A}_i$ towards brand i:

$$s_i = \frac{\mathscr{A}_i}{\sum_{j=1}^{M} \mathscr{A}_j} \in [0,1]. \tag{2.2}$$

$\mathscr{A}_i$ is typically modeled as a parametric function of a set of K marketing instruments $\{X_{ki}\}_{k=1}^{K} \in \mathbb{R}^K$, where X_{ki} gives the value of the k^{th} marketing instrument for brand i. We typically specify $\mathscr{A}_i$ as

$$\mathscr{A}_i = \exp\left(\beta_i + \sum_{k=1}^{K} \sum_{j=1}^{M} \gamma_{kij} X_{kj}\right), \tag{2.3}$$

where β_i measures the inherent attraction of brand i and $\gamma_{kij} \in \mathbb{R}$ measures the effect of the value of the k^{th} marketing instrument for brand j on brand i's attraction. Whether γ_{kij} is positive (negative) is indicative of the excitatory (inhibiting) relationship from brand j to brand i through marketing instrument X_{kj}.

MSAMs are interpreted via the model elasticity $e(s_i, X_{kj})$, the ratio of the percent change in the market share s_i given a percent change in the value of the k^{th} marketing instrument for brand j. For example, an elasticity of $e(s_i, X_{kj}) = 0.1$ means that a 1% increase in X_{kj} corresponds to a 0.1% increase in s_i. That is,

$$e(s_i, X_{kj}) = \frac{\partial s_i / s_i}{\partial X_{kj} / X_{kj}} = \frac{\partial s_i}{\partial X_{kj}} \cdot \frac{X_{kj}}{s_i}. \tag{2.4}$$

The elasticity $e(s_i, X_{kj})$ captures the overall effect of brand j's marketing instrument X_{kj} on brand i's market share s_i: both the *direct effect* of X_{kj} on s_i, controlled by γ_{kij}, and the *indirect effect* of X_{kj} on s_i through its effect on the attraction of other brands $\{j \neq i\}$.

2.4 The OMM Model

In this section, we develop a two-tier model of the opinion ecosystem. The first tier models the total size of the opinion attention market on multiple online platforms. The second tier models the market share of opinions on each platform. Next, we introduce a scheme for parameter estimation.

OMM consists of two tiers; the first tier, which we call the *opinion volume model*, tracks the size of the opinion attention market, while the second tier, the *opinion share model*, tracks the market shares of the different opinions. Table 2.1 summarises the notation for important variables in the OMM. The full table is available in Appendix A.1.

2.4.1 Opinion Volume Model

Suppose our opinion ecosystem consists of P social media platforms. The opinion volume model tracks the attention volume, i.e. the number of opinionated posts $N^p(t)$, on each platform $p \in \{1, \dots P\}$ and time $t \in \mathbb{N}$. We model $\{N^p(t)\}_p$ as a $P-$dimensional DTHP (defined analogous to the multivariate Hawkes process [47]) with conditional intensity $\{\lambda^p(t)\}_p$,

$$\lambda^p(t) = \mu^p \cdot S(t) + \sum_{q=1}^{P} \sum_{s<t} \alpha^{pq} \cdot f(t-s) \cdot N^q(s). \tag{2.5}$$

In contrast to Eq. (2.1), we use a time-varying exogenous signal $S(t)$, which accounts for the baseline volume of events of exogenous origin. The signal $S(t)$ accounts for natural tendencies and events (i.e., epidemics, elections) and typically cannot be controlled. We introduce a scaling term μ^p for each platform p such that $\mu^p \cdot S(t)$ represents the exogenous opinion count for platform p on time t.

Notation	Interpretation
P	number of social media platforms
M	number of opinion types
K	number of positive interventions
T	terminal time
Variables	
$S(t)$	input signal, volume of exogenous events
$X_k(t)$	input signal, k^{th} positive intervention
$s_i^p(t)$	market share of opin. i on platform p at time t
$\lambda_i^p(t)$	conditional intensity of opinion i on platform p
$\lambda^p(t\|i)$	intensity of opin. i on plat. p, assuming independence among opinions
$N_i^p(t)$	#posts with opinion i on platform p at time t
$e(s_i^p(t),\cdot)$	opinion share model elasticity
Data	
n_t^p / $n_{i,t}^p$	#posts on platform p at time t / with opinion i
$s_{i,t}^p$	fraction of opin. i posts on platf. p at time t
Parameters	
μ_j^p	exogenous scaling term for opin. j on platf. p
α^{pq}	excitation parameter for intra-platform ($p=q$) and inter-platform (for $p\neq q$) dynamics
θ	memory parameter, describing how fast an event is forgotten, $\theta\in[0,1]$
γ_{ik}^p	direct effect of the k^{th} intervention on share of opinion i on platform p
β_{ij}^{pq}	direct effect that opinion j on platform q has on share of opinion i on platform p.

Table 2.1: Summary of important quantities and notation in Chapter 2.

Since online platforms are not siloed and have significant user overlap, we allow the P platforms to interact via intra- and inter-platform excitation. The parameter $\alpha^{pq}>0$ sets the level of intra-platform (for $p=q$) and inter-platform (for $p\neq q$) excitation. Lastly, we set $N^p(t)\sim \text{Poi}(\lambda^p(t))$.

2.4.2 Opinion Share Model

With the attention volumes for each platform p estimated in the first tier, the second tier models the market share $s_i^p(t)$, calculated as the fraction of opinionated posts on platform p conveying opinion i. Given the limited attention market size, opinions compete for attention

within each platform.

Suppose that there are M different opinion types. We set $N_i^p(t)$ to be the number of opinionated posts conveying opinion i on platform p on time t, and $\lambda_i^p(t)$ to be its conditional intensity. Using the notion of limited attention [134], we relate $\lambda_i^p(t)$ to $\lambda^p(t)$ in Eq. (2.5) by introducing the market share $s_i^p(t) \in [0,1]$ as the fraction of opinion i posts on platform p. That is,

$$\lambda_i^p(t) = \lambda^p(t) \cdot s_i^p(t), \tag{2.6}$$

and $\sum_{i=1}^M s_i^p(t) = 1$.

Similar to Eq. (2.2), we model $s_i^p(t)$ with attraction $\mathscr{A}_i^p(t)$,

$$s_i^p(t) = \frac{\mathscr{A}_i^p(t)}{\sum_{j=1}^M \mathscr{A}_j^p(t)}. \tag{2.7}$$

Leveraging the MNL form in Eq. (2.3), we define attraction

$$\mathscr{A}_i^p(t) = \exp \mathscr{T}_i^p(t), \tag{2.8}$$

where $\mathscr{T}_i^p(t)$ consists of two parts, accounting for *interventions* and *endogenous* dynamics, and is described in detail below.

(*Interventions.*) First, we introduce a set of K positive interventions $\{X_k(t)\}_k$ that modify the opinion market shares in the opinion ecosystem. The interventions $\{X_k(t)\}_k$ have a different role to $S(t)$ in Eq. (2.5), as the latter modifies the attention market size. To reduce the influence of noise, we use the smoothed version of $\bar{X}_k(t)$ with the kernel f, given by

$$\bar{X}_k(t) = \sum_{s<t} f(t-s) \cdot X_k(s). \tag{2.9}$$

We introduce the parameter $\gamma_{ik}^p \in \mathbb{R}$ to measure the direct effect of the k^{th} intervention on the market share of opinion i on platform p. If γ_{ik}^p is positive (negative), then $X_k(t)$ reinforces (inhibits) opinion i on platform p.

(*Endogenous dynamics.*) Second, we model the contribution of endogenous dynamics on the attraction of opinions. To represent the prevalence of opinion j on platform q, we make use of the conditional intensity $\lambda^q(t|j)$,

$$\lambda^p(t|j) = \mu_j^p \cdot S(t) + \sum_{q=1}^P \sum_{s<t} \alpha^{pq} \cdot f(t-s) \cdot N_j^q(s), \tag{2.10}$$

which models the dynamics of opinion j independent of other opinions. We introduce the parameter $\beta_{ij}^{pq} \in \mathbb{R}$ to measure the direct effect that opinion j on platform q has on the

market share of opinion i on platform p. Similar to γ_{ik}^p, we allow β_{ij}^{pq} to be positive (negative), representing a reinforcing (inhibiting) relationship from opinion j to i on platform q and p, respectively.

Given these components, we model $\mathcal{T}_i^p(t)$ as

$$\mathcal{T}_i^p(t) = \underbrace{\sum_{k=1}^{K} \gamma_{ik}^p \cdot \bar{X}_k(t)}_{\text{interventions}} + \underbrace{\sum_{q=1}^{P} \sum_{j=1}^{M} \beta_{ij}^{pq} \cdot \lambda^q(t|j)}_{\text{endogenous}}, \tag{2.11}$$

where $\mu^p = \sum_{j=1}^{M} \mu_j^p$.

2.4.3 Inference

Over the observation period $t \in \{1, \ldots, T\}$, assume that we observe the exogenous signal $S(t)$, the K interventions $\{X_k(t)\}_k$, and the number $n_{i,t}^p$ of posts conveying opinion i on platform p for each i and p. Our goal is to estimate the parameter set $\mathbf{\Theta} = \{\mu_j^p, \alpha^{pq}, \theta, \gamma_{ik}^p, \beta_{ij}^{pq}\}$.

The structure of our two-tier model allows us to cast parameter estimation as a two-tier optimization problem. Let $\mathbf{\Theta}_1 = \{\mu^p, \alpha^{pq}, \theta\}$. The key observation here is that the first-tier parameter set $\mathbf{\Theta}_1$ can be estimated using only the opinion volume model in Eq. (2.5), independent of the opinion share model in Eq. (2.11). By optimizing the likelihood $\mathscr{L}_1(\mathbf{\Theta}_1 | \{n_t^p\}_{p,t})$ of the platform-level volumes $\{n_t^p\}_{p,t}$, we can obtain an estimate $\hat{\mathbf{\Theta}}_1$ of $\mathbf{\Theta}_1$. The likelihood is given by

$$\mathscr{L}_1(\mathbf{\Theta}_1 | \{n_t^p\}_{p,t}) \propto \sum_{t=1}^{T} \sum_{p=1}^{P} \left[n_t^p \log \lambda^p(t) - \lambda^p(t) \right]. \tag{2.12}$$

The second-tier parameter set $\mathbf{\Theta}_2 = \{\mu_j^p, \gamma_{ik}^p, \beta_{ij}^{pq}\}$ can be obtained by optimizing the likelihood $\mathscr{L}_2(\mathbf{\Theta}_2 | \hat{\mathbf{\Theta}}_1, \{n_{i,t}^p\}_{i,p,t})$ of the opinion volumes $\{n_{i,t}^p\}_{i,p,t}$, conditioned on our estimate of the first-tier parameters $\hat{\mathbf{\Theta}}_1$.

$$\mathscr{L}_2(\mathbf{\Theta}_2 | \mathbf{\Theta}_1, \{n_{i,t}^p\}_{i,p,t}) \propto \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \left[n_{i,t}^p (\log \lambda^p(t) + \log s_i^p(t)) - (\lambda^p(t) \cdot s_i^p(t)) \right] \tag{2.13}$$

Our full estimated parameter set is given by $\hat{\mathbf{\Theta}} = \hat{\mathbf{\Theta}}_1 \cup \hat{\mathbf{\Theta}}_2$. The technical details of the estimation and the derivation of the likelihoods $\mathscr{L}_1(\cdot)$ and $\mathscr{L}_2(\cdot)$ and gradients $\partial_{\mathbf{\Theta}_1} \mathscr{L}_1(\cdot)$ and $\partial_{\mathbf{\Theta}_2} \mathscr{L}_2(\cdot)$ are available in Appendix A.2.

Runtime Complexity. Evaluating $\mathscr{L}_1(\mathbf{\Theta}_1 | \{n_t^p\}_{p,t})$ in Eq. (2.12) has time complexity $\mathcal{O}(T^2 \cdot P^2)$. This can be seen by noting that calculating $\lambda^p(t)$ in Eq. (2.5) has complexity $\mathcal{O}(T \cdot P)$, which is nested in a $T \cdot P$ loop to evaluate Eq. (2.12). On the other hand, evaluating

$\mathscr{L}_2(\boldsymbol{\Theta}_2|\hat{\boldsymbol{\Theta}}_1, \{n_{i,t}^p\}_{i,p,t})$ has time complexity $\mathcal{O}(T^2 \cdot P \cdot M \cdot [K + P^2 \cdot M])$. First, we observe that the sum in Eq. (2.13) has complexity $\mathcal{O}(T \cdot P \cdot M \cdot \mathscr{S})$, where $\mathscr{S}$ is the complexity of evaluating $s_i^p(t)$. Note that in the inner loop of Eq. (2.13) the complexity of computing $s_i^p(t)$ dominates over $\lambda^p(t)$. From Eqs. (2.10) and (2.11), we see that $\mathscr{S} = \mathcal{O}(T \cdot K + T \cdot P^2 \cdot M)$. Plugging this expression for $\mathscr{S}$ in $\mathcal{O}(T \cdot P \cdot M \cdot \mathscr{S})$ yields $\mathcal{O}(T^2 \cdot P \cdot M \cdot [K + P^2 \cdot M])$.

Simulation. Suppose we are given the opinion volume $n_{i,0}^p$ at time $t = 0$ for each platform p and opinion i, such that $n_t^p = \sum_i n_{i,t}^p$. A sample of $n_{i,t}^p$ from OMM can be obtained by calculating the conditional intensity $\lambda_i^p(t)$ using Eq. (2.6), and then sampling $n_{i,t}^p$ from Poi $(\lambda_i^p(t))$. We obtain $\{n_{i,t}^p\}_{i,p,t}$ by repeating these steps over $\{1, \ldots, T\}$.

Numerical Considerations. To improve model fit in our real-world case studies, we implement three augmentations to the model and estimation method, outlined below and fully detailed in Appendix A.4. First, we modify the attraction $\mathscr{A}_i^p(t)$ in Eq. (2.7) to prevent numerical overflow/underflow. Second, we add a regularization term in the second-tier optimization problem in Section 2.4.3 to impose structural constraints on $\{\hat{\gamma}_{ik}^p\}$ and improve estimation. Third, we apply log-scaling on $\lambda^q(t|j)$ and standardize both $\lambda^q(t|j)$ and $\bar{X}_k(s)$ in Eq. (2.11) to solve scaling issues.

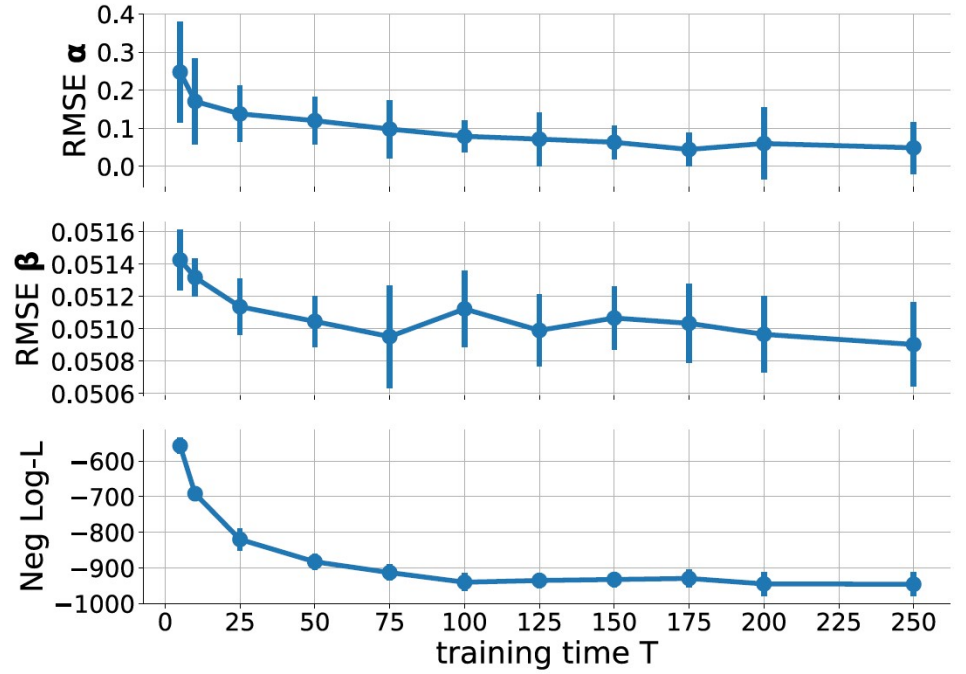
Stability Assumption. We implicitly assume that the opinion attention market is stable over the timeframe of the analysis, in the sense that the parameters $\boldsymbol{\Theta}$ governing the behavior of the process stay constant within the timeframe. In situations where this assumption is not expected to hold (e.g. extreme events) and parameters change within the timeframe, a change-point model extension [14] of the OMM is necessitated.

2.5 Learning with Synthetic Data

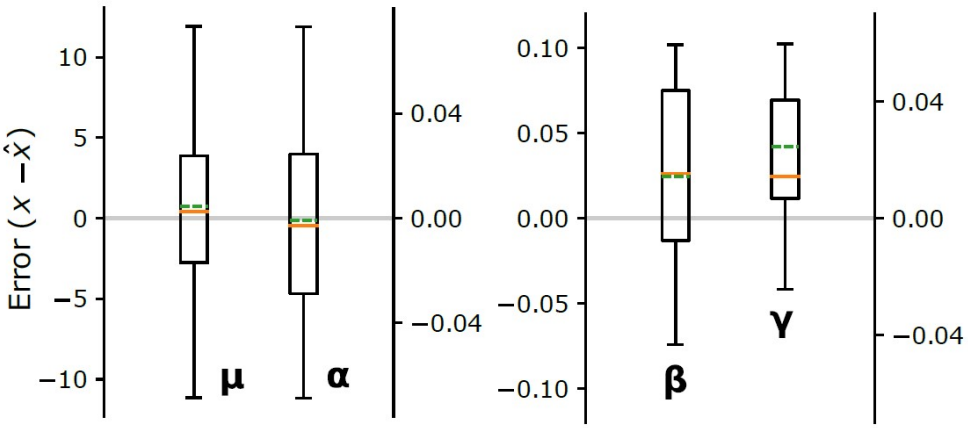
In this section, we consider the parameter estimation task with synthetic data. First, we discuss our experimental setup and the synthetic dataset. Next, we show that parameter recovery error decreases and stabilizes as we increase the training time T and the number of samples $n_{samples}$.

Experimental Setup. We set $P = M = K = 2$. We set $[\mu_1^1, \mu_2^1, \mu_1^2, \mu_2^2] = [15, 5, 5, 20]$, and $\theta = 0.5$ and draw $\alpha^{pq} \sim \text{Unif}(0, 0.5)$, $\beta_{ij}^{pq} \sim \text{Unif}(0, 0.1)$ and $\gamma_{ik}^p \sim \text{Unif}(0, 0.1)$. The exogenous signals are $S(t) = 1$, $X_1(t) = 5\sin(0.1x) + 5$, and $X_2(t) = 10\sin(0.05x + 1.25) + 10$.

We construct our synthetic dataset using the simulation algorithm in Section 2.4.3 to get 400 samples of opinion volumes $\left\{n_{i,t}^p\right\}_{i,p,t}$ for $t \in \{1, \ldots, T = 300\}$. We implement joint fitting [102]: we partition the 400 samples into 20 groups of $n_{samples} = 20$ samples each.



((a))



((b))

Figure 2.2: Parameter recovery results on synthetic data. In (a), we show the convergence of the RMSE of the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ estimates and the negative log-likelihood as we increase the training time T. In (b), we show the difference between our estimates for $\{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\}$ and the true values. Dashed green lines and orange lines are the mean and median values, respectively.

The likelihoods $\mathscr{L}_1(\boldsymbol{\Theta}_1)$ and $\mathscr{L}_2(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_1)$ of each group are maximised to get an estimate $\hat{\boldsymbol{\Theta}}$, yielding 20 sets of parameter estimates.

Model Evaluation. To study the convergence of our learning algorithm, we compute the root mean-squared error (RMSE) of our estimated $\hat{\boldsymbol{\Theta}} = \{\hat{\mu}_j^p, \hat{\alpha}^{pq}, \hat{\theta}, \hat{\gamma}_{ik}^p, \hat{\beta}_{ij}^{pq}\}$ with respect to the true $\boldsymbol{\Theta}$, following [120]. We report the average RMSE per parameter type, where the average is taken over the components of the matrix or tensor corresponding to the parameter type.

In Fig. 2.2(a), we see that training on a longer timeframe leads to lower RMSE for $\hat{\alpha}^{pq}$ and $\hat{\beta}_{ij}^{pq}$ and better model fit measured by the likelihood $\mathscr{L}_2$. Results for $\hat{\mu}_j^p$, $\hat{\theta}$ and $\hat{\gamma}_{ik}^p$, and on varying $n_{samples}$ are in Appendix A.3.

In Fig. 2.2(b), we plot the difference distribution between our estimates and the true values. We recover first-tier parameters $\{\hat{\mu}_j^p, \hat{\alpha}^{pq}\}$ well, as evidenced by our mean estimates coinciding with the true values. We observe a slight overestimation of $\{\hat{\gamma}_{ik}^p, \hat{\beta}_{ij}^{pq}\}$, given the nonconvexity of $\mathscr{L}_2$ and the high dimensionality of the second-tier parameter set.

2.6 Real-World Datasets

This section introduces two real-world datasets we have curated to evaluate the OMM.

2.6.1 Bushfire Opinions Dataset

We construct the *Bushfire Opinions dataset*, containing 90 days of Twitter and Facebook discussions about bushfires and climate change between Nov. 1, 2019 to January 29, 2020. The Facebook postings are a subset of the *SocialSense* dataset [61], which was collected with the approval of the Human Research Ethics Committee of the University of Technology Sydney (approval number: ETH19-3877); we select posts and comments about bushfires and climate change (*SocialSense* also contains discussions around COVID-19). Using CrowdTangle[2], we unobtrusively collected public far-right Australian Facebook discussions, identified via a digital ethnographic study (see [61] and Appendix A.7 for details). We build the Twitter discussions using the Twitter Academic v2 API; we collect tweets emitted between November 1, 2019 to January 29, 2020 that mention bushfire keywords such as *bushfire, arson, australiaburns* (see the full list in Appendix A.7). We use the AWS Location Service[3] to geocode users based on their free-text location and description fields and filter only for tweets from Australian users.

[2] https://www.crowdtangle.com/
[3] https://aws.amazon.com/location/

Our focus on the 2019-2020 Australian bushfires is motivated by the availability of human-annotated topics, opinions [61] and stance classifiers [96] trained on the same topic and timeframe. We use these classifiers to filter and label our dataset.

Moderate and Far-Right Opinion Labeling. To filter and label relevant Facebook and Twitter postings, we use the textual topic and opinion classifiers developed by [61], with a reported 93% accuracy in classifying Facebook and Twitter posts on bushfires and climate change. We select the following most prevalent six opinions, covering 95% of Twitter and 81% of Facebook postings:

0. Greens policies are the cause of the Australian bushfires.
1. Mainstream media cannot be trusted.
2. Climate change crisis is not real / is a UN hoax.
3. Australian bushfires and climate change are not related.
4. Australian bushfires were caused by random arsonists.
5. Bushfires are a normal summer occurrence in Australia.

Furthermore, we deploy the far-right stance detector introduced by [96] – which leverages a textual homophily measurement to quantify the similarity between Twitter users and known far-right activists. On the *Bushfire Opinions Twitter dataset*, the stance detector achieves a 5-fold CV AUC ROC score of 0.889. An opinion is labeled as *far-right* if the posting agrees with the opinion (denoted as +), or *moderate* if the posting disagrees with the opinion (-). We represent our opinion set as $\{(i-, i+) | i \in \{0, \ldots, 5\}\}$. In summary, we consider $P = 2$ platforms with 74,461 tweets and 7,974 Facebook postings labeled with $M = 12$ stanced opinions. We aggregate posting volumes by the hour, resulting in $T = 2,160$ time points over 90 days from Nov 1, 2019, to Jan 29, 2020.

Exogenous Signal S and Intervention X. The exogenous signal $S(t)$ (Eq. (2.5)) modulates the total size of the attention market in the first tier of OMM. We use the 5-day rolling average of the Google Trends query *bushfire+climate change* in Australia, normalized to a max value of 1. Google Trends captures the baseline interest on topics [110] and is a proxy for offline events (ex. actual bushfires and government measures) [75].

The interventions $\{X_k(t)\}$ modulate the market share of far-right and moderate opinions. Our interventions consist of two sources of news coverage: reputable (R) mainstream Australian publishers (e.g., The Sydney Morning Herald, Canberra Times, Crikey) and controversial (C) international publishers (e.g., Sputnik News, Breitbart, Red State). For each

opinion $i \in \{0,\ldots,5\}$, we consider a pair of interventions $(R_i(t), C_i(t))$, consisting of reputable and controversial daily news volumes discussing opinion i. We assemble the intervention set $\{X_k(t)\}$ ($K = 12$) so that the first six interventions correspond to $\{R_0(t),\ldots,R_5(t)\}$ while the last six correspond to $\{C_0(t),\ldots,C_5(t)\}$.

We sourced reputable Australian news publishers from the Reputable News Index (RNIX) [63]. We query Factiva [53] to obtain the daily news volume of these outlets for each of the six opinions using a keyword search. We similarly obtain the news volumes from controversial international publishers from NELA-GT-2019 [42] using a keyword search. We subtract the Google Trends signal from the news volumes for each intervention. We compute the standardized form of $X_k(t)$ as $\hat{X}_k(t) = \text{news}_k(t) - \frac{\max_t \text{news}_k(t)}{\max_t S(t)} S(t)$. For brevity, in the bushfire case study, we denote $\hat{X}_k(t)$ as $X_k(t)$ (i.e., always in standardized form). After standardization, $X_k(t)$ indicates whether reputable or controversial media over- or under-reports relative to the public's attention.

2.6.2 VEVO 2017 Top 10 Dataset

We assemble the *VEVO 2017 Top 10* dataset by aligning artist-level time series of YouTube views and Twitter post counts ($P = 2$) for the top $M = 10$ VEVO-affiliated artists over $T = 100$ days from Jan 2, 2017 to Apr 11, 2017.

The YouTube time series are obtained from the *VEVO Music Graph dataset* [126], containing daily view counts for music videos posted by verified VEVO artists in six English-speaking countries (USA, UK, Canada, Australia, New Zealand, and Ireland). We combine the view counts for all music videos that belong to a given artist to obtain artist-level YouTube view time series. For Twitter, we leverage the Twitter API to get daily counts of posts with text containing an input query. We obtain the artist-level Twitter post time series using the artist's name as the input query.

Unlike the single exogenous signal $S(t)$ in the Bushfire Opinions dataset, we use a different exogenous signal $S_i(t)$ for each artist i – the Google Trends for each artist i. Using the set $\{S_i(t)\}$ instead of a single $S(t)$ requires several small changes to Eq. (2.5), Eq. (2.10), and the model gradients. We fully detail these changes in Appendix A.4. We do not consider any interventions $\{X_k(t)\}$ as we seek to uncover endogenous interactions across artists.

2.7 Predictive Evaluation

This section evaluates the OMM's predictive capabilities on two real-world datasets. We introduce our prediction task, evaluation metrics and baselines, then present the results.

2.7.1 Model Setup

We use a temporal holdout strategy similar to prior literature [63, 100, 103]: we fit OMM on $\mathcal{T}_{obs}$ and evaluate performance on $\mathcal{T}_{pred}$. Backtesting is another viable alternate evaluation approach; however, it is significantly more computationally intensive, and we prefer the temporal holdout. For the bushfire case study, $\mathcal{T}_{obs} = \{1, \ldots, 1800\}$ where time is in hours (i.e., days 1-75 of our period of interest) and $\mathcal{T}_{pred} = \{1801, \ldots, 2160\}$ (i.e., days 76-90). For the VEVO case study, $\mathcal{T}_{obs} = \{1, \ldots, 75\}$ and $\mathcal{T}_{pred} = \{76, \ldots, 100\}$.

We consider two tasks: (1) opinion volume prediction and (2) opinion share prediction. For the first task, we predict the total volume of opinionated posts on the P platforms during the evaluation period. We measure performance using the platform-averaged symmetric mean absolute percentage error (SMAPE) of predicted volumes $\{\bar{n}_t^p | t \in \mathcal{T}_{pred}\}$ on platform p relative to the actual volumes $\{n_t^p | t \in \mathcal{T}_{pred}\}$,

$$\text{SMAPE} = \frac{1}{P} \sum_{p=1}^{P} \left(\frac{100\%}{360} \sum_{t=1801}^{2160} \frac{|\bar{n}_t^p - n_t^p|}{|\bar{n}_t^p| + |n_t^p|} \right). \tag{2.14}$$

The predicted opinion volumes $\{\bar{n}_t^p\}$ are obtained using the OMM simulation algorithm. We (1) condition on $\{n_{i,t}^p | t \in \mathcal{T}_{obs}\}$, (2) run the algorithm to sample $\{n_{i,t}^p\}$ on $\mathcal{T}_{pred}$, then (3) sum over opinion types $\{i\}$ to get predicted opinion volumes $n_t^p = \sum_i n_{i,t}^p$. We repeat $R = 5$ times, and average over the samples to obtain $\{\bar{n}_t^p | t \in \mathcal{T}_{pred}\}$.

For opinion share prediction, we predict the opinion market shares $\{s_{i,t}^p\}$ for each platform p on the evaluation period. To evaluate how well we predict opinion market shares, we calculate the KL divergence of predicted market shares $\{\bar{s}_t^p | t \in \mathcal{T}_{pred}\}$ (obtained similar to $\{\bar{n}_t^p\}$ described above) relative to actual market shares $\{s_t^p | t \in \mathcal{T}_{pred}\}$,

$$\text{KL}^p(t) = \sum_{i=1}^{M} s_{i,t}^p \log \frac{\bar{s}_{i,t}^p}{s_{i,t}^p}. \tag{2.15}$$

2.7.2 Baselines

We compare OMM with the discretized versions of the Correlated Cascades (CC) model [134] and Competing Products (CP) model [120] – the current state-of-the-art models in product

share modeling, covered in related works. For the bushfire study, we test the effectiveness of interventions by fitting OMM without $\{X_k(t)\}$ (indicated as OMM\X).

We also consider a feature-based predictive baseline – the multivariate linear regression (MLR), used previously for online popularity prediction [92, 103]. We build MLR with a one-week sliding window of three types of features: the previous event counts, exogenous signal $S(t)$ and interventions $\{X_k(t)\}$. The predictive targets are the event counts $\{n^p_{i,t}\}$ for each point on $\mathcal{T}_{pred}$. Analogous to OMM fitted without interventions $\{X_k(t)\}$, we additionally train MLR without $\{X_k(t)\}$ (indicated as MLR\X) for the bushfire case study.

OMM, CC and CP are generative models typically designed for explainability and are known to be suboptimal for prediction [76]. In contrast, feature-driven approaches (e.g., MLR) use machine learning to predict using training features. Such approaches are designed mainly for prediction and have weaker explainability since they do not model the data-generation process [76]. In this work, we are interested in the dual tasks of predicting and explaining opinion market shares, hence our focus on generative approaches.

2.7.3 Predicting Opinion Volumes

Fig. 2.3(a) showcases the observed (blue line) and modeled (orange line) opinion volumes for the bushfire dataset. We visually observe that OMM achieves a tight fit on both the training and the prediction period (hashed area). The VEVO dataset results are shown in Appendix A.5. We further compare OMM's predictive performances against baselines. The top row of boxplots in Fig. 2.4(a) and Fig. 2.4(b) shows the platform-averaged SMAPE of predicted volumes for the bushfire and VEVO datasets, respectively. We make two observations. First, in both case studies, OMM outperforms all baselines on opinion volume prediction. Second, OMM outperforms OMM\X, indicating the role of media coverage in shaping attention.

2.7.4 Predicting Opinion Market Shares

Fig. 2.3(b) visualizes the observed (left column) and fitted during training and predicted during testing (right column) opinion market shares for the bushfire dataset. We see that the opinion distribution on Twitter has significantly more variation than on Facebook, and that OMM closely captures the trend in opinion shares on both platforms. The VEVO dataset results are in Appendix A.5. The bottom rows of Fig. 2.4(a) and Fig. 2.4(b) show the KL-divergence of predicted market shares for the bushfire (Facebook and Twitter) and VEVO (YouTube and Twitter) datasets, respectively. We make several observations. First, on the bushfire dataset, performance is better for Twitter than Facebook ($\text{KL}^{TW}(t) < \text{KL}^{FB}(t)$)

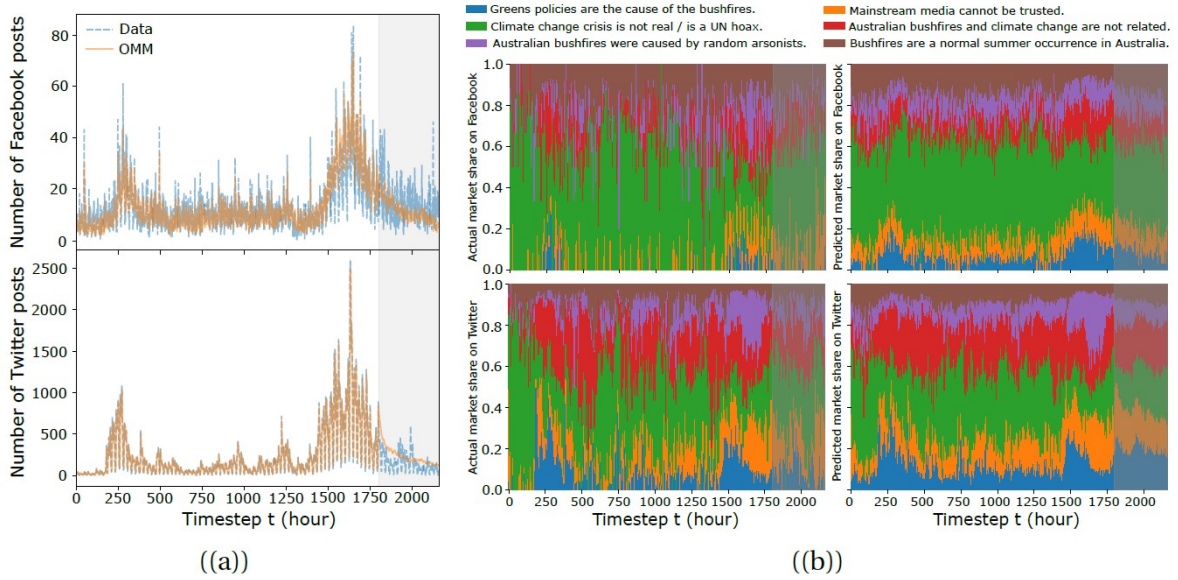


Figure 2.3: Fitting and predicting with OMM on the Bushfire Opinions dataset. We train OMM on the first 1800 timesteps and predict on timesteps 1801 to 2160 (shaded area). We show results for Facebook (top row) and Twitter (bottom row). (a) Actual (dashed blue lines) vs. fitted/predicted (orange lines) volumes; (b) Actual (left panels) and fitted during training and predicted during testing (right panels) opinion market shares on Facebook and Twitter. We aggregate the far-right and moderate opinions.

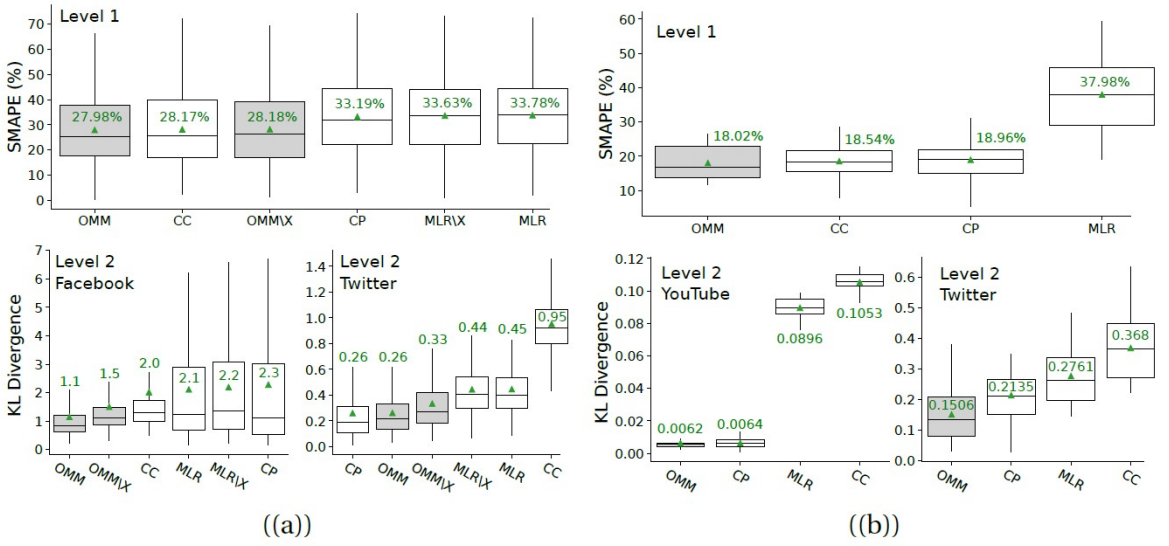


Figure 2.4: Predictive evaluation of OMM on (a) Bushfire Opinions and (b) VEVO 2017 Top 10 datasets. Boxplots are sorted left to right by the mean (shown with green triangle). Shaded boxplots correspond to versions of OMM. The top panels show the platform-averaged SMAPE of volumes on $\mathcal{T}_{pred}$. Bottom panels plot the KL divergence of predicted and actual market shares.

due to Facebook having lower opinion counts than Twitter. Similarly, on the VEVO dataset $\mathrm{KL}^{YT}(t) < \mathrm{KL}^{TW}(t)$. Second, OMM consistently outperforms all baselines on both datasets, except for Twitter on bushfires, where CP and OMM are comparable. CC performs poorly since it does not model asymmetric opinion interactions and assumes all opinions reinforce or inhibit one another. CP performs poorly on Facebook (Twitter) for the bushfire (VEVO) dataset due to CP not having the notion of limited total attention. Due to higher bushfire postings on Twitter, CP pays more attention to Twitter. Lastly, OMM with $\{X_k(t)\}$ outperforms OMM without $\{X_k(t)\}$ on the bushfire dataset, suggesting that mainstream and controversial media effectively shape the opinion ecosystem.

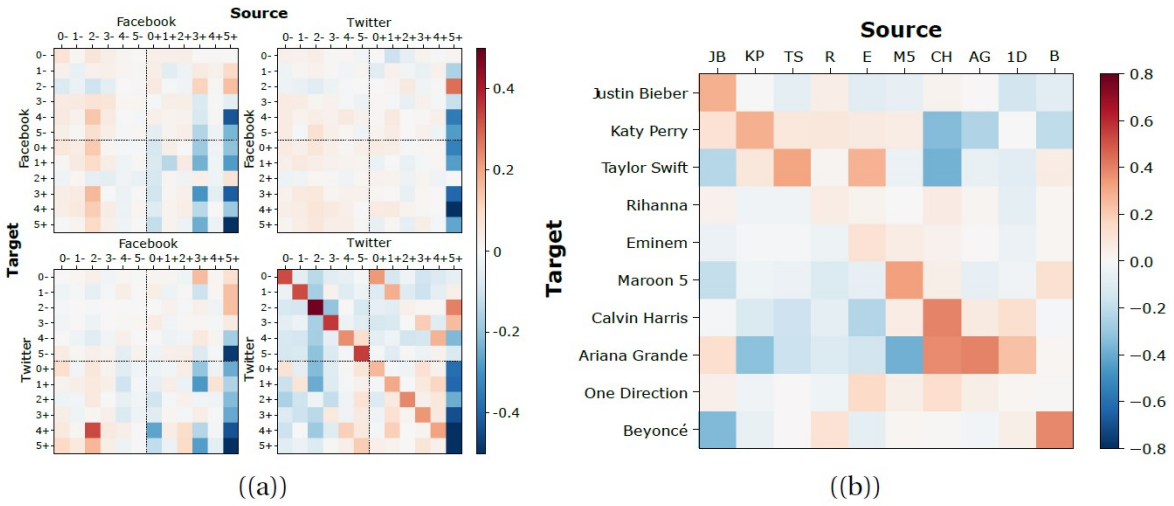
2.8 Interpreting OMM Elasticities

In this section, we leverage the fitted OMM to uncover interactions across opinions and platforms in the bushfire dataset and artists in the VEVO dataset.

2.8.1 Uncovering Opinions Interactions

To study opinion interactions in the bushfire dataset, we calculate the opinion share model elasticities (see Eq. (2.4)) accounting for the endogenous volume $\lambda^p(t|j)$ and the intervention $\bar{X}_k(s)$ (see Eq. (2.11)). The endogenous elasticities $e(s_i^p(t), \lambda^q(t|j))$ quantify the competition-cooperation interactions across opinions. The intervention elasticity $e(s_i^p(t), \bar{X}_k(t))$ quantifies the sensitivity of opinion market shares to intervention $X_k(t)$. We derive the elasticities and show results for $e(s_i^p(t), \bar{X}_k(t))$ in Appendix A.6. Fig. 2.5(a) reports the time averages of $e(s_i^p(t), \lambda^q(t|j))$.

First, we study intra-platform reinforcement (top-left & bottom-right in Fig. 2.5(a)). We see different behaviors for Facebook and Twitter. For Twitter, we have two observations. First, there is strong self-reinforcement for opinions (i.e., main diagonal), indicative of the echo chamber effect [23]. Second, there is significant cross-reinforcement among far-right sympathizers and opponents (i.e., diagonals on the upper-right & lower-left submatrices), implying exchanges or arguments between opposing camps. For Facebook, OMM detects little interaction among opinions, aside from the generally inhibitory effect of the opinions "Australian bushfires and climate change are unrelated" (3+) and "Bushfires are a normal summer occurrence" (5+) on other opinions. This is because Facebook far-right groups have limited interaction with the opposing side.



((a)) ((b))

Figure 2.5: Interpretability of OMM. (a) Endogenous elasticities $e(s_i^p(t), \lambda^q(t|j))$ across opinion pairs (i, j) on respective platforms (p, q) in the bushfire dataset. Elasticities have direction and should be read from column (source) to row (target) for the platform and within each matrix. For example, the bottom-right matrix corresponds to influences from Twitter to Twitter; the cell $\{4-, 4+\}$ ({row, column}) is the influence of opinion 4+ on 4−, positive and large meaning that 4+ has a strong reinforcing effect on 4−. (b) YouTube elasticities $e(s_i^{YT}(t), \lambda^{YT}(t|j))$ across artist pairs (i, j) in the VEVO 2017 Top 10 dataset.

2.8.2 How to Effectively Suppress Far-Right Opinions

The above implies that confrontation is not the most effective method to suppress far-right opinions, as it has the potential to backfire by bringing even more attention to them. A more effective method is boosting related counter-arguments; for instance, to suppress “Australian bushfires were caused by random arsonists" (4+) on Twitter, OMM indicates to promote “Climate change is real" (2-) and “Greens are not the cause of the bushfires" (0-). Boosting the opposite argument, i.e., “Australian bushfires were not caused by random arsonists" (4-), would backfire. The opinion “Bushfires are a normal summer occurrence in Australia" (5+) shows a different behavior: it reinforces most moderate opinions and inhibits far-right opinions. In particular, the “Bushfires are normal" opinion (5+) appears to trigger “Climate change is real" (2-), probably due to the diametric opposition nature of these opinions. The effect of 5+ on 2- holds across every pair of platforms. Additionally, on Facebook, “Australian bushfires and climate change are not related" (3+) has a similar effect on other opinions as the “Bushfires are normal" opinion (5+), probably due to the similarity of their topic content.

2.8.3 Cross-Platform Reinforcement

Cross-platform reinforcement is generally weak due to the Facebook far-right groups acting as a filter bubble. Apart from the effect of “Bushfires are normal" (5+) (see above), there is little cross-reinforcement among opinions from Twitter to Facebook. In the bottom-left matrix of Fig. 2.5(a), we see that “Australian bushfires and climate change are not related" (3+) affects other opinions in a similar way to “Bushfires are normal" (5+); furthermore, “Climate change is real" (2-) triggers “Australian bushfires were caused by arsonists" (4+).

2.8.4 Interactions Across VEVO Artists

Lastly, in Fig. 2.5(b), we shift our attention to the VEVO dataset and look at the YouTube-to-YouTube elasticities $e(s_i^{YT}(t), \lambda^{YT}(t|j))$ across our set of artists. The Twitter and cross-platform elasticities are available in Appendix A.6.

We highlight three key observations. First, there is strong self-reinforcement for most artists (i.e., the main diagonal), an intuitive result reflecting these popular artists' strong fanbase. Second, OMM picks up non-trivial artist interactions that correspond with real-world events – the animosity and friendship relations show up in their popularity dynamics. For instance, we see that Calvin Harris inhibits both Taylor Swift (the two broke up in 2016[4]) and Katy Perry (the two had a long-lasting feud[5], due to Harris pulling out of Perry's 2011 tour last minute). Similarly, Taylor Swift and Justin Bieber have a mutually inhibiting relationship. The two have a well-known uneasy relationship[6] since Justin Bieber and Selena Gomez used to date and the latter is one of Taylor Swift's close friends. Meanwhile, Calvin Harris and Ariana Grande have a reinforcing relationship, correlating with their collaboration “Heatstroke” released in March 2017. OMM picks up these relationships because we fit on online popularity driven by audience response. Fans of a given artist can choose to support or not support another artist based on real-world interactions, as indicated by the results above. Our third observation relates to the complexity of fanbase support for artists occupying the same genre: similar artists do not all just cooperate or compete for market share but can have unique pairwise relationships. For instance, Katy Perry, Taylor Swift and Ariana Grande occupy a similar niche (mainstream pop). However, our model uncovers that Taylor Swift and Katy Perry reinforce each other, while these two inhibit (and are inhibited by) Ariana Grande.

[4] people.com/celebrity/taylor-swift-calvin-harris-breakup-timeline/

[5] nme.com/news/music/katy-perry-ends-six-year-beef-calvin-harris-2128100

[6] people.com/music/justin-bieber-selena-gomez-relationship-look-back/

2.9 OMM as a Testbed for Interventions

The interventions $\{X_k(t)\}$ can lead to delayed effects in the opinion ecosystem due to the opinion dependency structure. For example, if an intervention is designed to boost a target opinion, it will indirectly boost all other opinions with a cooperative relationship with the target opinion. Furthermore, it will inhibit opinions with a competitive relationship with the target. Since elasticities only inform us of the *instantaneous* effect on opinion market shares, we perform a what-if exercise to study the role of interventions in the bushfire case study. We vary the size of the intervention and synthetically sample outcomes to observe the long-term effects of media coverage on the opinion ecosystem.

2.9.1 "What-if" Can Inform A/B Test Design

We train OMM on observational data; therefore, the inferred effects of interventions $\{X_k(t)\}$ are not causal impact estimates but rather evidence of causal effects. However, the previous section demonstrates that OMM can uncover complex relationships across opinions, providing compelling evidence that OMM is also able to uncover relationships between opinions and interventions. Therefore, the what-if exercise in this section showcases OMM as a testbed for interventions, usable for designing A/B testing that determines true causal effects. The OMM informs us of the effectiveness of interventions, allowing us to prioritize which specific interventions to test.

2.9.2 "What-if" Setup

We test the effect of interventions by synthetically increasing or decreasing their volumes past a given time point (see top panel of Fig. 2.1) and measuring the percentage change in far-right opinions. Let $k^* \in \{1,\dots,K\}$ be the index of the modulated intervention. We modulate $X_{k^*}(t)$ as $X^{(r)}_{k^*}(t) = X_{k^*}(t) + r \cdot \mu_{X_{k^*}} \cdot \mathbf{I}_{(t>1800)}$, where $\mathbf{I}_{(\cdot)}$ is the indicator function and $\mu_{X_{k^*}}$ is the mean volume of $X_{k^*}(t)$ on $\mathcal{T}_{obs}$. The parameter r controls the percent increase ($r > 0$) or decrease ($r < 0$) in media coverage beyond the change point $t = 1800$; $r = 0$ is the original $X_{k^*}(t)$. We run OMM with $X^{(r)}_{k^*}(t)$ for various r, and keep $X_k(t)$ fixed for $k \neq k^*$. We quantify the effects of intervention $X_{k^*}(t)$ as the average percent change (relative to $r = 0$) in the opinion market shares after the change point, i.e., $\mathcal{T}_{pred}$. We perform this procedure for all $k^* \in \{1,\dots,K\}$.

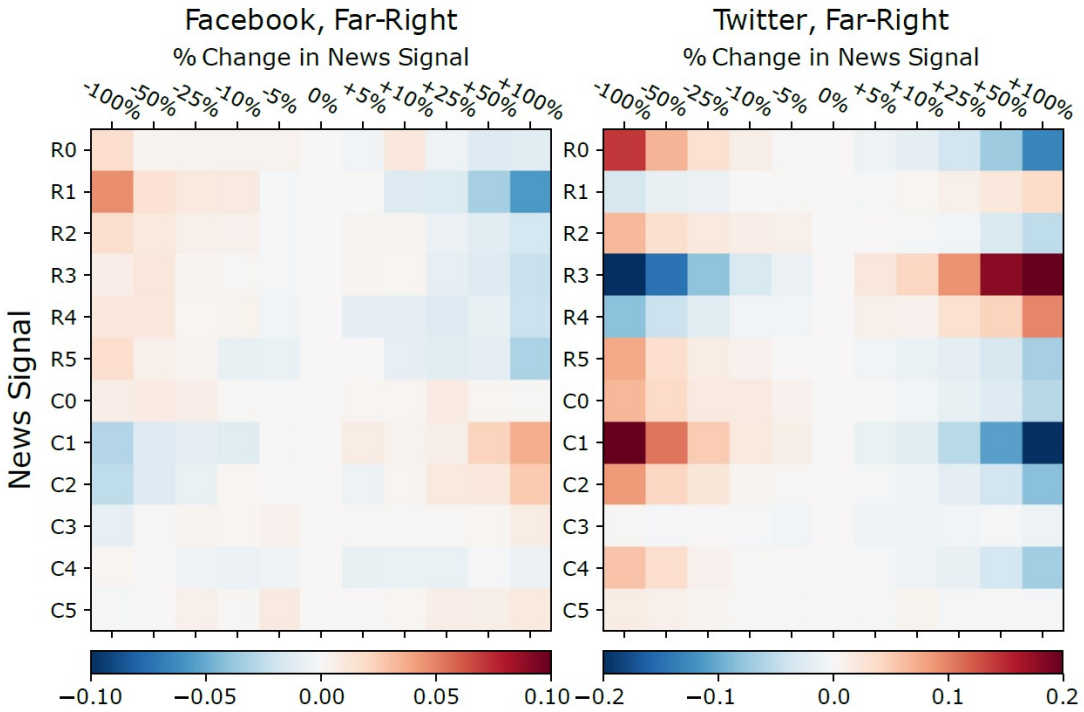


Figure 2.6: We modulate the volume of reputable (R) and controversial (C) news for each opinion (in $\{0,1,2,3,4,5\}$) from -100% to 100% of the mean volume and simulate OMM to see the percent change in the far-right (+) opinion market shares on Facebook (left) and Twitter (right).

2.9.3 How News Influences Far-Right Opinions

Fig. 2.6 shows the average percent changes in the market share of far-right opinions when modulating the interventions $\{R_i(t), C_i(t)\}$ one at a time for various r over 50 simulations. On Facebook, far-right opinions are suppressed by reputable news and reinforced by the majority of controversial news, except for news concerning "Greens policies are the cause of the Australian bushfires" (R_0) and "Australian bushfires were caused by arsonists" (R_4). On Twitter, both reputable and controversial news suppress far-right opinions, except for reputable news concerning "Australian bushfires/climate change are unrelated" (R_3), "Australian bushfires were caused by arsonists" (R_4) and to a lesser extent "Mainstream media cannot be trusted" (R_1).

We have two key insights. First, we see that the effect of the news on Facebook is modest compared to Twitter since the far-right public groups on Facebook behave as almost perfect filter bubbles in which news has little penetration. Second, indiscriminately increasing reputable news is not an effective strategy for suppressing far-right opinions on Twitter (see R_3 and R_4). Doing so can backfire since it brings even more attention to far-right users and

their narratives [89].

2.9.4 How to Effectively Use the Testbed

Assuming that A/B testing is performed by an entity in control of reputable news coverage (R_i here above), the results above indicate that the test should mainly concentrate on the effects of increasing R_1 (on Facebook), increasing R_0 and decreasing R_3 and R_4 (on Twitter). We leave as future work the design and execution of such an experiment. Our analysis in this chapter focuses on mitigating far-right opinions with media coverage. However, OMM can be leveraged as an intervention evaluation tool for information operations in other domains and fighting mis- & disinformation and online propaganda.

2.10 Summary and Discussion

This work introduces the Opinion Market Model (OMM), a novel two-tier model of the dynamics of the online opinion ecosystem. The first tier models the size of the attention market, and the second tier models opinions competing or cooperating for limited public attention under the influence of positive interventions. We develop algorithms to simulate and estimate OMM, showing the convergence using synthetic data. We demonstrate real-world applicability on a dataset of Facebook and Twitter discussions containing moderate and far-right opinions on bushfires and climate change [61] and a dataset of YouTube and Twitter attention volumes for popular artists on VEVO [126]. We show OMM predicts opinion market shares better than state-of-the-art baselines [120, 134] and uncovers latent competitive and cooperative interactions across opinions: self-reinforcement attributable to the echo chamber effect and interactions between far-right sympathizers and opponents. Lastly, we quantify the effect of reputable and controversial media coverage on Facebook and Twitter.

Scope of Study. This work focuses on the manifestation of far-right opinions in the context of the 2019-2020 Australian bushfires. Note that far-right ideology manifests in other political issues (e.g., gun control, LGBT rights, xenophobia), which we do not tackle here. Moreover, we do not focus on the general political science of far-right ideology since we are projecting onto a specific context.

AUTHOR DECLARATION

The following chapter contains content from the following publication.

Pio Calderon and Marian-Andrei Rizoiu. “What Drives Online Popularity: Author, Content or Sharers? Estimating Spread Dynamics with Bayesian Mixture Hawkes." *Accepted as Full Paper in ECML-PKDD 2024*.

Author Contributions: P.C. led the research for this study, managed data collection, and conducted the experiments. M.A.R. provided supervision throughout the project. P.C. and M.A.R. collaboratively developed the model, interpreted the results, and contributed to writing of the manuscript.

Production Note:
Signature removed prior to publication.

Pio Calderon

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

CHAPTER 3

WHAT DRIVES ONLINE POPULARITY: AUTHOR, CONTENT OR SHARERS? ESTIMATING SPREAD DYNAMICS WITH BAYESIAN MIXTURE HAWKES

The spread of content on social media is shaped by intertwining factors on three levels: the source, the content itself, and the pathways of content spread. At the lowest level, the popularity of the sharing user determines its eventual reach. However, higher-level factors such as the nature of the online item and the credibility of its source also play crucial roles in determining how widely and rapidly the online item spreads. In this work, we propose the Bayesian Mixture Hawkes (BMH) model to jointly learn the influence of source, content and spread. We formulate the BMH model as a hierarchical mixture model of separable Hawkes processes, accommodating different classes of Hawkes dynamics and the influence of feature sets on these classes. We test the BMH model on two learning tasks, cold-start popularity prediction and temporal profile generalization performance, applying to two real-world retweet cascade datasets referencing articles from controversial and traditional media publishers. The BMH model outperforms the state-of-the-art models and predictive baselines on both datasets and utilizes cascade- and item-level information better than the alternatives. Second, we perform a counter-factual analysis where we apply the trained publisher-level BMH models to a set of article headlines and show that effectiveness of headline writing style (neutral, clickbait, inflammatory) varies across publishers. The BMH model unveils differences in style effectiveness between controversial and reputable publishers, where we find clickbait to be notably more effective for reputable

publishers as opposed to controversial ones, which links to the latter's overuse of clickbait. Lastly, we introduce a two-step 'generate-then-evaluate' approach to optimise headlines before posting time, where we use text-generating AI to produce rewrites for a target headline, and then use the fitted BMH model to rank the rewrites based on predicted effectiveness. We run an experiment on Mechanical Turk and demonstrate that online respondents have a significant preference for the model-optimised headlines over their pre-optimised and previously published versions.

3.1 Introduction

Social media platforms have played an increasingly important role as distribution hubs for content. In 2023, it was reported that 69% of the U.S. adult population use social media as a news source [16], implying a significant shift in how information is consumed. Understanding how content propagates on these platforms – both the size and speed of dissemination – is vital since the impact is intrinsically tied to the level of online engagement the content receives. To command attention in today's digital age, it is not sufficient to craft high-quality content alone, but rather high-quality content that resonates with social media.

The spread of content online is influenced by factors at varying levels. At the lowest level, the breadth of a *diffusion cascade*, referring to the sequence of content shares triggered by a user, often hinges on the user's popularity as reflected by their follower count [4]. If a highly followed user shares an online item, it reaches a broader audience, increasing the likelihood that it will be shared. However, the cascade's growth is not solely dependent on user popularity. The nature and category of the shared content play crucial roles, as various topics may engage audiences in different ways [103, 118]. For news dissemination, the way an article headline is written, particularly the use of clickbait tactics to create an *information gap* to exploit the audience's curiosity [135], significantly impacts the total attention (i.e. *popularity*) the news article receives. Beyond cascade- and item-level factors, the reputation of the online item's source also affects how widely and quickly information spreads [87]. An article from a reputable source like The New York Times may spread more quickly and be taken more seriously than an article from a controversial, lesser-known blog due to the former's established credibility. Accurately modeling diffusion cascades of online content requires an approach that jointly considers these factors at different levels.

In this work, we address three open questions related to jointly modeling the influence of the source, item- and cascade-level factors on online content spread.

The first research question examines how these three levels influence the spread of online content. While prior studies have explored the effects of cascade features [115] and item-level variations [63], a comprehensive framework that jointly considers the three levels has yet to be developed. Our first question is: **Can we build a model for the spread dynamics of online content that accounts for the intertwining influence of its source, the content itself, and cascade-level factors?** To tackle this, we propose the Bayesian Mixture Hawkes (BMH) model, a novel source-level hierarchical mixture model of separable Hawkes processes that models diffusion cascades' size and temporal profile as a function of cascade- and item-level features. The left half of Fig. 3.1 showcases how the source-level BMH model

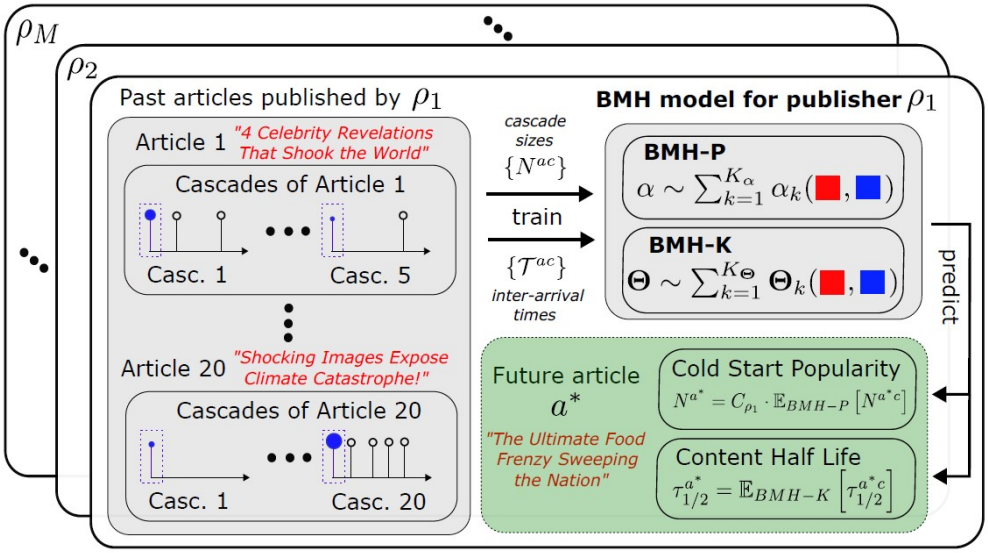


Figure 3.1: An *intuitive plate diagram* for the BMH model. *Left:* The BMH model is trained using a historical dataset: a collection of M publishers $\{\rho_1,\ldots,\rho_M\}$, items for each publisher (i.e. articles), and a set of diffusion cascades for each item. Each diffusion cascade consists of a timeline of events, here represented by a set of lollipops. *Upper Right:* The BMH is a publisher-level model that maps cascade features (shown in blue color) and article features (in red color) to a mixture of Hawkes processes. *Lower Right:* The trained BMH model (with the historical follower count distribution) can be used to infer spread dynamics of future articles based on their headlines.

learns across both the cascade and item levels from a hierarchically structured dataset (i.e., a set of items, cascade groups for each item, and feature sets attached to each). The BMH model is capable of learning different classes of Hawkes process dynamics, taking into account the ability of online content to trigger varied responses, from highly popular to largely unnoticed cascades, as well as those that fade quickly or diminish over time. The BMH learns the influence of feature sets on these classes in two ways: the location of each class in the Hawkes parameter space and the membership probability of each cascade belonging to each class. The trained BMH model can then be used to predict future items' popularity and spread dynamics from the same source (see the right half of Fig. 3.1). We test the BMH model on two hierarchical retweet cascade datasets that reference articles from controversial and reputable media publishers [63] and on two tasks: cold-start popularity prediction and temporal profile generalization performance. We show that the BMH outperforms the state-of-the-art in item popularity prediction (Dual Mixture Model [63]), Empirical Bayes approach [115] and predictive baselines for both tasks and datasets, and that the BMH model jointly leverages cascade- (i.e., the follower count of the seed user) and article-level (i.e., the article headline embedding vector) information better than

the benchmarks. Furthermore, our model ablation highlights the role of the initiating user in shaping the cascade dynamics related to controversial media, a factor less critical for cascades linked to reputable media. This distinction mirrors the diverse pathways of online information dissemination: controversial media often circulate within topical social groups [10, 54], with the initial endorser serving to validate the content, while for reputable media the publisher's reputation is the most important factor.

Our second open question relates to learning differences in the spread dynamics across news publishers: **Can we uncover across-publisher differences in how headline writing style (neutral, clickbait, inflammatory) affects published content's popularity and temporal profile?** We run a counter-factual analysis using the trained publisher-level BMH models and a labeled set of article headlines [66] to show the variation of headline style effectiveness across publishers. We find that the BMH model is able to capture nuanced publisher behavior, such as the effectiveness of inflammatory headlines for tabloids. The BMH model also unveils differences in the success of clickbait between controversial and reputable outlets, linking to existing research on clickbait fatigue and the diminishing relationship between clickbait effectiveness and volume [70, 135].

Our third open question relates to a real-world problem faced by journalists: **Can we use the BMH model to optimise the effectiveness of news headlines before posting time?** In Section 3.7 we introduce a two-step 'generate-then-evaluate' approach, where in the first step we leverage text-generating AI (e.g., GPT [84]) to produce rewrites for a given target headline, and in the second step we use the trained BMH model to rank the rewrites based on predicted cold-start popularity and half-life. We demonstrate the effectiveness of this procedure empirically through a Mechanical Turk [85] experiment, showing that model-optimised headlines have a significant higher selection rate than pre-optimised and previously published headlines.

The main contributions of the work are as follows:

1. The Bayesian Mixture Hawkes (BMH) model[1], a novel hierarchical mixture model of the joint influence of cascade- and item-level features on online item spread dynamics. On two news datasets, we show that the BMH outperforms the state-of-the-art and baselines in cold-start popularity prediction and temporal profile generalization performance.

2. A counter-factual analysis showing how headline writing style affects published content's spread dynamics. Using the BMH model we learn the differences in the effec-

[1]The Stan/CmdStanPy implementation of the BMH model is available at `https://github.com/behavioral-ds/bayesian-mixture-hawkes/`.

tiveness of headlines across publishers and show general trends across controversial and reputable media outlets.

3. A two-step procedure to optimise headlines before posting time, where we use text-generating AI to produce rewrites for a target headline, and then use the fitted BMH model to rank the rewrites based on predicted effectiveness. We demonstrate the effectiveness of this procedure through a Mechanical Turk experiment.

3.2 Related Work

In recent years, generative models, and specifically the Hawkes process [47], have been employed to model online information diffusion given their dual *predictive* and *interpretable* capabilities [5, 41, 71, 136]. However, the Hawkes process cannot incorporate feature sets in its base form since it relies only on observed temporal sequences to fit the model parameters. Numerous modifications to incorporate feature sets have been proposed to enhance model fit and predictive capabilities. A hybrid approach introduced in [76] integrates the Hawkes process with a scaling factor trained on cascade-level features to improve retweet cascade size prediction. The Empirical Bayes (EB) method [115] utilizes historical retweet sequences to link cascade features and the prior distribution of Hawkes process parameters, leading to better forecasting. The parametric Hawkes process [67] models the branching factor, i.e. the expected number of offsprings from a parent event, as a linear combination of event-level features. Lastly, the Tweedie-Hawkes process [68] improves on this by combining the Hawkes process with the Tweedie distribution to more realistically model the effect of event-level features on the branching factor. The proposed BMH model is a hierarchical model and can incorporate two levels of feature sets: the cascade- and the item (i.e., cascade-group)-level, which previous work does not cover.

Another relevant area is mixtures of point processes, employed when the data is suspected to be generated from multiple dynamical classes (i.e., parameter sets). In [129], the Hawkes process was combined with the Dirichlet distribution to model clusters of cascades. An online learning framework was introduced in [39] to fit mixtures of multivariate Hawkes processes to learn the interaction network across a set of actors. [108] introduces a generative model for mixtures of more complex point processes by using recurrent neural networks. Closest to our work is the Dual Mixture Model (DMM) [63], a generative model for cascade groups. Each cascade is sampled from a mixture of separable Hawkes processes learned jointly with their mixture probabilities. To the best of our knowledge, including feature sets into mixture models of point processes has not been explored: the BMH model solves this by learning the influence of features on the mixture components.

3.3 Preliminaries

We discuss two point process models that form the foundation of the BMH model. Section 3.3.1 presents the Hawkes Process (HP) [47], a temporal point process model that displays self-exciting behavior. Section 3.3.2 introduces the Dual Mixture Model (DMM) [63],

an approach to jointly model groups of cascades. An introduction to Bayesian hierarchical modeling, which we employ to model hierarchical data, is included in Section 3.3.3.

3.3.1 Hawkes Process

The Hawkes process (HP) [47] is a temporal point process widely used to model phenomena that display self-excitation, i.e., the likelihood of an event increases as more events occur. The HP is specified using the conditional intensity function $\lambda(t|\mathcal{H})$, the event rate at any time t conditioned on the history $\mathcal{H} = \{t_j | t_j < t\}$ of past events up to that point, i.e. $\lambda(t|\mathcal{H}) = \mu + \sum_{j=1}^{N} \alpha \cdot g(t - t_j|\boldsymbol{\Theta})$. For brevity, we drop the condition on the event history and write $\lambda(t|\mathcal{H})$ as $\lambda(t)$. Under this parametrization, a Hawkes process $\mathcal{HP}(\mu, \alpha, \boldsymbol{\Theta}|g)$ is identified with the parameters μ, α and $g(\cdot|\boldsymbol{\Theta}) : \mathbb{R}^+ \to \mathbb{R}^+$. The parameter $\mu \geq 0$ is the arrival rate of events triggered by external sources, the branching factor $\alpha \geq 0$ is the expected number of offsprings generated by a single parent event which controls the level of self-excitation from previous events, and the memory kernel $g(\cdot|\boldsymbol{\Theta})$ models the temporal decay of influence of previous events on future events controlled by the parameter set $\boldsymbol{\Theta}$. In this work, we utilize the power law kernel parametrized by $\boldsymbol{\Theta} = \{\theta, d\}$, given by $g(t|\theta, d) = \theta \cdot d^{\theta} \cdot (t + d)^{-(1+\theta)}$. Other common choices for the memory kernel are the exponential kernel $g(t|\theta) = \theta \cdot e^{-\theta t}$ and the Reyleigh kernel $g(t|\theta) = e^{-\frac{1}{2}\theta \cdot t^2}$. We focus on the power law as it has been shown in [76] to outperform these alternatives in popularity prediction. HP estimation and prediction is discussed in detail in Appendix B.1.1.

Given a collection of *complete* cascades $\mathbb{H} = \{\mathcal{H}_i\}$ where each $\mathcal{H}_i$ is completely observed (i.e. terminal time $T_i \to \infty$), and assuming no exogenous events (i.e. $\mu = 0$), the HP log-likelihood $\mathcal{L}(\alpha, \boldsymbol{\Theta}|\mathbb{H})$ splits into two log-likelihoods [63],

$$\mathcal{L}(\alpha, \boldsymbol{\Theta}|\mathbb{H}) = \mathcal{L}(\alpha|\mathbb{H}) + \mathcal{L}(\boldsymbol{\Theta}|\mathbb{H}), \tag{3.1}$$

$$\mathcal{L}(\alpha|\mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \log\left[\alpha^{N_i - 1} e^{-N_i \alpha}\right], \quad \mathcal{L}(\boldsymbol{\Theta}|\mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \sum_{t_j \in \mathcal{H}_i, j \geq 1} \log \sum_{t_z < t_j} g(t_j - t_z|\boldsymbol{\Theta}),$$

where we set $N_i = |\mathcal{H}_i|$. Under this case, Hawkes process estimation splits into two independent problems, hence the term *separable Hawkes process*. The first problem (popularity estimation) utilizes the cascade sizes $\{N_i\}$ to estimate the branching factor α by maximizing $\mathcal{L}(\alpha|\mathbb{H})$. It was shown in [63] that maximizing $\mathcal{L}(\alpha|\mathbb{H})$ is equivalent to maximizing $\sum_{\mathcal{H}_i \in \mathbb{H}} \log \mathbb{B}(N_i|\alpha)$, where $\mathbb{B}(\cdot|\alpha)$ is the Borel distribution [11]. The second problem (kernel estimation) uses the interevent-time distribution $\mathcal{T} = \{t_j - t_z\}_{t_z < t_j, t_j \in \mathcal{H}, \mathcal{H} \in \mathbb{H}}$ to estimate $\boldsymbol{\Theta}$ by maximizing $\mathcal{L}(\boldsymbol{\Theta}|\mathbb{H})$.

3.3.2 Dual Mixture Model

Maximizing Eq. (3.1) yields the best-fitting Hawkes parameter set $\{\alpha, \mathbf{\Theta}\}$. However, this approach assumes that all cascades stem from a singular parameter set, an assumption which may not hold if there are multiple dynamical classes of cascade behavior. The Dual Mixture Model (DMM) [63] was proposed to model a cascade group $\mathbb{H}$ with a mixture of K separable Hawkes processes of different parameter sets to account for different dynamical classes. Under separability, the DMM splits into two submodels: the Borel mixture model (BMM) for popularity estimation and the kernel mixture model (KMM) for kernel estimation. The BMM assumes that there exist K popularity classes accounting for the cascade sizes $\{N_i\}$, where the i^{th} class is represented by the branching factor α_i^* with probability p_i^B, i.e. $M^B = \{(\alpha_i^*, p_i^B)\}_{i=1}^K$. Similarly, the KMM assumes that there are K kernel classes accounting for the interevent-time distribution $\mathscr{T}$, where the j^{th} class is represented by the kernel parameter set $\mathbf{\Theta}_j^*$ with probability p_j^g, i.e. $M^g = \{(\mathbf{\Theta}_j^*, p_j^g)\}_{i=1}^K$. The DMM is the Cartesian product of M^B and M^g, i.e. $M = \{(\alpha_i^*, \mathbf{\Theta}_j^*, p_i^B \cdot p_j^g) | (\alpha_i^*, p_i^B) \in M^B, (\mathbf{\Theta}_j^*, p_j^g) \in M^g\}$. DMM estimation and prediction is discussed in detail in Appendix B.1.2.

3.3.3 Bayesian Hierarchical Modeling

Let θ be a parameter set of a generative process $\mathscr{P}$ and $\mathscr{D}$ be a sample from $\mathscr{P}$. Bayesian inference involves (1) quantifing our prior belief on θ through a prior distribution $\mathbb{P}(\theta)$, which could be uninformative or based on expert opinion, and then (2) updating $\mathbb{P}(\theta)$ using the data $\mathscr{D}$, with the likelihood function $\mathscr{L}(\mathscr{D}|\theta)$ serving as our weight on θ. Our result is the posterior distribution $\mathbb{P}(\theta|\mathscr{D})$, which combines our beliefs on θ based on our prior and the data:

$$\mathbb{P}(\theta|\mathscr{D}) \propto \mathscr{L}(\mathscr{D}|\theta) \cdot \mathbb{P}(\theta). \tag{3.2}$$

One advantage of Bayesian inference is its ability to accommodate the hierarchical structure of our dataset. For example, suppose that we have N data points $\{x_i\}$ which are sampled from some generative model $\mathscr{P}(\theta)$. Additionally, we are given information that each data point belong to one of m related groups. We can handle this information in three ways.

First, we ignore it and assume that all groups are drawn from the same generative model, i.e. $x_i \sim \mathscr{P}(\theta)$. This approach ignores variability across groups.

Second, we assume that the groups are independent from one another and fit a separate θ_j for each group, i.e. $x_i \sim \mathscr{P}(\theta_j)$. This approach ignores the fact that the groups are related.

The third approach, Bayesian hierarchical modeling, offers a compromise between these two by allowing variation across groups. Here, we assume that each group j has its own θ_j parameter, and $x_{j[i]} \sim \mathscr{P}(\theta_j)$, where $j[i]$ is read as 'the group data point i belongs to'. We assume that $\{\theta_j\}$ are not independent but are samples from a group-level distribution $\mathscr{Q}$ parametrized by a group-level parameter θ_{group}, i.e. $\theta_j \sim \mathscr{Q}(\theta_{group})$. Under this hierarchical framework, $\mathscr{Q}(\theta_{group})$ acts a prior for each parameter θ_j. Specifying a prior distribution for θ_{group} completes the Bayesian hierarchical model. Our posterior is a joint distribution over each group's parameter θ_j and the group-level parameter θ_{group}.

3.4 Bayesian Mixture Hawkes (BMH) Model

In this section, we develop the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of separable Hawkes processes to learn the effect of cascade-level and item-level features on cascade spread dynamics. We first describe the dataset structure that the BMH model is tailored to handle, then discuss the BMH model's objectives and the approach we adopt to address each. We then present the two components of the BMH: the popularity submodel in Section 3.4.1 and the kernel submodel in Section 3.4.2.

Assume that we are given the following dataset. First, we have a collection of items, denoted as $\mathscr{A}$, from a shared source ρ, where each item $a \in \mathscr{A}$ is characterized by the feature vector $\vec{y}^a \in \mathbb{R}^{N_y}$. If ρ is a news publisher, then $\mathscr{A}$ can represent a collection of news articles and $\vec{y}^a$ the embedding vector for article a's headline. Second, we have a set of complete cascades $\mathbb{H}^a$ for each item $a \in \mathscr{A}$, where cascade $\mathscr{H}^{ac} \in \mathbb{H}^a$ has size N^{ac}, interevent distribution $\mathscr{T}^{ac}$, and is described by the feature vector $\vec{x}^{ac} \in \mathbb{R}^{N_x}$. In our news example, $\mathbb{H}^a$ can represent discussions on Twitter related to article a, which we obtain by collecting all retweet cascades initiated with a tweet linking article a's URL. The feature vector $\vec{x}^{ac}$ can be taken as the follower count of the cascade's initiator.

We model the generative process of $\mathscr{H}^{ac}$ using a separable power-law HP with parameter set $(\alpha^{ac}, \mathbf{\Theta}^{ac})$, i.e. $\mathscr{H}^{ac} \sim \mathscr{HP}(\alpha^{ac}, \mathbf{\Theta}^{ac} | g)$. We construct the BMH as a model for $(\alpha^{ac}, \mathbf{\Theta}^{ac})$ with three goals: (1) jointly learn across the item set $\mathscr{A}$, (2) learn the relationship between $\vec{y}^a$ and $(\alpha^{ac}, \mathbf{\Theta}^{ac})$, and (3) learn the link between $\vec{x}^{ac}$ and the same parameters. We handle goal (1) by using a two-level Bayesian hierarchical model to jointly fit across each item $a \in \mathscr{A}$ and to tie together cascade- and item-level information. For goals (2) and (3), we consider a mixture of separable HPs with K_α classes for α^{ac} and $K_{\mathbf{\Theta}}$ classes for $\mathbf{\Theta}^{ac}$. We learn the influence of $\vec{y}^a$ and $\vec{x}^{ac}$ on $\{\alpha^{ac}, \mathbf{\Theta}^{ac}\}$ through the centers and membership probabilities of the K_α popularity classes and $K_{\mathbf{\Theta}}$ kernel classes.

Due to the separability of the underlying HP, the BMH divides into two independent models: (1) BMH-P, the *popularity* submodel for α^{ac}, and (2) BMH-K, the *kernel* submodel for $\mathbf{\Theta}^{ac}$. Table 3.1 lists the notation for important variables in the BMH and the mapping to real-world quantities in the datasets in Section 3.5. The full table of notation is presented in Table B.1.

Table 3.1: Summary of important quantities and notation in Chapter 3.

Parameter	Interpretation	Real-World Mapping
$a/\mathscr{A}$	item/s produced by source ρ	news article/s from publisher ρ
$\mathscr{H}^{ac}/\mathbb{H}^{a}$	cascade/s related to item a	retweet cascade/s for article a
$\vec{y}^{a}$	item-level features of a	headline embedding for article a
N^{a}	item popularity of a	overall tweet count for article a
$\vec{x}^{ac}$	cascade-level features of $\mathscr{H}^{ac}$	# followers of $\mathscr{H}^{ac}$ seed user
N^{ac}	cascade size of $\mathscr{H}^{ac}$	
$\mathscr{T}^{ac}$	intereevent-time distribution of $\mathscr{H}^{ac}$	
$(\alpha^{ac}, \mathbf{\Theta}^{ac})$	HP parameter set generating $\mathscr{H}^{ac}$	
$\tau^{ac}_{1/2}$	diffusion half-life of $\mathscr{H}^{ac}$	
$K_\alpha / K_{\mathbf{\Theta}}$	# of BMH-P/-K classes	
$z^{ac}_{\alpha,k} / z^{ac}_{\mathbf{\Theta},k}$	class k membership probability	
$\delta_{\alpha,k}/\delta_{\theta,k}$	baseline logit(α), log(θ) for class k	
$\delta_{z_{\alpha,k}}/\delta_{z_{\mathbf{\Theta},k}}$	baseline class k mem. probability	
$\vec{\gamma}_{\alpha,k}/\vec{\gamma}_{\theta,k}$	effect of $\vec{y}^{a}$ on class k center	
$\vec{\gamma}_{z_{\alpha,k}}/\vec{\gamma}_{z_{\mathbf{\Theta},k}}$	effect of $\vec{y}^{a}$ on class k mem. prob.	
$\vec{\beta}_{\alpha,k}/\vec{\beta}_{\theta,k}$	effect of $\vec{x}^{ac}$ on class k center	
$\vec{\beta}_{z_{\alpha,k}}/\vec{\beta}_{z_{\mathbf{\Theta},k}}$	effect of $\vec{x}^{ac}$ on class k mem. prob.	

3.4.1 BMH-P, the Popularity Submodel

The branching factor α^{ac} is modeled as the mixture random variable

$$\text{logit}(\alpha^{ac}) = \delta^{a}_{\alpha,k} + \vec{\gamma}_{\alpha,k} \cdot \vec{y}^{a}, \tag{3.3}$$

with membership probability $z^{ac}_{\alpha,k}$ $(k = 1, \ldots, K_\alpha)$,

$$z^{ac}_{\alpha,k} = \frac{\exp(\delta^{a}_{z_{\alpha,k}} + \vec{\beta}^{a}_{z_{\alpha,k}} \cdot \vec{x}^{ac} + \vec{\gamma}_{z_{\alpha,k}} \cdot \vec{y}^{a})}{\sum_{k'=1}^{K_\alpha} \exp(\delta^{a}_{z_{\alpha,k'}} + \vec{\beta}^{a}_{z_{\alpha,k'}} \cdot \vec{x}^{ac} + \vec{\gamma}_{z_{\alpha,k'}} \cdot \vec{y}^{a})}. \tag{3.4}$$

The intercept $\delta^{a}_{\alpha,k}$ in Eq. (3.3) sets the centering of logit(α^{ac}) for popularity class k. In Eq. (3.4), we designate $k = 1$ as the reference class (i.e. $\delta^{a}_{z_{\alpha,1}} = \vec{\beta}^{a}_{z_{\alpha,1}} = \vec{\gamma}_{z_{\alpha,1}} = 0$); parameters for $k > 1$ control deviation from class $k = 1$. The intercept $\delta^{a}_{z_{\alpha,k}}$ controls the baseline proportion of class k'. The influence of item features on logit(α^{ac}) and class k membership are estimated by $\vec{\gamma}_{\alpha,k}$ and $\vec{\gamma}_{z_{\alpha,k}}$, respectively, while the influence of cascade features on class k membership is estimated by $\vec{\beta}^{a}_{z_{\alpha,k}}$. Note that $\vec{\gamma}_{\alpha,k}, \vec{\gamma}_{z_{\alpha,k}}$ are shared across $\mathscr{A}$ while $\vec{\beta}^{a}_{z_{\alpha,k}}$ is estimated per a.

For brevity, we collect the parameter vector specific to item a as

$$\vec{p}^{a}_{\alpha} = [\delta^{a}_{\alpha,1}, \ldots, \delta^{a}_{\alpha,K_\alpha}, \delta^{a}_{z_{\alpha,2}}, \ldots, \delta^{a}_{z_{\alpha,K_\alpha}}, \vec{\beta}^{a}_{z_{\alpha,2}}, \ldots, \vec{\beta}^{a}_{z_{\alpha,K_\alpha}}]^{\top},$$

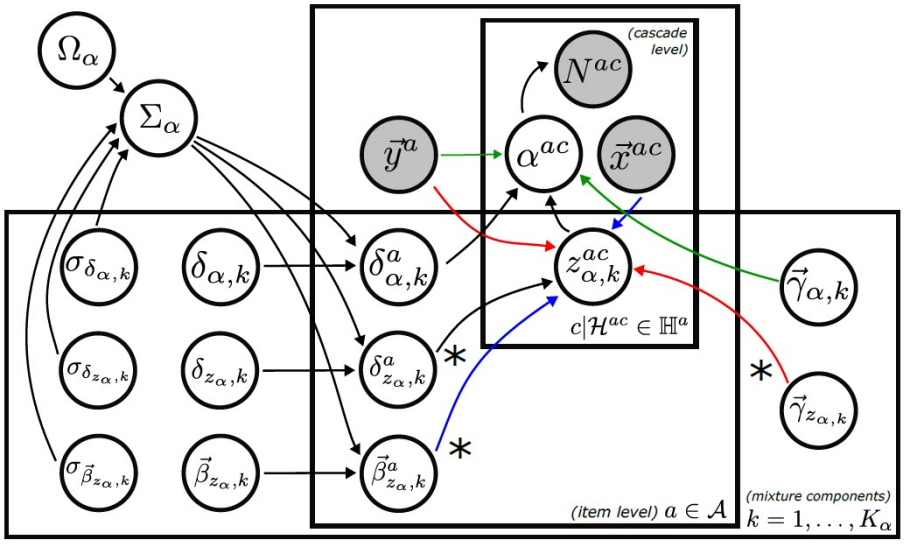


Figure 3.2: Plate diagram of the BMH-P model. Shaded nodes are observables while empty nodes are latent variables. Paired colored edges indicate source nodes appearing as a product in the target node. For instance, the green edges indicate that $\vec{\gamma}_{\alpha,k}$ and $\vec{y}^a$ appear as $\vec{\gamma}_{\alpha,k} \cdot \vec{y}^a$ in the expression for α^{ac} in Eq. (3.3). The same concept holds for the blue and red edges. Edges marked with * indicate dependence of the target node on the source node indexed with k and the entire set $\{1, \cdots, K_\alpha\}$. For instance, in Eq. (3.4) $z^{ac}_{\alpha,k}$ depends on $\vec{\beta}^a_{z_{\alpha,k}}$ (see the numerator) and $\vec{\beta}^a_{z_{\alpha,k'}}$ for $k' \in \{1, \cdots, K_\alpha\}$ (see the denominator).

with dimensionality

$$|\vec{p}^a_\alpha| = K_\alpha + (K_\alpha - 1) \cdot (1 + N_x). \tag{3.5}$$

We link item a with $\mathscr{A}$ by assuming that $\vec{p}^a_\alpha$ is drawn from a source-level multivariate normal (MVN) distribution with mean $\vec{p}_\alpha$ and covariance matrix Σ_α,

$$\vec{p}^a_\alpha \sim \text{MVN}\left(\vec{p}_\alpha, \Sigma_\alpha\right), \quad \Sigma_\alpha = \text{D}_\alpha \cdot \Omega_\alpha \cdot \text{D}_\alpha, \quad \text{D}_\alpha = \text{diag}(\sigma_{\vec{p}_\alpha}), \tag{3.6}$$

where Ω_α is a correlation matrix and $\sigma_{\vec{p}_\alpha}$ is a vector of standard deviations corresponding to $\vec{p}_\alpha$.

The plate diagram for the BMH-P model is shown in Fig. 3.2. Variable pairs that appear as a product term are colored green, red and blue in Eqs. (3.3) and (3.4), visualized in Fig. 3.2 as source nodes with green, red and blue edges.

Inference and Prediction. Let $\mathscr{P}_\alpha$ be the parameter set for the BMH-P model. From Eqs. (3.5) and (3.6) we see that $|\mathscr{P}_\alpha| = |\vec{p}^a_\alpha| \cdot (|\mathscr{A}| + 2 + |\vec{p}^a_\alpha|) + (2K_\alpha - 1) \cdot N_y$, where the first term $|\vec{p}^a_\alpha| \cdot |\mathscr{A}|$ accounts for the individual parameters for each item in $\mathscr{A}$, the second term $|\vec{p}^a_\alpha| \cdot 2$ accounts for $\vec{p}_\alpha$ and $\sigma_{\vec{p}_\alpha}$, the third term $|\vec{p}^a_\alpha|^2$ accounts for Ω_α, and the last term accounts for $\{\vec{\gamma}_{\alpha,k}, \vec{\gamma}_{z_{\alpha,k}}\}$. In Table 3.2, we compare the number of parameters of the BMH-P model

Table 3.2: Model complexity comparison. We compare the number of parameters in the BMH model against the baseline models: the publisher-level joint HP (see Appendix B.1.1) and the DMM [63]. Note that the other baselines, namely the EB [115] and CR models, are regression-based, and their parameter count depends on the specific regression model used.

	Popularity Submodel	Kernel Submodel
Joint HP [47]	1	2
DMM [63]	$2 \cdot K_\alpha \cdot \lvert\mathscr{A}\rvert$	$4 \cdot K_{\mathbf{\Theta}} \cdot \lvert\mathscr{A}\rvert$
BMH	$\lvert\vec{p}^{\,a}_{\alpha}\rvert \cdot (2 + \lvert\mathscr{A}\rvert + \lvert\vec{p}^{\,a}_{\alpha}\rvert) + (2K_\alpha - 1) \cdot N_y$	$2 \cdot K_{\mathbf{\Theta}} \cdot (4 + \lvert\mathscr{A}\rvert) + \lvert\vec{p}^{\,a}_{z_{\mathbf{\Theta}}}\rvert \cdot (2 + \lvert\mathscr{A}\rvert + \lvert\vec{p}^{\,a}_{z_{\mathbf{\Theta}}}\rvert) + (2K_{\mathbf{\Theta}} - 1) \cdot N_y$

with the baseline models in Section 3.5, namely the publisher-level joint Hawkes process (see Appendix B.1.1) and the DMM [63]. The added complexity of the BMH-P model stems from the additional parameters that model the source-level distribution and the influence of item- and cascade-level features.

From the set of cascade sizes $\{N_{ac}\}_{\mathscr{H}^{ac} \in \mathbb{H}^a, a \in \mathscr{A}}$, we estimate the posterior distribution $\mathbb{P}(\mathscr{P}_\alpha | \{N_{ac}\}_{ac}) \propto \exp(\mathscr{L}(\mathscr{P}_\alpha | \{N_{ac}\}_{ac})) \cdot \mathbb{P}(\mathscr{P}_\alpha)$, where $\mathbb{P}(\mathscr{P}_\alpha)$ is the prior for $\mathscr{P}_\alpha$ and $\mathscr{L}(\mathscr{P}_\alpha | \{N_{ac}\}_{ac})$ is the log-likelihood of $\mathscr{P}_\alpha$ given the cascade sizes (derived in Appendix B.2.2),

(3.7)

$$\mathscr{L}(\mathscr{P}_\alpha | \{N_{ac}\}_{ac}) = \log \mathbb{P}(\{N_{ac}\}_{ac} | \mathscr{P}_\alpha) = \sum_{\mathscr{H}^{ac} \in \mathbb{H}^a, a \in \mathscr{A}} \log \sum_{k=1}^{K_\alpha} z^{ac}_{\alpha,k} \cdot \mathbb{B}(N_{ac} | \text{inv-logit}(\delta^a_{\alpha,k} + \vec{\gamma}_{\alpha,k} \cdot \vec{y}^{\,a})),$$

where $\mathbb{B}(\cdot|\alpha)$ is the Borel distribution. Setting $n^{cascades} = |\{N_{ac}\}_{\mathscr{H}^{ac} \in \mathbb{H}^a, a \in \mathscr{A}}|$ as the total number of cascades in our dataset, the runtime complexity of evaluating $\mathscr{L}(\mathscr{P}_\alpha | \{N_{ac}\}_{ac})$ is $\mathscr{O}(n^{cascades} \cdot K_\alpha \cdot [N_x + N_y])$, where N_x and N_y are the dimensionalities of our cascade and item feature vectors, respectively.

Informative priors have to be set on $\{\delta_{\alpha,k}, \delta_{z_{\alpha,k}}\}$ to identify the K_α classes in the α parameter space. $\delta_{\alpha,k}$ and $\delta_{z_{\alpha,k}}$ identify the center and baseline proportion of the k^{th} class, respectively. Weakly informative priors are set for the other parameters in $\mathscr{P}_\alpha$. To sample the posterior distribution $\mathbb{P}(\mathscr{P}_\alpha | \{N_{ac}\}_{ac})$, we implement[1] the BMH-P model in Stan [15], which uses the No-U-Turn Sampler (NUTS), a Hamiltonial Monte Carlo technique well-suited for sampling from high-dimensional target distributions. The cost of obtaining an independent sample from a $|\mathscr{P}_\alpha|$-dimensional distribution is roughly $\mathscr{O}(|\mathscr{P}_\alpha|^{\frac{5}{4}})$ [49]. We use CmdStanPy [116] to run Stan code through Python.

Using the average cascade count for items in $\mathscr{A}$, denoted as $\hat{C}_\rho$, and the empirical distribution of the cascade feature vector $\vec{x}^{ac}$, denoted as $\hat{f}_\rho(x)$, the fitted BMH-P model can be used to estimate the cold-start popularity $\hat{N}^{a^*}$ of an out-of-sample item a^* with feature

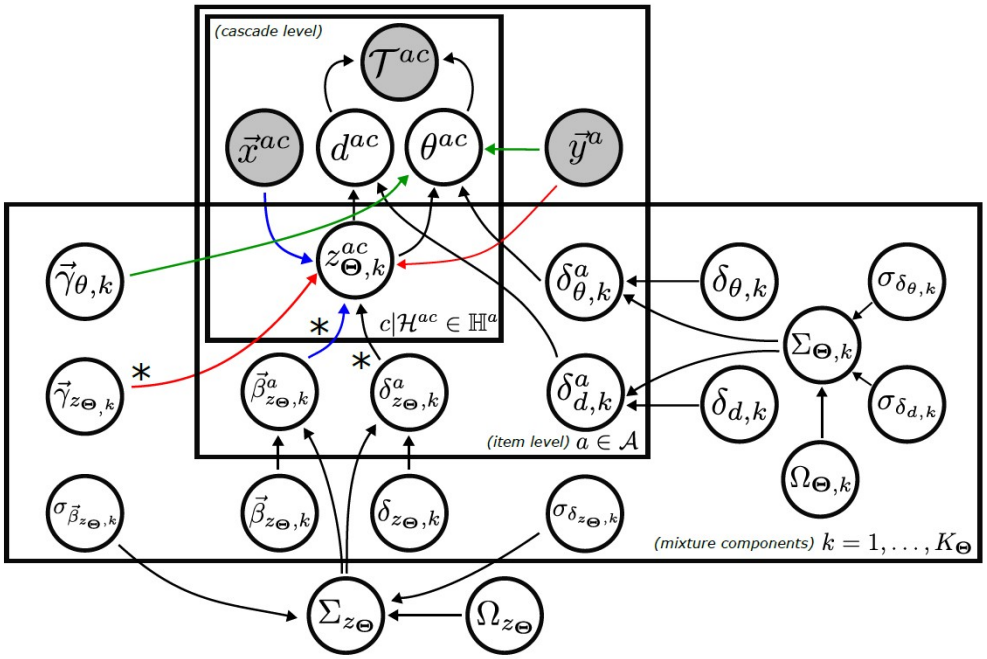


Figure 3.3: Plate diagram of the BMH-K model. Shaded nodes are observables while empty nodes are latent variables. Paired colored edges indicate source nodes appearing as a product in the target node. For instance, the green edges indicate that $\vec{\gamma}_{\theta,k}$ and $\vec{y}^a$ appear as the product $\vec{\gamma}_{\theta,k} \cdot \vec{y}^a$ in the expression for θ^{ac} in Eq. (3.9). The same concept holds for the blue and red edges. Edges marked with * indicate dependence of the target node on the source node indexed with k and the entire set $\{1,\cdots,K_\Theta\}$. For instance, in Eq. (3.10) $z^{ac}_{\Theta,k}$ depends on $\vec{\beta}^a_{z_{\Theta,k}}$ (see the numerator) and $\vec{\beta}^a_{z_{\Theta,k'}}$ for $k' \in \{1,\cdots,K_\Theta\}$ (see the denominator).

vector $\vec{y}^{a^*}$:

$$\hat{N}^{a^*} \approx \hat{C}_\rho \cdot \sum_{x=0}^{\infty} \sum_{k=1}^{K_\alpha} z^{a^*,c}_{\alpha,k} \cdot \left[1 + \exp\left(\delta^{a^*}_{\alpha,k} + \vec{\gamma}_{\alpha,k} \cdot \vec{y}^{a^*}\right)\right] \cdot \hat{f}_\rho(x), \tag{3.8}$$

where we assume that $\vec{x}^{ac} = x \in \mathbb{N}$ (see Appendix B.2.2).

3.4.2 BMH-K, the Kernel Submodel

Under the power-law, the kernel parameter set generating $\mathcal{H}^{ac}$ is $\mathbf{\Theta}^{ac} = [\theta^{ac}, d^{ac}]^\intercal$. We model $\mathbf{\Theta}^{ac}$ as a pair of mixture random variables taking the value

$$\log(\theta^{ac}) = \delta^a_{\theta,k} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^a, \quad \log(d^{ac}) = \delta^a_{d,k} \tag{3.9}$$

with probability $z^{ac}_{\mathbf{\Theta},k}$ $(k = 1,\ldots,K_\Theta)$, where

$$z^{ac}_{\mathbf{\Theta},k} = \frac{\exp(\delta^a_{z_{\mathbf{\Theta},k}} + \vec{\beta}^a_{z_{\mathbf{\Theta},k}} \cdot \vec{x}^{ac} + \vec{\gamma}_{z_{\mathbf{\Theta},k}} \cdot \vec{y}^a)}{\sum_{k'=1}^{K_\Theta} \exp(\delta^a_{z_{\mathbf{\Theta},k'}} + \vec{\beta}^a_{z_{\mathbf{\Theta},k'}} \cdot \vec{x}^{ac} + \vec{\gamma}_{z_{\mathbf{\Theta},k'}} \cdot \vec{y}^a)}. \tag{3.10}$$

In Eq. (3.10) we designate $k = 1$ as the reference class (i.e. $\delta^a_{z_{\mathbf{\Theta},1}} = \vec{\beta}^a_{z_{\mathbf{\Theta},1}} = \vec{\gamma}_{z_{\mathbf{\Theta},1}} = 0$).

Collect the parameter vectors for BMH-K as

$$\vec{p}^a_{\mathbf{\Theta},k} = [\delta^a_{\theta,k}, \delta^a_{d,k}]^\top,$$
$$\vec{p}^a_{z_\mathbf{\Theta}} = \left[\delta^a_{z_{\mathbf{\Theta},2}}, \ldots, \delta^a_{z_{\mathbf{\Theta},K_\mathbf{\Theta}}}, \vec{\beta}^a_{z_{\mathbf{\Theta},2}}, \ldots, \vec{\beta}^a_{z_{\mathbf{\Theta},K_\mathbf{\Theta}}}\right]^\top,$$

with dimensionalities

$$|\vec{p}^a_{\mathbf{\Theta},k}| = 2 \tag{3.11}$$

$$|\vec{p}^a_{z_\mathbf{\Theta}}| = (K_\alpha - 1) \cdot (1 + N_x), \tag{3.12}$$

The power law kernel having two parameters (i.e. θ^{ac}, d^{ac}) makes it challenging to estimate a joint source-level MVN distribution as we did for BMH-P. To simplify, we assume independence of $(\delta^a_{\theta,k}, \delta^a_{d,k})$ across classes. For each kernel class k, we assume $\vec{p}_{\mathbf{\Theta},k}$ is drawn from a source-level MVN distribution with mean $\vec{p}_{\mathbf{\Theta},k} = [\delta_{\theta,k}, \delta_{d,k}]^\top$ and covariance matrix $\Sigma_{\mathbf{\Theta},k}$. Lastly, we assume $\vec{p}^a_{z_\mathbf{\Theta}}$ is drawn from an MVN distribution with mean $\vec{p}_{z_\mathbf{\Theta}}$ and covariance matrix $\Sigma_{z_\mathbf{\Theta}}$.

$$\vec{p}^a_{\mathbf{\Theta},k} \sim \text{MVN}\left(\vec{p}_{\mathbf{\Theta},k}, \Sigma_{\mathbf{\Theta},k}\right), \quad \Sigma_{\mathbf{\Theta},k} = \text{D}_{\mathbf{\Theta},k} \cdot \Omega_{\mathbf{\Theta},k} \cdot \text{D}_{\mathbf{\Theta},k}, \quad \text{D}_{\mathbf{\Theta},k} = \text{diag}(\sigma_{\vec{p}_{\mathbf{\Theta},k}}) \tag{3.13}$$

$$\vec{p}^a_{z_\mathbf{\Theta}} \sim \text{MVN}\left(\vec{p}_{z_\mathbf{\Theta}}, \Sigma_{z_\mathbf{\Theta}}\right), \quad \Sigma_{z_\mathbf{\Theta}} = \text{D}_{z_\mathbf{\Theta}} \cdot \Omega_{z_\mathbf{\Theta}} \cdot \text{D}_{z_\mathbf{\Theta}}, \quad \text{D}_{z_\mathbf{\Theta}} = \text{diag}(\sigma_{z_\mathbf{\Theta}}), \tag{3.14}$$

where $\sigma_{\vec{p}_{\mathbf{\Theta},k}}, \sigma_{\vec{p}_{z_\mathbf{\Theta}}}$ are standard deviation vectors and $\Omega_{\mathbf{\Theta},k}, \Omega_{z_\mathbf{\Theta}}$ are correlation matrices.

The plate diagram for the BMH-K model is shown in Fig. 3.3. Variable pairs that appear as a product term are colored green, red and blue in Eqs. (3.9) and (3.10), visualized in Fig. 3.3 as source nodes with green, red and blue edges.

Inference and Prediction. Let $\mathscr{P}_\mathbf{\Theta}$ be the parameter set for the BMH-K model. From Eqs. (3.11) to (3.14) we see that $|\mathscr{P}_\mathbf{\Theta}| = K_\mathbf{\Theta} \cdot |\vec{p}^a_{\mathbf{\Theta},k}| \cdot (|\mathscr{A}| + 2 + |\vec{p}^a_{\mathbf{\Theta},k}|) + |\vec{p}^a_{z_\mathbf{\Theta}}| \cdot (|\mathscr{A}| + 2 + |\vec{p}^a_{z_\mathbf{\Theta}}|) + (2K_\mathbf{\Theta} - 1) \cdot N_y$, following similar reasoning for $|\mathscr{P}_\alpha|$. In Table 3.2 we compare the model complexity of the BMH-K model with the baseline models in Section 3.5.

From the interevent-time distributions $\{\mathscr{T}^{ac}\}_{ac}$, we estimate the posterior distribution $\mathbb{P}(\mathscr{P}_\mathbf{\Theta}|\mathscr{T}^{ac}) \propto \exp(\mathscr{L}(\mathscr{P}_\mathbf{\Theta}|\{\mathscr{T}^{ac}\}_{ac}) \cdot \mathbb{P}(\mathscr{P}_\mathbf{\Theta})$. The log-likelihood of $\mathscr{P}_\mathbf{\Theta}$ given $\{\mathscr{T}^{ac}\}_{ac}$ (derived in Appendix B.2.3) is given by

$$\mathscr{L}(\mathscr{P}_\mathbf{\Theta}|\{\mathscr{T}^{ac}\}_{ac}) = \sum_{\mathscr{H}^{ac} \in \mathbb{H}^a, a \in \mathscr{A}} \log \sum_{k=1}^{K_\mathbf{\Theta}} z^{ac}_{\mathbf{\Theta},k} \cdot f(\mathscr{H}^{ac}|e^{\delta^a_{\theta,k} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^a}, e^{\delta^a_{d,k}}), \tag{3.15}$$

where $f(\mathscr{H}|\theta, d) = \prod_{t_j \in \mathscr{H}} \sum_{t_z < t_j} g(t_j - t_z|\theta, d)$. Setting $n^{cascades} = |\{N_{ac}\}_{\mathscr{H}^{ac} \in \mathbb{H}^a, a \in \mathscr{A}}|$ and $n^{events} = \max_{\mathscr{H}^{ac} \in \mathbb{H}^a, a \in \mathscr{A}} N_{ac}$ as the total number of cascades and the size of the longest cascade, respectively, the worst-case runtime complexity of evaluating $\mathscr{L}(\mathscr{P}_\mathbf{\Theta}|\{\mathscr{T}^{ac}\}_{ac})$ is $\mathscr{O}(n^{cascades} \cdot K_\mathbf{\Theta} \cdot [N_x + (n^{events})^2 \cdot N_y])$.

Informative priors have to be set on $\{\delta_{\theta,k}, \delta_{d,k}, \delta_{z_{\mathbf{\Theta},k}}\}$ to identify the $K_{\mathbf{\Theta}}$ classes in the (θ, d) parameter space. $(\delta_{\theta,k}, \delta_{d,k})$ and $\delta_{z_{\mathbf{\Theta},k}}$ identify the center and baseline proportion of the k^{th} class, respectively. Weakly informative priors are set for the other parameters in $\mathscr{P}_{\mathbf{\Theta}}$. Similar to the BMH-P model, we implement[1] the BMH-K model in Stan and CmdStanPy to sample from the posterior distribution $\mathbb{P}(\mathscr{P}_{\mathbf{\Theta}}|\mathscr{T}^{ac})$.

The BMH-K model predicts the half-life $\hat{\tau}^{a^*}_{1/2}$ of an out-of-sample item a^* as (see Appendix B.2.3),

$$\hat{\tau}^{a^*}_{1/2} \approx \sum_{x=0}^{\infty} \sum_{k=1}^{K_{\mathbf{\Theta}}} z^{a^*,c}_{\mathbf{\Theta},k} \cdot e^{\delta^{a^*}_{d,k}} \cdot \left[2^{\exp(\delta^{a^*}_{\theta,k} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^{a^*})} - 1 \right] \cdot \hat{f}_{\rho}(x). \tag{3.16}$$

Table 3.3: Statistics of the predictive evaluation datasets.

	$CNIX-Fit$	$CNIX-Test$	$RNIX-Fit$	$RNIX-Test$
#publishers	41	41	28	28
#articles	72,009	40,506	2,682	18,116
#cascades	4,620,509	1,874,729	244,596	460,504
#tweets	42,546,067	18,235,185	1,573,909	5,139,967

3.5 Predictive Evaluation

In this section, we introduce two evaluation datasets (Section 3.5.1) and assess the BMH model's performance on two tasks: cold-start popularity prediction (Section 3.5.2) and temporal profile generalization performance (Section 3.5.3), i.e. evaluating the likelihood of the interevent distribution of future cascades.

3.5.1 Datasets

We use two datasets from [63] for predictive evaluation, consisting of collections of Twitter retweet cascades that link articles from online news sources. The Controversial News Index (*CNIX*) dataset consists of retweet cascades mentioning articles from 41 online news publishers known for controversial content, such as `https://www.breitbart.com/`. Conversely, the Reputable News Index (*RNIX*) follows the same structure as the $CNIX$ dataset but gathers cascades linked to articles from 28 reputable publishers, such as `https://www.news.com.au/`. The tweets for both datasets were collected by the QUT Digital Media Research Centre by retrospectively querying the Twitter search endpoint for URL mentions of the articles between June 30, 2017 and Dec 31, 2019. In Table 3.1 we link quantities in these datasets with variables in the BMH model.

Both *CNIX* and *RNIX* are temporally split into *Fit* (i.e. training) and *Test* (i.e. evaluation) datasets. The first contains tweets published from Jun 30, 2017 to Jan 1, 2019, while the second contains tweets from Feb 1, 2019 to Dec 31, 2019. A one-month gap between *Fit* and *Test* ensures that cascades in the training data are finished before the test period. Table 3.3 shows summary statistics.

We use the standardized 32-dimensional embedding of a's headline (i.e. PCA-reduced, *all-MiniLM-L6-v2* [99]) as our article feature vector $\vec{y}^a$, and the standardized log-follower count of the cascade's seed user as the cascade feature vector $\vec{x}^{ac}$.

3.5.2 Cold-Start Popularity Prediction

Our first task is evaluating the ability of the BMH-P model to predict cold-start popularity of unpublished content. With publisher ρ's trained BMH-P model, we predict the future popularity N^{a^*} of an out-of-sample article a^* with Eq. (3.8). To guide the selection of the number of mixture components K_α, we fit the BMM to each publisher in $RNIX$. We observe that the BMM-fitted $\{\alpha_i^a\}$ distribution is bimodal, corresponding to clusters of popular and unpopular cascades. See Appendix B.3.1 for full details. Using this result, we fit a BMH-P model for each publisher in $CNIX$ and $RNIX$ in Stan with the hyperparameter $K_\alpha = 2$. The full set of priors for the BMH-P model is listed in Appendix B.3.2. Note that we use a Laplace prior on $\vec{\gamma}_{\alpha,1}, \vec{\gamma}_{\alpha,2}, \vec{\gamma}_{z_{\alpha,2}}$ to impose regularization given the high dimensionality of the article feature vector ($|\vec{y}^a| = 32$) we consider.

To evaluate the predictive power of $\vec{x}^{ac}$ and $\vec{y}^a$, apart from the full model as developed in Section 3.4.1 (which we call $\alpha(\vec{y}^a) + z(\vec{x}^{ac}, \vec{y}^a)$) we fit three simpler variants of BMH-P: (1) $\alpha(\vec{y}^a) + z(\vec{y}^a)$, where we set $\vec{x}^{ac} = 0$ in Eq. (3.4); (2) $\alpha(\varnothing) + z(\vec{y}^a)$, where set $\vec{x}^{ac} = 0$ in Eq. (3.4) and $\vec{y}^{ac} = 0$ in Eq. (3.3); and (3) $\alpha(\varnothing) + z(\varnothing)$, where we set $\vec{x}^{ac} = 0$ in Eq. (3.4) and $\vec{y}^{ac} = 0$ in Eqs. (3.3) and (3.4).

We compare the performance of the BMH-P model to three approaches: (1) the DMM [63], (2) the empirical Bayes (EB) approach [115], and (3) feature-based cascade-size (CR) regression models (i.e. a neural network with one hidden layer of 100 nodes) built using scikit-learn [88]. For EB and CR, we fit two variants: one using only article features (i.e. EB(y) and CR(y)) and another using both cascade and article features (i.e. EB(x,y) and CR(x,y)). We report the Average Relative Error (ARE) over the set of articles in the $Test$ datasets. Let N^a and $\hat{N}^a$ be the actual and predicted popularity of article a, then $ARE(a) = \frac{|\hat{N}^a - N^a|}{N^a}$.

Results. In the top half of Table 3.4, we summarize cold-start popularity prediction performance of the model variants for $CNIX$ and $RNIX$. In both datasets the variants with only article-level features $\vec{y}^a$ and without the cascade-level features $\vec{x}^{ac}$ show minimal performance gain ($RNIX$) or even worse performance ($CNIX$) over the no-feature $\alpha(\varnothing) + z(\varnothing)$ model. The full model $\alpha(\vec{y}^a) + z(\vec{x}^{ac}, \vec{y}^a)$ significantly outperforms each simpler variant, highlighting the importance of the seed user's popularity as a predictor of final popularity [4].

We compare the performance of the best-performing BMH-P model with the benchmarks in the top row of Fig. 3.4(a) and Fig. 3.4(b). We can see that the BMH-P model outperforms each benchmark based on median performance. We note that in each task, the benchmarks that only have article features ($CR(y)$ and $EB(y)$) outperform the corresponding benchmarks that also include cascade features ($CR(x,y)$ and $EB(x,y)$). However, our

Table 3.4: Cold-start popularity prediction and model generalization results. We show the median ($25^{th}, 75^{th}$ quantiles) for BMH variants with different feature components removed. The best score across variants is in bold. Lower is better.

Popularity (ARE)	*CNIX*	*RNIX*
$\alpha(\varnothing) + z(\varnothing)$	0.707 (0.334, 1.513)	0.644 (0.335, 0.921)
$\alpha(\varnothing) + z(\vec{y}^a)$	0.708 (0.336, 1.497)	0.666 (0.339, 1.033)
$\alpha(\vec{y}^a) + z(\vec{y}^a)$	0.738 (0.370, 1.316)	0.643 (0.325, 0.953)
$\alpha(\vec{y}^a) + z(\vec{x}^{ac}, \vec{y}^a)$	**0.646** (0.313, 0.935)	**0.635** (0.342, 0.932)
Generalization (NLL)	*CNIX*	*RNIX*
$\theta(\varnothing) + z(\varnothing)$	-3.841 (-5.293, -2.717)	-2.564 (-3.231, -2.031)
$\theta(\varnothing) + z(\vec{y}^a)$	-3.782 (-4.873, -2.683)	-2.550 (-3.226, -1.988)
$\theta(\vec{y}^a) + z(\vec{y}^a)$	-3.649 (-4.816, -2.617)	**-2.689** (-3.492, -2.117)
$\theta(\vec{y}^a) + z(\vec{x}^{ac}, \vec{y}^a)$	**-4.013** (-5.766, -2.714)	-2.645 (-3.450, -2.063)

ablation results show that the best-performing BMH-P model includes both the cascade and article features. This implies that the added structure of the BMH-P model jointly leverages the article- and cascade-level information better than the benchmarks.

3.5.3 Temporal Profile Generalization Performance

Our second task is evaluating the performance of the BMH-K model in capturing the inter-arrival distribution of future cascades of unpublished articles. Given publisher ρ's trained BMH-K model, we calculate the log-likelihood $\mathscr{L}(\mathscr{P}_{\mathbf{\Theta}}|\{\mathscr{T}^{a^*c}\})$ of the inter-arrival distribution $\{\mathscr{T}^{a^*c}\}$ of an unpublished article a^*.

To guide the selection of the number of mixture components $K_{\mathbf{\Theta}}$, we fit the KMM to each publisher in *RNIX*. We observe that the KMM-fitted $\{\theta_i^a, d_I^a\}$ distribution is trimodal, corresponding to clusters of usual, slow- and fast-diffusing cascades cascades. See Appendix B.3.1 for the full details. Using this result, we fit a BMH-K model for each publisher in *CNIX* and *RNIX* in Stan with the hyperparameter $K_{\mathbf{\Theta}} = 3$. The full set of priors for the BMH-K model is listed in Appendix B.3.3. Note that we use a Laplace prior on $\vec{\gamma}_{\mathbf{\Theta},2}, \vec{\gamma}_{\mathbf{\Theta},3}, \vec{\gamma}_{z_{\mathbf{\Theta},2}}, \vec{\gamma}_{z_{\mathbf{\Theta},3}}$ to impose regularization given the high dimensionality of the article feature vector ($|\vec{y}^a| = 32$) we consider.

In addition to the full BMH-K model developed in Section 3.4.2 (which we call $\theta(\vec{y}^a) + z(\vec{x}^{ac}, \vec{y}^a)$) we fit three progressively simpler variants analogous to the ablation for the BMH-P model: $\theta(\vec{y}^a) + z(\vec{y}^a)$, $\theta(\varnothing) + z(\vec{y}^a)$, and $\theta(\varnothing) + z(\varnothing)$. To evaluate performance, we calculate the loglikelihood $\mathscr{L}(\mathscr{P}_{\mathbf{\Theta}}|\{\mathscr{T}^{ac}\})$ of inter-arrival times $\{\mathscr{T}^{ac}\}_{a \in \mathscr{A}}$ over articles in the *Test*

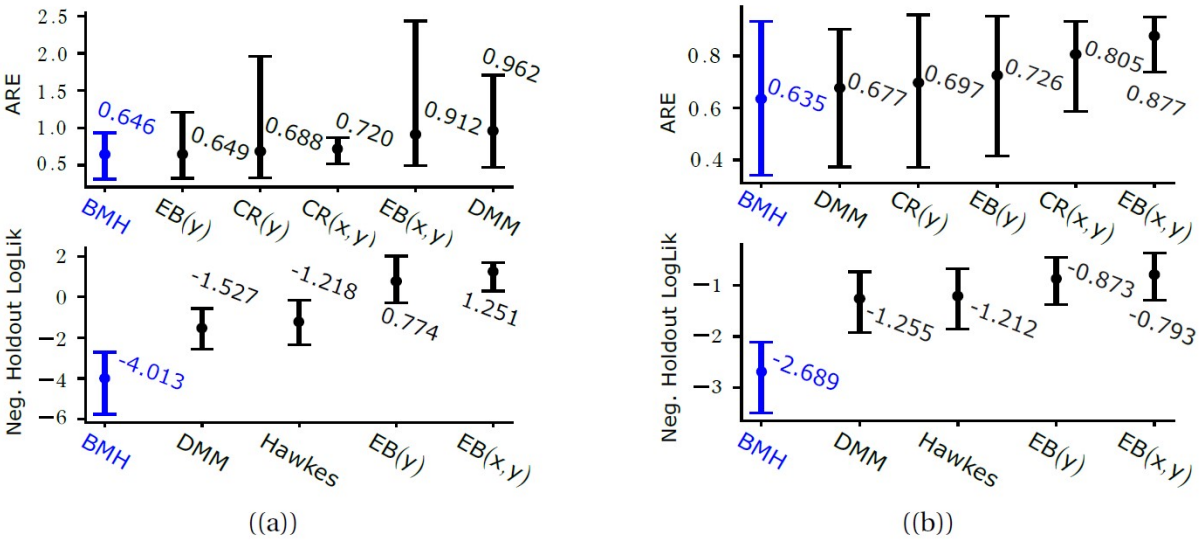


Figure 3.4: Predictive performance for (a) CNIX and (b) RNIX. The dots indicate the median and the error bars give the $25^{th}/75^{th}$ quantiles. We compare the BMH with the DMM [63], EB [115], cascade-size (CR) models, and the joint HP.

datasets. Since we are evaluating on likelihood, we use generative models as benchmarks: the DMM, EB(y), EB(x,y), and publisher-level joint HP (see Appendix B.1.1).

Results. In the lower half of Table 3.4, we see that for $CNIX$ each additional model component improves the log-likelihood, and that the full model $\alpha(\vec{y}^a) + z(\vec{x}^{ac}, \vec{y}^a)$ has the best performance. For $RNIX$ we observe that the variant without the seed user follower count, i.e., $\theta(\vec{y}^a) + z(\vec{y}^a)$, has the best performance. This finding suggests that in cascades related to reputable media articles, the seed user is not as influential in determining how long a cascade unfolds. In contrast, for controversial media articles, the seed user plays a significant role. We posit this is because the more fringe messaging in controversial media spreads through topical social groups (like conspiracy theorists, QAnon sympathizers and far-right supporters) [10, 54]. As a result, the first endorser is particularly important to legitimize content within the group. This is in contrast with the publicizing of traditional media articles on social media, where the most important factor is the publisher's reputation. In the bottom row of Fig. 3.4(a) and Fig. 3.4(b), we see that similar to the popularity prediction task, the BMH-K model outperforms all benchmarks on median performance for both datasets.

3.6 What-If? Headline Style Profiling

This section performs a counter-factual analysis to show that BMH successfully captures the relationship between headline writing style (i.e. neutral, clickbait or inflammatory) and content popularity and half-life. We run a 'What-If?' experiment, taking headlines of different writing styles and using the trained BMH models to infer how these headlines would perform under different publishers.

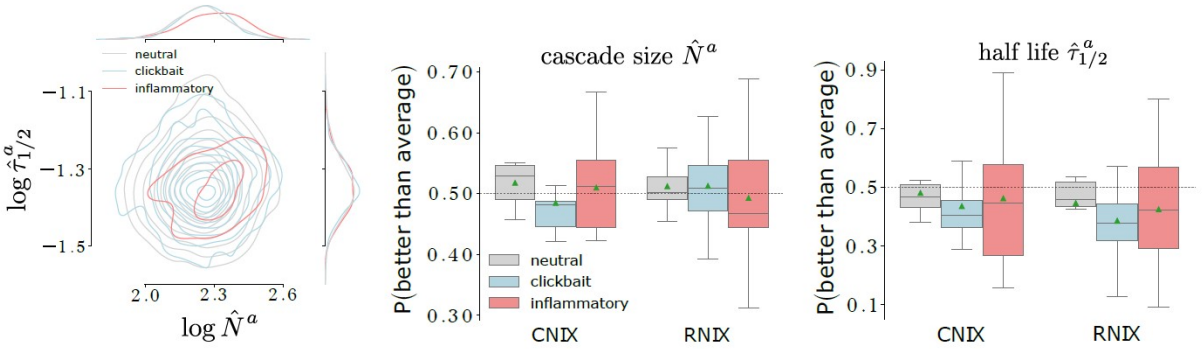
3.6.1 Dataset and Publisher Models

We utilize $HEADLINES$, a dataset of 1,227 article headlines collected using the news aggregation platform The Daily Edit [66]. The headlines come from four topics (Top Stories, Australia, Finance, and Climate Change) and six media sources (Daily Telegraph, Sky News, Sunday Morning Herald, The Guardian, news.com.au). Each headline was examined and sorted into one of three categories based on its informational and emotional content: neutral (N=727), clickbait (N=438) and inflammatory (N=62). Neutral headlines are detailed and appropriate, avoiding unnecessary information or emotive language, e.g. *'Australia's top military officer in the UK speaks ahead of Queen's funeral.'* Clickbait lacks informational and/or emotive quality without being misleading or inflammatory, often designed to attract attention, e.g. *'Bizarre sight spotted amid Aussie floods.'* Inflammatory headlines contain unnecessary details, often on serious topics, and may include inappropriate emotional language or details that reinforce negative stereotypes, e.g. *'Absolutely disgraceful': AFL fans blasted.'*

We use the trained publisher-level BMH models in Section 3.5 to predict performance of article headlines for each publisher in $CNIX$ and $RNIX$: expected cascade size (Eq. (3.8)) (setting $\hat{C}_\rho = 1$) and half-life (Eq. (3.16)). We use the variants that include only item features (i.e. $\alpha(\vec{y}^a) + z(\vec{y}^a)$ for BMH-P and $\theta(\vec{y}^a) + z(\vec{y}^a)$ for BMH-K) since cascade features are not available in this counter-factual setting.

3.6.2 Results

We apply the trained the BMH-P/-K models of each publisher ρ in $\{CNIX, RNIX\}$ to each of the 1,227 article headlines in $HEADLINES$ to infer the article's peformance if it were published under ρ. We summarize the predictions with a publisher-level performance heatmap ($\log \hat{N}^a$ vs. $\log \hat{\tau}^a_{1/2}$), where we differentiate the performance of neutral, clickbait and inflammatory headlines by aggregating the predictions of each headline style as contour



((a)) ((b)) ((c))

Figure 3.5: (a) Distribution of predicted half-life $\log \hat{\tau}^a_{1/2}$ vs. cascade size $\log \hat{N}^a$ for each article in *HEADLINES* using the `news.com.au` BMH model. (b and c) Probability that an article performs better than the publisher average, for each headline style across $CNIX$ and $RNIX$: (b) cascade size $\hat{N}^a$; (c) half life $\hat{\tau}^a_{1/2}$.

plots. Fig. 3.5(a) exemplifies the performance heatmap for the $RNIX$ publisher `news.com.au`. For this news source, we see that inflammatory headlines appear to have much higher popularity than neutral or clickbait headlines, while there is not much difference in half-life across headline styles. This is somewhat expected, as this publisher is known for its tabloid tendencies, focusing on "celebrity gossip, travel, lifestyle, sport, business, technology, money, and real estate", according to Media Bias Fact Check (MBFC) [74]. MBFC also rates its factual reporting as "MOSTLY FACTUAL" due to the occasional use of poor sources. We observe differences in the patterns for the headline styles across publishers (see Appendix B.4), implying that effective headlines for one publisher might not be effective for another, and that the BMH model learns these differences.

To summarise the differences across the categories $CNIX$ and $RNIX$, we compute the probability that each headline performs better – has a larger predicted cascade size or longer predicted half-life based on the BMH– than the publisher average based on the publisher's historical data. In Figs. 3.5(b) and 3.5(c) we show the distribution of these probabilities for each category and headline style.

We have three observations for the popularity probabilities in Fig. 3.5(b). First, we see that for $CNIX$, neutral headlines are effective (i.e. median better-than-average probability $> 50\%$). In contrast, clickbait headlines are ineffective (i.e. median better-than-average probability $< 50\%$). We link this result to the known inverse U-shaped relationship between clickbait volume and audience engagement [135], where too little or too much clickbait leads to suboptimal attention, suggesting the existence of a *sweet spot* for clickbait use. The over-prevalence of clickbait in controversial media outlets results in clickbait fatigue

among readers [70], leading to diminished effectiveness of clickbait headlines observed in Fig. 3.5(b).

Second and interestingly, we see that for $RNIX$ clickbait tends to perform better than neutral headlines. This is explained by Rony et al [105], who show that traditional news-oriented media consist of only 22% clickbait headlines while unreliable media consists of 39% clickbait based on a large sample of headlines. Since reputable media publishers have lower clickbait volume than controversial outlets, they are closer to the sweet spot for clickbait usage, retaining its effectiveness for drawing audience engagement. We do see a larger variance for clickbait for $RNIX$ compared to $CNIX$, suggesting that clickbait effectiveness is inconsistent and may not resonate universally, linking to the fact that clickbait strategies are only successful with certain audience segments [78].

Third, we observe large variance of performance for inflammatory headlines in both categories, indicative of the polarizing nature of this headline style. Inflammatory headlines tend to perform better in controversial outlets.

For the half-life probabilities (Fig. 3.5(c)), we see similar results, except that neutral headlines in both categories have higher half-life than clickbait, demonstrating the ephemerality of clickbait [69] irrespective of where it is published.

3.7 Headline Optimization with the BMH Model

In this section, we demonstrate the application of the trained BMH model to optimize online news article headlines for performance prior to publication. We propose a two-step *generate-then-evaluate* approach. The first step involves using text-generating AI (i.e. GPT) to produce rewrites for an input headline. The second step uses the trained BMH model to predict cold-start popularity and half life, which we use to rank the effectiveness of the rewrites. We demonstrate the effectiveness of this approach empirically through a Mechanical Turk (MTurk) [85] experiment where we applied the two-step procedure to previously published headlines and found that online respondents preferred the model-optimized versions.

Since the BMH-P and BMH-K publisher-level models in Section 3.5 are trained at the publisher level, it is necessary to specify a particular publisher for which we optimize performance. In this experiment, we selected `news.com.au` as the target publisher.

3.7.1 Generate-then-Evaluate Approach

We propose a two-step approach to optimize headlines. For a particular headline we aim to optimize, we generate a pool of 100 rewrites by prompting GPT3.5 with the following text: *"Imagine you are the editor of a big online news website. Come up with 100 creative variations of the following headline that you think will pull in an audience. Make sure not every output variation is in the Title: Subtitle format. Make sure the original headline and output variations have similar length."* We then use the trained BMH-P and BMH-K models to obtain the predicted popularity and half life of each rewrite and the original headline. These predictions are converted to relative improvement over the original headline scores and averaged. Let a^0 and a represent the original and AI-rewritten headlines, respectively. The performance score of a is computed as

$$s^a = \frac{1}{2} \cdot \left(\frac{\hat{N}^a - \hat{N}^{a^0}}{\hat{N}^{a^0}} + \frac{\hat{\tau}^a_{1/2} - \hat{\tau}^{a^0}_{1/2}}{\hat{\tau}^{a^0}_{1/2}} \right), \tag{3.17}$$

where $\hat{N}^a$ and $\hat{\tau}^a_{1/2}$ are given by Eq. (3.8) and Eq. (3.16), respectively.

Finally, we use $\{s^a\}$ as an effectiveness metric to rank $\{a\}$, the set of headline rewrites for the original headline a^0. We set the model-optimised version a^* as the top-performing rewrite, i.e. $a^* = \text{argmax}_a s^a$.

3.7.2 Seed Headline Selection

We apply the generate-then-evaluate approach to each headline a^0 in the $HEADLINES$ dataset introduced in Section 3.6.1 to obtain its model-optimised version a^* and the average percentage improvement s^{a^*}. We choose the 100 headline pairs (a^0, a^*) that have the highest average % improvement s^{a^*} for our experiment.

3.7.3 MTurk Experiment

We conduct an experiment on the crowdsourcing platform Mechanical Turk, asking online participants to choose between the original and model-optimised versions of the headlines. We randomly split the 100 headline pairs into a set of 10 unique questionnaires (called Human Intelligence Tasks (HITs) on MTurk), where each item in the questionnaire is a choice between the original and model-optimised headlines, the participant prompted with the task "*Select the news headline that grabs and holds your attention.*" We ran data collection for 13 iterations, each iteration consisting of 100 filled-in questionnaires (i.e. 10 uniquely completed responses, called assignments on MTurk, for each of the 10 HITs), except for the final two which had 200 filled-in questionnaires (i.e. 20 assignments each).

Participant Filtering. The MTurk platform allows selection of participants based on certain selection criteria, with the goal of weeding out bad workers. For this experiment, we select participants with approval rating exceeding 98%, with more than 5000 HITs completed ($> 5,000$) and located in majority-English-speaking countries, UK, USA, Canada and Australia.

Response Filtering. As an added safeguard against low quality responses, we implement three additional heuristic filtering steps on the workers. First, on each HIT we additionally include 5 attention check items, where instead of running the original headline through the generate-then-evaluate system we simply reorder the words text into an incoherent sentence. For instance, for the headline "*stars collide in battle for jillaroos no.1 jersey at the world cup*", we jumble the words into "*battle collide in stars for cup no.1 jersey at the jillaroos*". For a particular HIT response to be accepted as valid, he must have at least 4 of the 5 attention checks (80%) to be correct, i.e. selecting the original headline. Second, to filter out workers that quickly blitz through the HIT, we filter out HIT responses in the lower 10th percentile in terms of the time taken to complete the HIT. Lastly we also filter out responses that deviate greatly from the average response for each HIT. We compute the inter-worker agreement for each of the 10 HITs and only consider responses filled out by workers that have an above-average inter-worker agreement score.

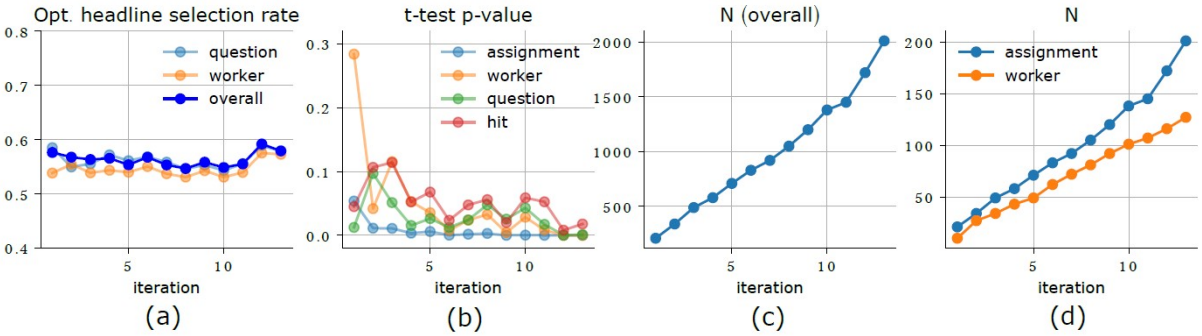


Figure 3.6: (a) Optimal headline selection rate aggregated across three levels (question, worker, overall) across the 13 iterations of the MTurk experiment. (b) One-tailed t-test p-value for the optimal headline selection rate being higher than random. (c) Number of questions answered across the 13 iterations. (d) Number of assignments and workers across the 13 iterations.

3.7.4 Results

Our MTurk sample consists of 13 iterations of data collection, with a total of 2010 binary responses between the original and model-optimised versions, 201 assignments of the 10 HITs, and 127 unique workers. In Fig. 3.6(c) and (d) we show the the overall size, total number of assignments and unique workers of the MTurk sample across the 13 iterations.

We use two metrics to summarise the results of the experiment. First, to quantify the preference for the model-optimised versions, we compute the optimised-headline selection rate, given as

$$\text{OPT} = \frac{\text{\# optimised headline selected}}{\text{\# headlines considered}}.$$

We aggregate the selection rates across three levels: question ($N = 100$), worker ($N = 127$) and overall ($N = 2010$). Note that the sample sizes for the worker and overall levels are at the final iteration. For instance, the worker-level selection rate is given by

$$\text{worker-level OPT} = \frac{1}{127}\sum_{k=1}^{127}\text{OPT(worker k)}.$$

Second, to quantify the significance of the model-optimised headline preference, we compute the p-value of the one-tailed t-test for the optimal headline selection rate being greater than random, i.e. OPT > 50%. We compute this value on four levels: assignment ($N = 201$), question ($N = 100$), worker ($N = 127$) and overall ($N = 2010$).

Fig. 3.6(a) and (b) show a summary of the results. In Fig. 3.6(a), we see that the optimised-headline selection rate consistently stays above the 50% line (i.e. no preference) across every iteration, ending up at an overall selection rate of 57.91% in the final iteration. Relative

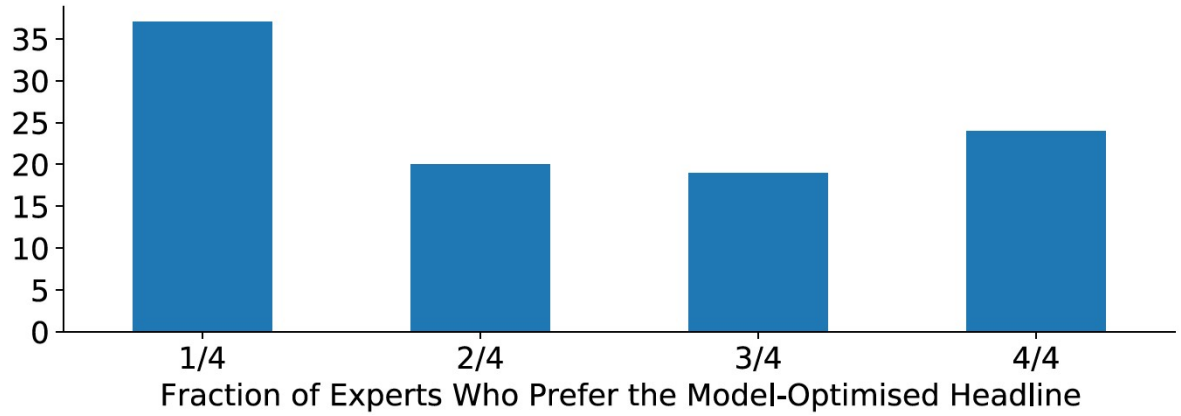


Figure 3.7: Fraction of journalists ($N = 4$) who prefer the model-optimised headline over the set of 100 (original, model-optimised) headline pairs in the MTurk experiment. The majority of the headlines in the dataset are preferred by 1 out of 4 journalists (=25% optimal-headline selection rate), indicative of journalists preferring the original headlines over the model-optimised ones.

to random choice, this is a 15.82% improvement, signifiying consistent preference of the model-optimised headlines over the original headlines. In Fig. 3.6(b) we show the via the p-value of the one-sided t-test for model-optimised headline preference. Note that the p-value generally decreases as we accrue a higher sample size; running the experiment for more iterations shows that the t-test p-values drop and stabilise, and by the last iteration, all p-values considered are less than 0.05, signifiying significance of the model-optimised headline preference at the 0.05 level for all aggregations considered.

3.7.5 Do Content Consumers and Producers Have Diverging Preferences?

In addition to the MTurk experiment, we asked four journalists to provide their preferences for the 100 headline pairs $\{(a^0, a^*)\}$ evaluated in the MTurk experiment. Fig. 3.7 shows a summary of the response at the question level. We see here a divergence in results from the MTurk experiment. Most of the headlines in the dataset are preferred by only 1 out of 4 journalists, indicating a 25% model-optimised headline selection rate. Indeed, by aggregating across all responses we see an overall optimal selection rate of 43%, indicative of the journalists' preference of the original versions of the headlines.

Our results show an apparent difference in preference between content producers (i.e. journalists) and consumers (i.e. MTurk workers). Content producers make assumptions on what *effective* content is for their intended audience, which might not hold in reality [94].

We can move from producer-optimised headlines (i.e. $\{a^0\}$) to consumer-optimised ones (i.e.$\{a^*\}$) by optimising directly on consumer feedback through the BMH model, which have a higher guarantee of content effectiveness on the consumer side.

3.8 Conclusion

This chapter proposes the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of Hawkes processes capable of learning the influence of item- and cascade-level features on spread dynamics. We demonstrate the applicability of the BMH model on two retweet cascade datasets that reference articles from reputable and controversial online news sources and show that the BMH model outperforms benchmark models in cold-start popularity prediction and temporal profile generalization performance. We apply the trained BMH models to a dataset of article headlines written in different headline styles and show differences in performance of headline styles across reputable and controversial outlets. Lastly, we demonstrate the effectiveness of the BMH model in an MTurk experiment, showing that online respondents have a significant preference for the BMH-optimised headlines over pre-optimised and previously published headlines.

Limitations and Future Work. We use the Hawkes process as the building block of the BMH model since it does not require the branching structure of diffusion cascades for inference. This choice is driven by data limitations on Twitter, where the branching structure of content shares is not accessible.

We propose two improvements. First, the BMH model assumes that α and $\boldsymbol{\Theta}$ depend only on cascade- and content-level features. We can allow α and $\boldsymbol{\Theta}$ to vary per event by including event-level features, which can be achieved by using the parametric Hawkes process [67] or Tweedie-Hawkes [68]. Second, the BMH model assumes a fixed number of popularity/kernel classes, obtained empirically by pre-fitting with the DMM. We can learn the manifest number of components directly from the data by assuming an infinite number of components via nonparametric Bayesian methods, such as using a Dirichlet Process prior [79].

In Section 3.7 we tested the capability of the BMH model as a cold-start headline optimization tool by combining it with generative AI (e.g. ChatGPT [84]). We aim to develop the capabilities further and apply it to other social media platforms such as Facebook, which introduce new data challenges.

AUTHOR DECLARATION

The following chapter contains content from the following publication.

Pio Calderon, Alexander Soen, and Marian-Andrei Rizoiu. “Linking Across Data Granularity: Fitting Multivariate Hawkes Processes to Partially Interval-Censored Data." *Under Review, IEEE Transactions on Computational Social Systems.*

Author Contributions: P.C. led the research for this study, managed data collection, and conducted the experiments. M.A.R. provided supervision throughout the project. P.C. and M.A.R. interpreted the results. P.C., A.S. and M.A.R. collaboratively developed the model and contributed to writing of the manuscript.

Production Note:
Signature removed prior to publication.

Pio Calderon

Production Note:
Signature removed prior to publication.

Alexander Soen

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

AUTHOR DECLARATION

The following chapter contains content from the following publication.

Quyu Kong, Pio Calderon, Rohit Ram, Olga Boichak, and Marian-Andrei Rizoiu. “Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems." *Proceedings of the ACM Web Conference 2023*, pp. 1813-1821. 2023.

Author Contributions: Q.K. led the research for this study, managed data collection, and conducted the experiments. M.A.R. provided supervision throughout the project. P.C. and R.R. contributed to model baselines. Q.K. and M.A.R. collaboratively developed the model. O.B. provided domain knowledge expertise and qualitative interpretation. Q.K., O.B. and M.A.R. interpreted the results and contributed to writing of the manuscript.

Production Note:
Signature removed prior to publication.

Quyu Kong

Production Note:
Signature removed prior to publication.

Pio Calderon

Production Note:
Signature removed prior to publication.

Rohit Ram

Production Note:
Signature removed prior to publication.

Olga Boichak

Production Note:
Signature removed prior to publication.

Marian-Andrei Rizoiu

CHAPTER 4

LINKING ACROSS DATA GRANULARITY: FITTING MULTIVARIATE HAWKES PROCESSES TO PARTIALLY INTERVAL-CENSORED DATA

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 which 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. Firstly, 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 popularity estimation algorithm Hawkes Intensity Process (HIP) [103]. 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. Lastly, 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 between COVID-19 cases and news reporting.

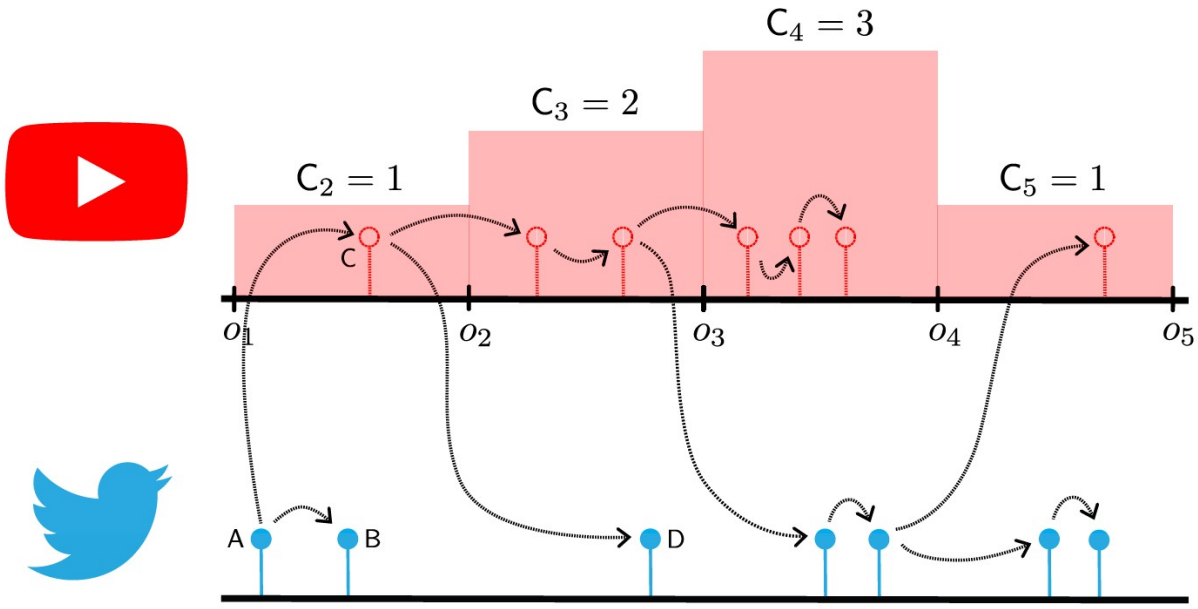


Figure 4.1: **Example of multi-platform interaction** between view events on YouTube (red lollipops) and tweets on Twitter (blue lollipops). The data is partially interval-censored, as YouTube does not expose individual views, but only the view counts C_i's over the predefined intervals $[o_i, o_{i+1})$ (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.

4.1 Introduction

The Hawkes process, introduced by [47], is a temporal point process that exhibits the *self-exciting* property, *i.e.*, the occurrence of one event increases the likelihood of future events. 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 [101]. The multivariate Hawkes process (MHP) [47] 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. 4.1 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 [14], we usually observe the aggregated daily counts of reported cases instead of detailed case information. Another reason is space limitations — for network traffic data [111], 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. 4.1, where the individual views are interval-censored, and we only observe aggregated daily counts.

This chapter 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 is presented as event times, the MHP can be fitted using MLE [27]. However, if the data is partially or completely interval-censored, MLE cannot fit the MHP process parameters because it lacks the independent increments property [102]. Given interval-censored counts, one could approach fitting the Hawkes process naively by sampling event times uniformly over the intervals [119]. 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, [102] 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 [60, 76, 133, 136], YouTube [26, 103], Reddit [64] — 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) [103] 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 [130], the 2010 cholera outbreak in Haiti [22], and the early spread of COVID in 2020 in various provinces in China [131]. 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. [130, 131] 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. [22] 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)[1], a novel multivariate temporal point process that operates on partially interval-censored data. We answer the first question in Section 4.4.1, 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 multivariate Hawkes process through parameter equivalence.

We address the second question in Section 4.7 by using PCMHP to predict the popularity of YouTube videos on both YouTube and Twitter. We demonstrate that PCMHP consistently

[1] Implementation available at `https://github.com/behavioral-ds/pmbp_implementation`.

outperforms the related HIP method [103], 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 4.8 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 (UK, 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 (US, Italy, Sweden, India, and the Philippines).

In Section 4.9, we briefly describe an alternative notion of the partially interval-censored setup which we tackled in another work [62] where I was a coauthor. Here, we extend the deep-learning-based Transformer Hawkes (TH) [137] to the partially interval-censored setup by introducing the Interval-Censored Transformer Hawkes (IC-TH) model.

4.2 Related Work

A significant portion of 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 [58] 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, [59] 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 [19], solving the issue in [59]. 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 [111] and [112] 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 [59] in synthetic experiments. Another sample-based approach, the recursive identification with sample correction (RISC) algorithm, was introduced in [106], 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 [6]. A Hawkes process incorporating spatio-temporal covariates was estimated using an EM algorithm in [20], while the least-squares approach was utilized in [107] 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 [14] 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 [9, 17, 114]. 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 [30]. This data type is prevalent in health and clinical research [72], 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 [62, 102], where event times are inaccessible and we instead observe event counts over predefined time intervals, as exemplified by Youtube views in Fig. 4.1. 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.

4.3 Preliminaries

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 $\mathbf{N}(t)$ represent the number of events that occurred up until time t and $\mathscr{H}_{t^-}^j$ be the set of all events that occur in dimension j up until t, for $j \in \{1,\dots,d\}$. We further denote the union of all history dimensions as $\mathscr{H}_{t^-} := \bigcup_{j=1}^d \mathscr{H}_{t^-}^j$. The d-dimensional conditional intensity function $\boldsymbol{\lambda}(t|\mathscr{H}_{t^-})$ is defined as

$$\lambda^j(t \mid \mathscr{H}_{t^-}) = \lim_{h\to 0^+} \frac{1}{h}\mathbb{P}\left\{N^j(t+h) - N^j(t) = 1 \mid \mathscr{H}_{t^-}\right\},$$

which gives the instantaneous probability of a dimension j event occurring in the increment $[t, t+dt)$, conditioned on all events that happen before t. For brevity and whenever it is clear from context, we drop the explicit conditioning on the history $\mathscr{H}_{t^-}$ and write $\boldsymbol{\lambda}(t) := \boldsymbol{\lambda}(t|\mathscr{H}_{t^-})$.

4.3.1 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 multivariate Hawkes process (MHP) is a point process where each dimension tracks the dynamics of each event type. In addition to being self-exciting, 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

$$\boldsymbol{\lambda}(t) := \boldsymbol{\mu}(t) + \sum_{j=1}^{d} \sum_{t_k^j \in \mathscr{H}_{t^-}^j} \boldsymbol{\varphi}^j(t - t_k^j), \tag{4.1}$$

where $\boldsymbol{\mu}(t)$ is the (deterministic) background intensity, a d-dimensional non-negative vector for each t controlling the arrival of external events into the system. The matrix $\boldsymbol{\varphi}(t)$ is called the Hawkes kernel, a $d \times d$ matrix of functions that characterizes the self- and cross-excitation across the event types representing the d dimensions. Let $\boldsymbol{\varphi}^j(t)$ represent 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 $\boldsymbol{\lambda}(t)$ defined in Eq. (4.1) is a stochastic function dependent on the history of any particular realization $\mathscr{H}_{t^-}$. Given a fixed $\mathscr{H}_{t^-}$, the intensity before or

equal to t is deterministically calculated using Eq. (4.1). 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 denote the index set $\{1,\dots,d\}$. If we assume $\varphi^{ij}(t) = \alpha^{ij} f^{ij}(t)$, $\alpha^{ij} \geq 0$, $f^{ij}(t) \geq 0$, and $\int_0^\infty f^{ij}(t)\mathrm{d}t = 1$ for $(i,j) \in D \times D$. We call α^{ij} the branching factor from j to i and the matrix $\boldsymbol{\alpha} = (\alpha^{ij}) \in (\mathbb{R}^+)^{d\times d}$ the branching matrix. The branching factor α^{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 [26]. 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 [76, 102, 109], which takes the form $\varphi^{ij}(t) = \alpha^{ij}\theta^{ij}\exp(-\theta^{ij}t)$, where θ^{ij} controls the rate of influence decay from j to i. More prerequisite details on the MHP are provided in Appendix C.1.

4.3.2 Mean Behavior Poisson Process

Consider a univariate Hawkes process with conditional intensity $\lambda(t)$. The Mean Behavior Poisson (MBP) process introduced by [102] is the inhomogeneous Poisson process with conditional intensity

$$\xi(t) := \mathbb{E}_{\mathcal{H}_{t^-}}[\lambda(t)]. \tag{4.2}$$

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 [102] that $\xi(t)$ follows the self-consistent equation

$$\xi(t) = \mu(t) + (\varphi * \xi)(t), \tag{4.3}$$

where $*$ denotes convolution. Furthermore, the mapping $\mu(t) \mapsto \xi(t)$ in Eq. (4.3) defines a linear time-invariant (LTI) system [90], 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 [102] to be

$$\xi(t) = \left(\delta(t) + \sum_{n=1}^{\infty} \varphi^{\otimes n}(t)\right) * \mu(t), \tag{4.4}$$

where $\otimes n$ 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 Appendix C.1.

4.3.3 Hawkes Intensity Process

The Hawkes intensity process (HIP), introduced in [103], 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 $\mathscr{P}[0,T) = \bigcup_{k=1}^{m}[o_{k-1}, o_k)$, where $o_0 = 0$ and $o_m = T$, and the associated view counts $\{\mathsf{C}_k\}_{k=1}^{m}$, the HIP model $\hat{\xi}[\cdot;\boldsymbol{\Theta}]$ is fitted by finding the parameter set $\boldsymbol{\Theta}$ that minimizes the sum of squares error $\sum_{k=1}^{m}\left(\mathsf{C}_k - \hat{\xi}[o_k;\boldsymbol{\Theta}]\right)^2$ of the following recursive formula for $\hat{\xi}[o_k;\boldsymbol{\Theta}]$,

$$\hat{\xi}[o_k;\boldsymbol{\Theta}] = \mu[o_k] + \sum_{s=0}^{k-1}\varphi(o_k - o_s;\boldsymbol{\Theta})\cdot\hat{\xi}[o_s;\boldsymbol{\Theta}].$$

The use of brackets emphasizes that the quantities are discretized over a partition of time. It was shown in [102, 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.*, $\hat{\xi}[o_k;\boldsymbol{\Theta}]\cdot(o_k - o_{k-1})$ is approximated as $\hat{\xi}[o_k;\boldsymbol{\Theta}]$.

Table 4.1: Important notation used in Chapter 4.

Symbol	Meaning
$\boldsymbol{\alpha}$	Hawkes branching matrix
$\rho(\boldsymbol{\alpha})$	spectral radius of $\boldsymbol{\alpha}$
$f^{ij}(t)$	exponential kernel from dimension j to i
$\boldsymbol{\theta}$	exponential kernel decay parameter matrix
$\boldsymbol{\mu}(t)$	deterministic background intensity
$\boldsymbol{\varphi}(t)$	Hawkes kernel, where $\varphi^{ij}(t) = \alpha^{ij} f^{ij}(t)$
$\boldsymbol{\lambda}(t)$	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^c	set of Hawkes dimensions for PCMHP (d, e)
$\mathscr{H}^j_{t^-}$	event sequence history on dimension $j \in D$
$\mathscr{H}^A_{t^-}$	union of event sequence histories on $A \subset D$
$\boldsymbol{\xi}_E(t)$	conditional intensity for PCMHP (d, e)
$\boldsymbol{\Xi}_E(t)$	compensator for PCMHP (d, e)
$\boldsymbol{\Theta}$	parameter set for PCMHP (d, e)
T	terminal time
$\mathscr{L}(\boldsymbol{\Theta}; T)$	log-likelihood function for PCMHP (d, e)
$\mathscr{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}; T)$	gradient of log-likelihood function for PCMHP (d, e)
Δ^P	time axis partition length for numerical convolution
γ^h	convergence threshold for infinite sum truncation

4.4 Partially Censored Multivariate Hawkes Process

For convenience, the list of notation that we use in this work is provided in Table 4.1.

4.4.1 Formulation

We define the Partially Censored Multivariate Hawkes Process PCMHP(d,e) with intensity $\boldsymbol{\xi}_E(t)$ as follows. A key idea of the PCMHP(d,e) is to fix the history of different dimensions. As such we denote the history union over a subset of dimensions $A \subset \{1 \dots d\}$ as $\mathscr{H}_{t^-}^A := \bigcup_{j \in A} \mathscr{H}_{t^-}^j$.

Definition 4.1. Consider a d-dimensional Hawkes process with conditional intensity $\boldsymbol{\lambda}(t)$ as defined in Eq. (4.1). Given a nonnegative integer $e \leq d$ and the index sets $D := \{1,\dots,d\}$, $E := \{1,\dots,e\}$ and $E^c := \{e+1,\dots,d\}$, the Partially Censored Multivariate Hawkes Process PCMHP(d,e) is the temporal point process whose conditional intensity $\boldsymbol{\xi}_E(t)$ is the expectation of $\boldsymbol{\lambda}(t)$ conditioned on the set of event histories $\mathscr{H}_{t^-}^{E^c}$ in the E^c dimensions and averaged over the set of event histories $\mathscr{H}_{t^-}^{E}$ in the E dimensions. That is,

$$\boldsymbol{\xi}_E(t) := \boldsymbol{\xi}\left(t \mid \mathscr{H}_{t^-}^{E^c}\right) = \mathbb{E}_{\mathscr{H}_{t^-}^{E}}\left[\boldsymbol{\lambda}(t) \mid \mathscr{H}_{t^-}^{E^c}\right]. \tag{4.5}$$

The PCMHP(d,e) intensity is a stochastic function due to its dependence on the current realization of $\mathscr{H}_{t^-}^{E^c}$; on the E dimensions we take the expectation over all possible realizations of $\mathscr{H}_{t^-}^{E}$, similar to the MBP intensity in Eq. (4.2).

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. 4.1 for instance E would be YouTube views and E^c the set of tweets.

Is the PCMHP (d,e) Poisson? Due to its dependence on the history of the E^c dimensions, the PCMHP (d,e) is not a Poisson process. From Eq. (4.5), the PCMHP(d,e) can be interpreted as a collection of d processes, where $\xi_E^j(t)$ follows a Hawkes process for $j \in E^c$ and follows an inhomogeneous Poisson process (conditional on the event history of the E^c dimensions) for $j \in E$. In fact, the PCMHP(d,e) generalizes both the MHP (by setting $e=0$) and the MBP process (by setting $e=d$).

Convolutional Formula. Consider the kernel

$$\boldsymbol{\varphi}(t) = \begin{bmatrix} \boldsymbol{\varphi}^1(t) & \dots & \boldsymbol{\varphi}^e(t) & \boldsymbol{\varphi}^{e+1}(t) & \dots & \boldsymbol{\varphi}^d(t) \end{bmatrix},$$

setting $\boldsymbol{\varphi}^j(t)$ to be the j^{th} column of $\boldsymbol{\varphi}(t)$. Similarly, let

$$\boldsymbol{\varphi}_E(t) = \begin{bmatrix} \boldsymbol{\varphi}^1(t) & \dots & \boldsymbol{\varphi}^e(t) & 0 & \dots & 0 \end{bmatrix}$$

and

$$\boldsymbol{\varphi}_{E^c}(t) = \begin{bmatrix} 0 & \dots & 0 & \boldsymbol{\varphi}^{e+1}(t) & \dots & \boldsymbol{\varphi}^d(t) \end{bmatrix}.$$

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

Theorem 4.1. *The conditional intensity* $\boldsymbol{\xi}_E(t)$ *of the* PCMHP *(*d, e*) process is given by*

$$\boldsymbol{\xi}_E(t) = \left[\boldsymbol{\delta}(t) + \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(t)\right] * \left[\boldsymbol{\mu}(t) + \sum_{j \in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}_{E^c}^{j}(t - t_k^j)\right]. \tag{4.6}$$

In general, $\boldsymbol{\xi}_E(t)$ does not admit a closed form solution because of the complexity of the infinite convolution sum of $\boldsymbol{\varphi}_E(t)$ (an interpretation of which is provided in Appendix C.2). However, in the special case of PCMHP$(2,1)$ with the exponential kernel, a closed-form solution for $\boldsymbol{\xi}_E(t)$ exists, derived in Appendix C.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 ρ (*i.e.* magnitude of the largest eigenvalue) of the branching matrix is less than one, *i.e.* $\rho(\boldsymbol{\alpha}) < 1$ [81]. Here we introduce the regularity conditions applicable for the PCMHP(d, e).

Consider the following submatrices of $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{EE} = (\alpha^{ij})_{(i,j)\in E\times E}, \qquad \boldsymbol{\alpha}^{EE^c} = (\alpha^{ij})_{(i,j)\in E\times E^c},$$
$$\boldsymbol{\alpha}^{E^cE} = (\alpha^{ij})_{(i,j)\in E^c\times E}, \qquad \boldsymbol{\alpha}^{E^cE^c} = (\alpha^{ij})_{(i,j)\in E^c\times E^c}.$$

The following are three conditions which ensure subcriticality of PCMHP(d, e).

Theorem 4.2. *The* PCMHP(d, e) *with branching matrix* $\boldsymbol{\alpha}$ *is subcritical if the following conditions hold.*

$$\rho(\boldsymbol{\alpha}^{EE}) < 1 \tag{4.7}$$

$$\rho(\boldsymbol{\alpha}^{E^cE^c}) < 1 \tag{4.8}$$

$$\rho(\boldsymbol{\alpha}^{E^cE}(I - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{EE^c}) < 1. \tag{4.9}$$

We note that the regularity conditions for PCMHP(d, e) in Theorem 4.2 cover the MHP and the MBP as special cases.

Proofs of Theorem 4.1 and Theorem 4.2 and a discussion on the nonlinear extension of the PCMHP are provided in Appendix C.4.

4.4.2 Inference

We now consider the problem of estimating the PCMHP parameter set $\boldsymbol{\Theta}$ 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 $\boldsymbol{\mu}(t) = \boldsymbol{\nu}$, the PCMHP$(d,e)$ parameter set to be estimated is given by $\boldsymbol{\Theta} = \{\boldsymbol{\nu}, \boldsymbol{\alpha}, \boldsymbol{\theta}\}$ with size $|\boldsymbol{\Theta}| = d + 2 \cdot d^2$.

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 $o_0^j < o_1^j < \ldots < o_{n^j}^j$ such that for o_k^j where $k \geq 1$, we observe the volume C_k^j of dimension j events that occurred during the interval $[o_{k-1}^j, o_k^j)$. Meanwhile, for $j \in Q^c$, we observe event sequences $\mathscr{H}_{T^-}^j = \{t_1^j < t_2^j < \ldots < t_{n^j}^j\}$. In practice, the observation partition $\bigcup_{j\in Q}\bigcup_{k=1}^{n^j}[o_{k-1}^j, o_k^j)$ is not a model hyperparameter but is determined by real-world dataset availability constraints. For instance, in Fig. 4.1 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 PCMHP(d,e) process to the above-defined data using the log-likelihood function derived below. The proof is available in Appendix C.5.

Theorem 4.3. *Given event times $\mathscr{H}_{T^-}^{Q^c}$, event volumes $\bigcup_{j\in Q}\{\mathsf{C}_k^j\}_{k=1}^{n^j}$, and a* PCMHP$(d,e)$ *model such that $E \supseteq Q$, the negative log-likelihood of parameter set $\boldsymbol{\Theta}$ can be written as*

$$\mathscr{L}(\boldsymbol{\Theta}; T) = \sum_{j\in Q} \mathscr{L}_{\text{IC-LL}}^j(\boldsymbol{\Theta}; T) + \sum_{j\in Q^c} \mathscr{L}_{\text{PP-LL}}^j(\boldsymbol{\Theta}; T), \tag{4.10}$$

where

$$\mathscr{L}_{\text{IC-LL}}^j(\boldsymbol{\Theta}; T) = \sum_{i=1}^{n^j} \left[\Xi_E^j(o_{i-1}^j, o_i^j; \boldsymbol{\Theta}) - \mathsf{C}_i^j \log \Xi_E^j(o_{i-1}^j, o_i^j; \boldsymbol{\Theta})\right], \tag{4.11}$$

$$\mathscr{L}_{\text{PP-LL}}^j(\boldsymbol{\Theta}; T) = -\sum_{t_k^j \in \mathscr{H}_{T^-}^j} \log \xi_E^j(t_k^j; \boldsymbol{\Theta}) + \Xi_E^j(T; \boldsymbol{\Theta}), \tag{4.12}$$

and $\boldsymbol{\Xi}_E(t)$ represents the compensator, i.e.*, the intensity $\boldsymbol{\xi}_E(t)$ integrated over 0 to t.*

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 $\mathscr{L}_{\text{IC-LL}}^j(\boldsymbol{\Theta}; T)$; if $j \in Q^c$ (event-times) then $\mathscr{L}_{\text{PP-LL}}^j(\boldsymbol{\Theta}; T)$ 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 4.6.

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, $\mathrm{n}^{E^c} = \sum_{j \in E^c} |\mathscr{H}_{T^-}^j|$, $\mathrm{n}^{Q^c} = \sum_{j \in Q^c} |\mathscr{H}_{T^-}^j|$, $\mathrm{n}^E = \sum_{j \in E} n^j$, and $\mathrm{n}^Q = \sum_{j \in Q} n^j$. Let C denote a constant independent of the dimension of the PCMHP and the data. Evaluating $\mathscr{L}(\boldsymbol{\Theta}; T)$ has a runtime complexity of $\mathscr{O}\left((C + \mathrm{n}^{E^c}) \cdot (\mathrm{n}^Q + \mathrm{n}^{Q^c})\right)$ (see Appendix C.5 for more details). In the case $E = Q = \emptyset$, the runtime complexity reduces to $\mathscr{O}((\mathrm{n}^{E^c})^2)$, consistent with the MHP. If $E = Q = D$, runtime complexity reduces to $\mathscr{O}(\mathrm{n}^E)$, consistent with the MBP (i.e. Poisson) process.

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

Gradient-based optimization tools – including IPOPT [123] that we use in our experiments in Section 4.7 and Section 4.8 — usually require the gradient. To approximate $\mathscr{L}(\boldsymbol{\Theta}; T)$ and its gradient $\mathscr{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}; T)$, we propose a numerical scheme in Appendix C.10 and Appendix C.11. We show in Appendix C.10 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.* $\mathscr{O}\left(\mathrm{n}^E \cdot \lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e\right)$. 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^c}, *i.e.* $\mathscr{O}([\mathrm{n}^E + \mathrm{n}^{E^c} + \lceil \frac{T}{\Delta^P} \rceil] \cdot \mathrm{n}^{E^c} \cdot d)$. In both cases, the number of partition intervals $\lceil \frac{T}{\Delta^P} \rceil$ only appears linearly. For partially interval-censored datasets with high frequency

data, the number of observed event times n^{E^c} is the most important determinant of runtime complexity given that it appears quadratically, while the number of observation intervals n^E and the number of partition intervals $\lceil \frac{T}{\Delta^P} \rceil$ only appear linearly.

Lastly, for the purpose of sampling from the PCMHP (d, e), we propose a modification of the thinning algorithm [82] in Appendix C.12.

4.5 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 within 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 pre-existing 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* (e.g. the MHP), 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 [34] or the discrete-time Hawkes process [14]) assume evenly spaced data [32]. If the censor intervals within or across dimensions do not line up, we would need to perform further data alteration, such as interpolation [98], 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 $\mathcal{O}((\mathrm{n}^{E^c})^2)$. 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 — has the potential to produce spurious event interactions across dimensions, particularly for wide censor intervals.

4.6 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 $\boldsymbol{\mu}(t) = \boldsymbol{\nu}$. 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 Appendix C.14.

Throughout this section, we refer to $\{\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\nu}\}$ and $\{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\nu}}\}$ 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. Lastly, we present the recovery results for the individual parameters $\{\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\nu}\}$.

4.6.1 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 (the PCMHP(d,e)); and (2) the interval-censoring of the timestamped MHP data. Since the intensity $\boldsymbol{\xi}_E(t)$ of the PCMHP (d,e) has to be estimated numerically (see Section 4.4.2), numerical approximation error also contributes to type (1) information loss. The numerical approximation error is minimal for sufficiently small Δ^P and γ^h (see Appendix C.9) and vanishes if $\Delta^P, \gamma^h \to 0$.

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 below). 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.

4.6.2 Dataset

Given (d, e), we construct two synthetic datasets: the *timestamp dataset* and the *partially interval-censored dataset*. The former consists of samples from a d-dimensional MHP. The latter is identical to the former, except that it has the E dimensions interval-censored.

We start by estimating the parameter recovery of a 2-dimensional MHP process using PCMHP$(2,1)$. We consider an MHP with $\rho(\boldsymbol{\alpha}) = 0.75$ and parameters $\alpha^{11} = 0.32$, $\alpha^{12} = 0.5$, $\alpha^{21} = 0.3$, $\alpha^{22} = 0.4$, $\theta^{11} = 0.5$, $\theta^{12} = 1.0$, $\theta^{21} = 0.5$, $\theta^{22} = 1.25$ and $\nu^1 = \nu^2 = 0.1$ We set $\rho(\boldsymbol{\alpha}) \in \{0.5, 0.75, 0.9\}$. We also test another parameter combination with $\rho(\boldsymbol{\alpha}) = 0.5$ (*i.e.* subcritical) and $\rho(\boldsymbol{\alpha}) = 0.9$ (*i.e.* approaching the critical regime) in Appendix C.14.

For a given parameter set, we sample 2500 event sequences $\mathscr{H}^1_{100} \cup \mathscr{H}^2_{100}$ over the time interval $[0, 100)$ using the MHP thinning algorithm [82]. Following a procedure similar to prior literature [102], we partition the 2500 event sequences into 50 groups, and each group of $N_{sequences} = 50$ events is used for joint fitting, yielding a single parameter set estimate. In total, we obtain 50 sets of parameter estimates from the sample.

We construct the partially interval-censored dataset by interval-censoring $\mathscr{H}^1_{100}$, 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 $\{[0,2), [2,4), \ldots, [98,100)\}$.

4.6.3 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:

- *timestamp dataset:* we use the point-process log-likelihood on all dimensions, defined in Eq. (4.12): $\sum_{j=1}^{d} \mathscr{L}^j_{\text{PP-LL}}(\boldsymbol{\Theta}; T)$.

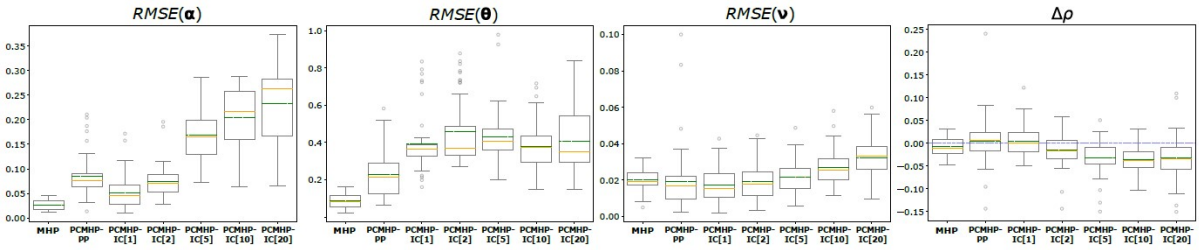


Figure 4.2: Comparison of performance metrics in the parameter recovery experiment across model fits: MHP (*i.e.* the data-generating process), PCMHP-PP and PCMHP-IC for varying interval sizes (1, 2, 5, 10 and 20). (left to right) RMSE for each parameter type $\{\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\nu}\}$ and spectral radius estimation error $\Delta\rho$. Samples are drawn from a 2-dimensional MHP with spectral radius $\rho(\boldsymbol{\alpha}) = 0.75$. Hyperparameters are $T = 100$ and $N_{sequences} = 50$. The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively.

- *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 Eq. (4.10)): $\sum_{j=1}^{e} \mathscr{L}^{j}_{\text{IC-LL}}(\boldsymbol{\Theta}; T) + \sum_{j=e+1}^{d} \mathscr{L}^{j}_{\text{PP-LL}}(\boldsymbol{\Theta}; T)$.

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 also specify k – the length of the observation window – as PCMHP-IC[k].

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

4.6.4 Results

Fig. 4.2 shows RMSE($\boldsymbol{\alpha}$), RMSE($\boldsymbol{\theta}$), RMSE($\boldsymbol{\nu}$) and $\Delta\rho$ across model fits. 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. 4.2, the gap between the first two boxplots (*i.e.* MHP vs. 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 vs. partially interval-censored

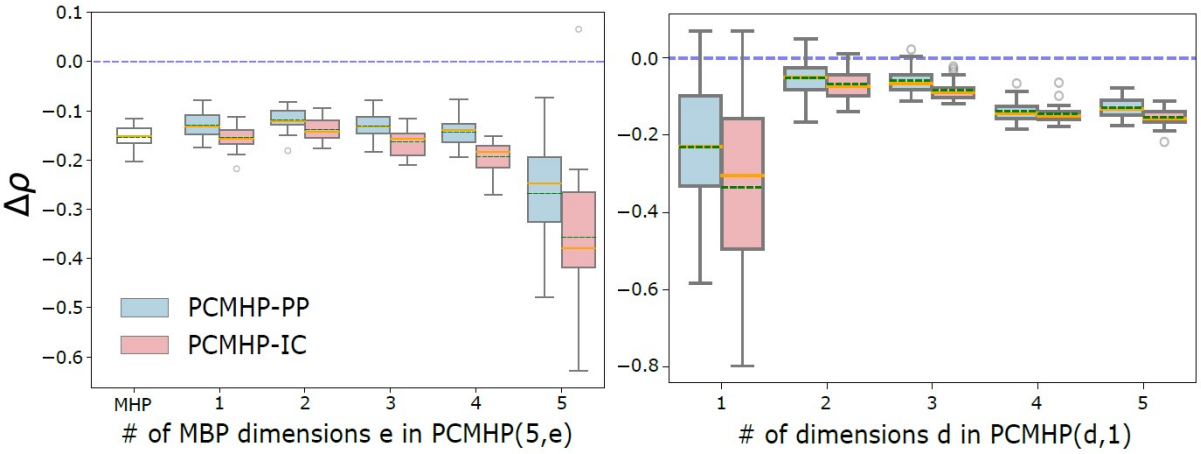


Figure 4.3: (Left) Relating the spectral radius estimation error $\Delta\rho$ of PCMHP$(5, e)$ and the number of MBP dimensions e. Note that PCMHP$(5, 0)$ is the MHP (*i.e.* the data-generating process). (Right) Relating the spectral radius estimation error $\Delta\rho$ of PCMHP$(d, 1)$ and the model dimensionality d. In both plots, samples are drawn from a d-dimensional MHP with spectral radius $\rho(\boldsymbol{\alpha}) = 0.92$. Hyperparameters are $T = 100$, $N_{sequences} = 20$ and intervalsize=1. We fit two models for each PCMHP column: PCMHP $-$ PP (*i.e.* PCMHP fit on timestamp data on all dimensions) and PCMHP $-$ IC (*i.e.* PCMHP fit on interval-censored data on the first e dimensions and timestamp data on the last $d - e$ dimensions). The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively.

data) indicates interval censoring error. The gaps between succeeding boxplots indicate the effect of wider observation windows.

Fig. 4.2 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$, the model mismatch error appears negligible; it is only for higher values of the observation window length (≥ 5) that the performance starts degrading due to information loss error. Both error types are present for $\boldsymbol{\theta}$. See Appendix C.14 for individual parameter fits.

Though we observe that the generating parameters are not always correctly recovered, we see in the rightmost subplot of Fig. 4.2 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 $\rho(\boldsymbol{\alpha})$ 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.

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 $\rho(\boldsymbol{\alpha}) = 0.92$. Results for other error metrics are in Appendix C.14.

In the left subplot of Fig. 4.3, 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. 4.3, 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.*, $d - e \neq 0$).

4.7 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 $[0, T^{train})$. 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.*, $E = \{views, shares\}$. The third dimension (tweets) is provided as event times, *i.e.* $E^c = \{tweets\}$. 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 $[T^{train}, T^{test})$.

4.7.1 Interval-Censored Forecasting with PCMHP

To each YouTube video corresponds a partially interval-censored Hawkes realization. A straightforward approach to predict the unfolding of the realization during $[T^{train}, T^{test})$ is to sample timestamps from PCMHP$(3, 2)$ on each of the three dimensions, conditioned on data before T^{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.

Below is an efficient procedure to calculate expected counts that leverages the compensator Ξ_E and requires sampling only the E^c dimensions (*i.e.*, tweets). Let $\mathscr{P}[T^{train}, T^{test}) = \bigcup_{i=1}^{P-1}[o_i, o_{i+1})$, where $o_1 = T^{train}$ and $o_P = T^{test}$, be a partition of $[T^{train}, T^{test})$.

1. Sample only the E^c dimensions on $[T^{train}, T^{test})$.

2. Compute expected counts on $\mathscr{P}[T^{train}, T^{test})$ as $\{\Xi_E(o_{i+1})$ - $\Xi_E(o_i) | i \in 1 \cdots P-1\}$.

3. Compute the average of $\{\Xi_E(o_{i+1})$ - $\Xi_E(o_i) | i \in 1 \cdots P-1\}$ 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 Appendix C.13.

4.7.2 Dataset, Experimental Setup and Evaluation

We use two subsets of the ACTIVE dataset [103] for model fitting and evaluation. The first subset – dubbed ACTIVE 20% – contains a 20% random sample of the ACTIVE dataset [103], *i.e.*, 2,834 videos published between 2014-05-29 and 2014-12-26. 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 [26, 126]; 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 Appendix C.15

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 ($T^{train} = 90$). It is known that generative models are suboptimal for prediction [76] 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 Appendix C.15.

The days $91-120$ are used for evaluation ($T^{test} = 120$). We measure prediction performance using the Absolute Percentile Error (APE) metric [103], which accounts for the long-tailness of online popularity – *e.g.*, the impact of an error of 10,000 views is very different for a video getting 20,000 views per day compared to a video getting 2 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.

4.7.3 Models and Baseline

We consider two 3-dimensional PCMHP models: PCMHP(3,2) and PCMHP(3,3). The former treats the tweets as a Hawkes dimension (see Definition 4.1) 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 Hawkes Intensity Process (HIP) [103], a parametric popularity prediction model discussed in Section 4.3. 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 $[T^{train}, T^{test})$ to get forecasts for the view counts on $[T^{train}, T^{test})$. 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.* $\frac{1}{\sum_{t=1}^{90} t}\sum_{t=1}^{90} t \cdot \#\text{tweets}(t)$ and $\frac{1}{\sum_{t=1}^{90} t}\sum_{t=1}^{90} t \cdot \#\text{shares}(t)$, which assigns a higher weight to more recent counts.

4.7.4 Results

Fig. 4.4 illustrates the fits of PCMHP(3,2) and the baseline HIP [103] for a sample video from `ACTIVE`. Visibly, we see that PCMHP(3,2) and HIP have comparable fits of the popularity dynamics (left column) during the training period (unshaded area), but PCMHP(3,2) outputs a much tighter fit during the test period (gray shaded area). We also observe two advantages of PCMHP. First, being a multivariate process that captures endogenous dynamics across its dimensions, PCMHP(3,2) provides a prediction for future share and tweet counts (center and left columns), in addition to the number of views. In contrast, HIP treats views (i.e. popularity) as exogenously driven by tweets and shares and thus can only predict the views' dimension. Second, PCMHP can quantify the uncertainty of the popularity prediction by

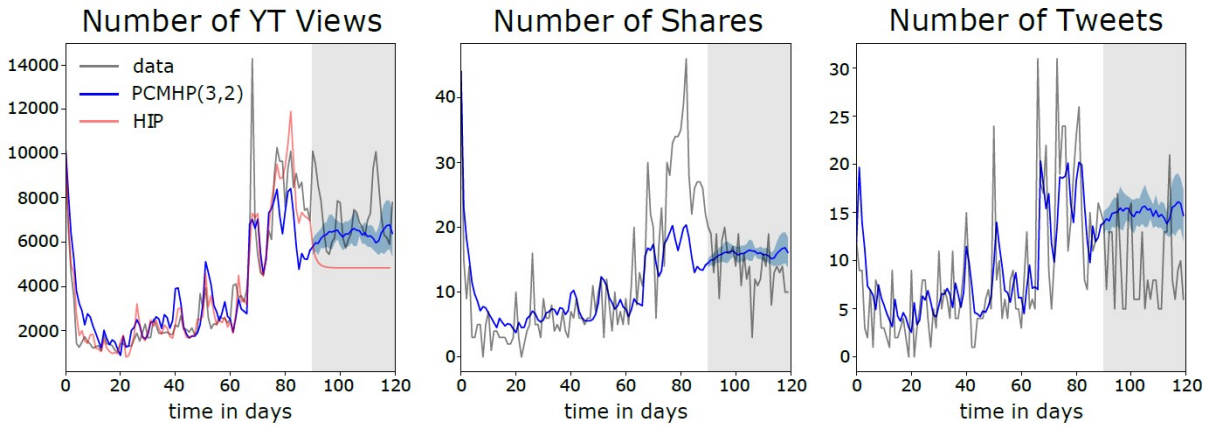


Figure 4.4: Comparison of fits and predictions of our proposal PCMHP(3,2) and the baseline HIP [103] for views (left), shares (center) and tweets (right) for a sample video from ACTIVE: a trailer for the 2014 movie Whiplash (id $7d\_jQycdQGo$). The first 90 days are used to fit model parameters, while 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.

Table 4.2: Performance comparison of PCMHP(3,3), PCMHP(3,2) and HIP on (a) a random sample that comprises 20% of the videos in ACTIVE, and (b) the set of dynamic videos from ACTIVE: mean, median, and standard deviation of the percentile errors for each model. Best-performing score in bold.

	ACTIVE20% (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	10.86	**7.28**	9.31
Median	**2.55**	4.69	4.96	4.82	**3.79**	4.73
StdDev	**7.13**	8.34	9.89	14.24	**9.58**	11.89

sampling multiple unfoldings of a realization and computing the variance of the samples (shown as the blue shaded area in Fig. 4.4).

In Table 4.2, we tabulate the mean, median and standard deviation of percentile errors for PCMHP(3,3), PCMHP(3,2), and HIP on ACTIVE 20% and DYNAMIC VIDEOS. We observe that the PCMHP flavors consistently outperform the baseline HIP on both datasets. Visibly, on ACTIVE 20%, PCMHP(3,3) outperforms PCMHP(3,2). This is because most videos in ACTIVE 20% do not exhibit much activity during the test period. Consequently, as a nonhomogeneous Poisson process with no self-excitation, PCMHP(3,3) fits better such flat trends than the self-exciting PCMHP(3,2) and HIP models. On DYNAMIC VIDEOS 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 4.5 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).

4.8 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.

4.8.1 Dataset

We curate and align two data sources.

The first dataset contains COVID daily case counts from the Johns Hopkins University [29]. 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: UK, USA, Brazil, China, France, Germany, India, Italy, Spain, Sweden, and the Philippines. We select the same countries as [14], to which we add the Philippines.

The second dataset contains timestamps of COVID-19-related news articles provided by the NLP startup Aylien [3]. 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*). Lastly, 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 10 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.

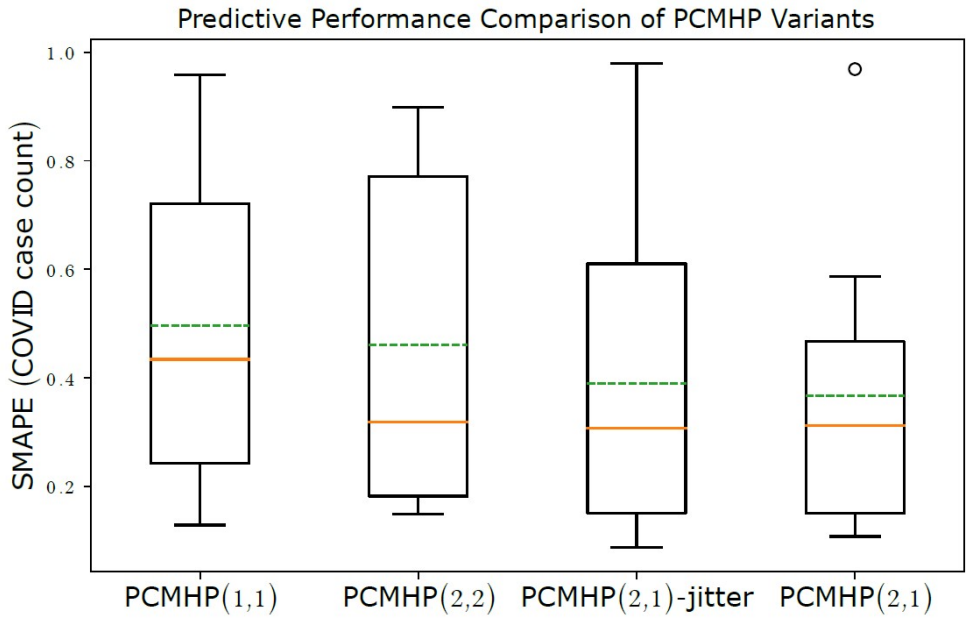


Figure 4.5: Performance comparison of PCMHP(1,1), PCMHP(2,2), PCMHP(2,1)-jitter and PCMHP(2,1) on the COVID case count prediction task over our sample of 11 countries. The dashed line and solid line indicate the mean and median estimates, respectively.

4.8.2 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. Lastly, to test whether 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 [119], and fit PCMHP(2,1) to this dataset. We call this baseline PCMHP(2,1)-jitter.

Similar to Section 4.7, we split our timeframe into a training period $[0, T^{train} = 90)$ and a testing period $[T^{train}, T^{test} = 120)$. 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 $SMAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{|A_t|+|F_t|}$, where F_t and A_t are the forecasted and actual values at time t, respectively.

Across our sample of 11 countries, we see in Fig. 4.5 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

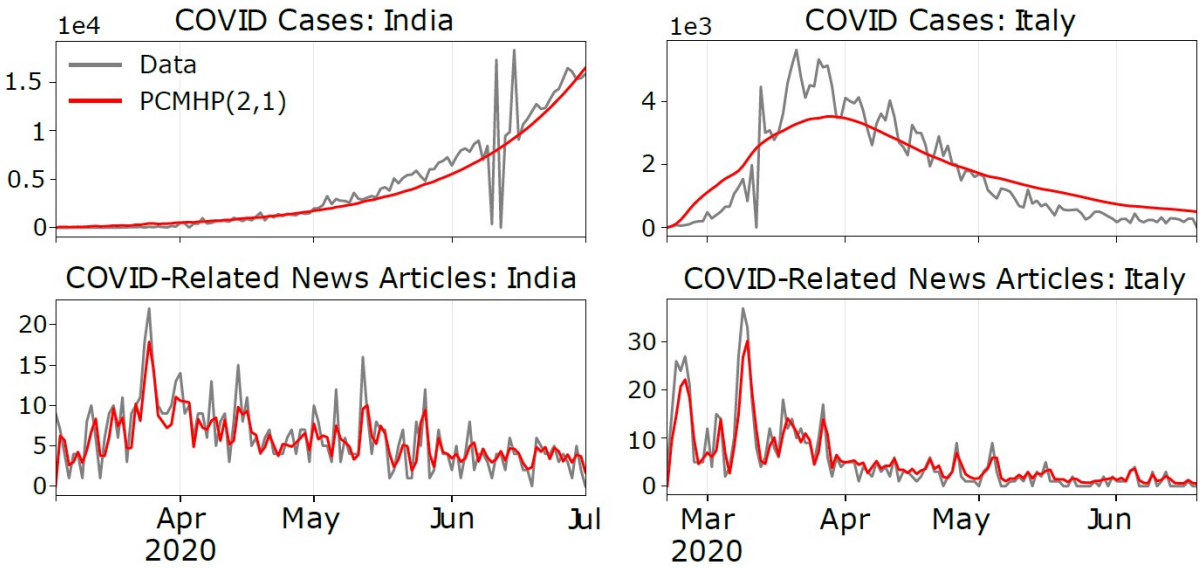


Figure 4.6: Observed and PCMHP(2,1)-fitted daily COVID-19 case counts (top row) and COVID-19-related news articles (bottom row) for India (left column) and Italy (right) during the early stage of the outbreak.

the complex COVID case count dynamics. This supports our claim in Section 4.5 that 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 vs. PCMHP(2,1)).

4.8.3 Results

Fig. 4.6 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 Appendix C.16. Visible from Fig. 4.6, PCMHP(2,1) captures well the dynamics of both countries. Based on the sample-based fit score introduced in Appendix C.16, 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.

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.

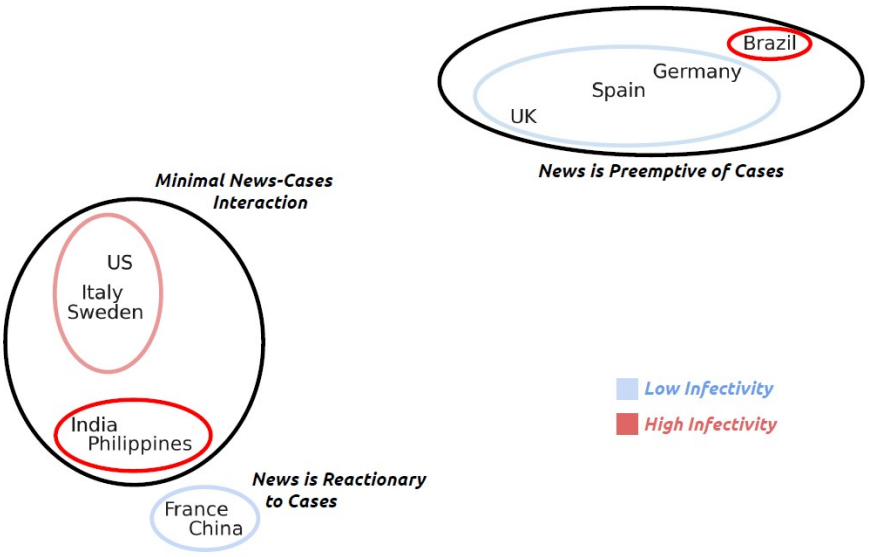


Figure 4.7: Labeled tSNE visualization of the clusters obtained from the fitted PCMHP(2, 1) parameters across the 11 countries we consider.

Table 4.3: K-means cluster centroids on the parameters obtained by fitting PCMHP(2, 1) on the case count and news article dataset.

Cluster	θ^{11}	θ^{12}	θ^{21}	θ^{22}	α^{11}	α^{12}	α^{21}	α^{22}	ν^1	ν^2
UK, German, Spain	0.76	0.12	1.82	1.84	0.79	3.60	0.02	0.42	0.03	0.28
Brazil	0.13	0.01	1.89	2.46	1.08	5.48	0	0.38	0	1.47
China, France	0.62	4	1.70	3.55	0.68	0.73	0.3	0.39	0	0.05
US, Italy, Sweden	0.67	0.22	1.61	1.51	0.93	0.73	0.006	0.65	0.29	0.59
India, Philippines	0.12	2.28	2.15	1.88	1.35	0.54	0.007	0.58	0.08	0.66

The $k = 5$ clusters are shown in Table 4.3 and visualized in Fig. 4.7 using t-SNE [121]. The first cluster (the UK, Germany, Spain) has high α^{12} and low α^{11}. The second cluster – made solely of Brazil – has both a high α^{12} and a very high α^{11}. With high α^{12}, the two clusters contain countries where *news strongly preempts cases*. The third cluster (China and France) has a high α^{21} indicative of *news playing a reactive role to cases*. The fourth cluster (US, Italy, Sweden) and fifth cluster (India, Philippines) both have low α^{12} and α^{21}, indicating *little interaction between news and cases*. We notice that COVID infectiousness is much higher in the fifth cluster (India, Philippines), with α^{11} greater than one (each case generates more than one case) and θ^{11} lowest across all clusters (slow decay, therefore long influence from cases to cases). Our fits indicate that India and the Philippines are countries particularly affected by COVID-19 in the early days.

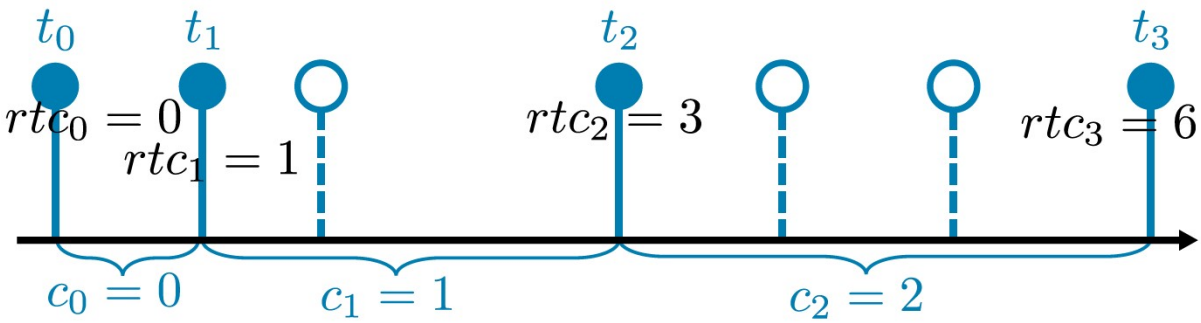


Figure 4.8: Partially interval-censored data handled by IC-TH [62]: observed events are illustrated with lollipops featuring solid lines, while unobserved retweet events are depicted with dotted lines. The Twitter API provides only retweet counts (i.e., rtc_0, rtc_1, ...), while the exact timestamps for the unobserved events are missing. This graphic was pulled from [62].

4.9 Interval-Censored Transformer Hawkes

In this section, we briefly introduce an alternative notion of the partially interval-censored setup and a deep learning-based methodology known as the Interval-Censored Transformer Hawkes (TH) architecture to handle this setup. This approach, detailed in [62], is presented here for completeness.

Alternative Notion of Partially Censored Data. The partially interval-censored setup considered in [62] is motivated by the sampled-down effect from the Twitter API [127] which returns only 1% of actual tweets in the streaming API. Piecing together the missing event counts is achievable by leveraging the *retweeted_count* field in the tweet metadata returned by the API, which gives the number of times the tweet has been retweeted. We show a sample timeline in Fig. 4.8. Here, we have a one-dimensional timeline of events given by $\{t_0, t_1, t_2, t_3\}$ returned by the downsampled Twitter API. To determine the number of tweets between observed events t_i, we can use the *retweeted_count* field, yielding $\{rtc_0, rtc_1, rtc_2, rtc_3\}$, where rtc_i is the number of tweets up to (but excluding) event i. We can then comptue the number of *missing* events between t_{i+1} and t_i as c_i. Hence, our dataset consists of both tweet timestamps $\{t_i\}$ and missing event counts between tweets $\{c_i\}$. Note the difference in setup between Fig. 4.1 and Fig. 4.8. In Fig. 4.1 timestamps and event counts occur in different dimensions, while in Fig. 4.8 they exist on the same timeline.

Transformer Hawkes. The TH model is a deep learning-based model of event sequences $\mathcal{H} = \{t_1, t_2, \ldots\}$ that uses a self-attention-based mechanism to model long-term dependencies. The conditional intensity of the TH model [137] is given by

$$\lambda^{TH}(t|\mathcal{H}) = f(\mathbf{w}^\top \mathbf{h}(t)), \tag{4.13}$$

where f is the softmax function, $\mathbf{w}$ is a weight matrix to be inferred, and the hidden state $\mathbf{h}(t)$ is computed as

$$\begin{aligned}
&\mathbf{h}(t_j) = \mathbf{H(j,:)},\\
&\mathbf{H(j,:)} = \mathrm{ReLU}(\mathbf{S}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2,\\
&\mathbf{S} = \mathrm{Concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)\mathbf{W}^O,\\
&\mathrm{head}_i = \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}_i^Q(\mathbf{X}\mathbf{W}_i^K)^\top}{\sqrt{d_k}}\right)\mathbf{X}\mathbf{W}_i^V,
\end{aligned}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K \in \mathbb{R}^{d_m \times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_m \times d_v}$, $\mathbf{W}^O \in \mathbb{R}^{hd_v \times d_m}$, $\mathbf{W}_1 \in \mathbb{R}^{d_m \times h}$, $\mathbf{b}_1 \in \mathbb{R}^h$,$\mathbf{W}_2 \in \mathbb{R}^{h \times d_m}$ and $\mathbf{b}_2 \in \mathbb{R}^{d_m}$ are learned weights, d_m is the embedding dimension, d_k, d_v are the hidden dimensions, and h is the number of heads. The input $\mathbf{X}$ is obtained via the temporal encoding

$$\mathbf{X}(j,i) = \begin{cases} \cos(t_j/1000^{\frac{i-1}{d_m}}) & \text{if } i \text{ is odd,} \\ \sin(t_j/1000^{\frac{i}{d_m}}) & \text{if } i \text{ is even,} \end{cases} \tag{4.14}$$

which maps each event timestamp t_j into a d_m-dimensional feature vector $\mathbf{X}_j$.

Interval-Censored Transformer Hawkes. The Transformer Hawkes (TH) architecture only models event timestamps. To accommodate the mix of event counts and timestamps in Fig. 4.8, we generalize TH to the IC-TH. First, we represent our mixed dataset with the triple (o_i, d_i, c_i) for the i^{th} observation. Here, o_i is observation time, d_i is the observation duration, and c_i is the number of events (observed or missing) within $(o_i,\, o_i + d_i)$. Denoting dt as an infinitesimal interval, observed events can then be represented as $(t_i - dt, 2dt, 1)$, while missing event counts are represented as $(t_i, t_{i+1} - t_i, c_i)$. A full retweet cascade can then be represented as $\mathcal{H} = \{(t_i, d_i, c_i) | i = 0, \ldots, m, o_i + d_i \leq T\}$.

With this representation, we can express the conditional intensity of the IC-TH as

$$\xi^{IC-TH}(t|\mathcal{H}) = \mathbb{E}_{\mathcal{H}^u}[\lambda(t)] = f(\mathbf{w}^\top \mathbf{h}(t)), \tag{4.15}$$

where $\mathcal{H}^u = \{(t_i, d_i, c_i) | d_i > 2dt, c_i > 0\}$ is the subset of $\mathcal{H}$ consisting of the missing event counts and the right-hand side is given by Eq. (4.13). Note that for the IC-TH, the temporal encoding in Eq. (4.14) is augmented with the duration d_i and event count c_i as masks to $\mathbf{X}_i$ (see [62] for full details). As special cases, note that if $\mathcal{H}^u = \varnothing$ or $\mathcal{H}^u = \mathcal{H}$, Eq. (4.15) reduces to $\lambda(t)$ in Eq. (4.1) or $\xi(t)$ in Eq. (4.2), respectively.

Given $\mathcal{H}$ (and setting $\mathcal{H}^c = \{(t_i, d_i, c_i) | d_i = 2dt, c_i = 1\}$ as the history subset of observed event times), we can infer the parameters of the IC-TH by maximizing the following log-likelihood:

$$\mathscr{L}_{\text{IC-TH-LL}}(\boldsymbol{\Theta}) = \sum_{t_i \in \mathcal{H}^u} c_i \log \Xi(t_i, t_{i+1}; \boldsymbol{\Theta}) + \sum_{t_i \in \mathcal{H}^c} \log \xi(t_i; \boldsymbol{\Theta}) - \sum_{t_i \in \mathcal{H}} \log \Xi(t_i, t_{i+1}; \boldsymbol{\Theta}), \tag{4.16}$$

where $\Xi(t_i, t_{i+1}) = \int_{t_i}^{t_{i+1}} \xi(z) dz$.

4.10 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 chapter, we derive the conditional intensity function of the PCMHP 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 multivariate Hawkes process 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 multivariate Hawkes process 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 multivariate Hawkes process 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 Hawkes Intensity Process [103].

Third, to demonstrate 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.

Future Work. There are three 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 PCMHP 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. Lastly, 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 whilst also being flexible enough to be used in the partially interval-censored setting.

CHAPTER 5

CONCLUSION

In this chapter, we summarize the main contributions of this thesis and discuss potential avenues for future research.

5.1 Thesis Summary

This thesis investigates stochastic models of information spread in online social systems to uncover their latent mechanisms, predict future online diffusions, and evaluate the effects of external interventions. By integrating contextual and domain-specific expert opinions, we introduce structural assumptions into the Hawkes process, leading to better model interpretability and accuracy.

The main contributions of this thesis are:

- **A two-tier finite attention model of the online opinion ecosystem that jointly models inter-opinion dynamics and the effect of positive interventions.** In Chapter 2 we introduced a two-tier model of the online opinion ecosystem called the Opinion Market Model (OMM), where the first tier models the size of the opinion attention market using a multivariate discrete-time Hawkes process, while the second tier employs the market share attraction model to capture cooperation and competition among opinions and the influence of positive interventions. Validated through synthetic and real-world datasets, the OMM outperforms state-of-the-art models and uncovers latent opinion interactions. Lastly, we demonstrated the OMM's capability as

a testbed for interventions via a counterfactual analysis, where we varied the volume of reputable and controversial media coverage and observed the resulting effect on far-right opinion market share on Facebook and Twitter.

- **A hierarchical mixture model to jointly learn the influence of source-, content- and spread-level factors on the spread of content on social media.** In Chapter 3 we developed the Bayesian Mixture Hawkes (BMH) model, a hierarchical mixture model of separable Hawkes processes to jointly learn the influence of source, content, and cascade-level factors on the spread dynamics of online items. We tested the BMH on two learning tasks, cold-start popularity prediction and temporal profile generalization tasks, and on two real-world retweet cascade datasets referencing articles from controversial and reputable media publishers, demonstrating that the BMH model outperforming state-of-the-art and baseline models and leverage cascade- and item-level features better than the alternatives. Through a counterfactual analysis with the BMH model we show differences in the effectiveness of different headline writing styles (neutral, clickbait, inflammatory) across reputable and controversial publishers. Lastly, we introduced a two-step procedure to optimise headlines before posting time, where text-generating AI is leveraged to produce rewrites for a target headline and the fitted BMH model's predictions on cold-start effectiveness are used to rank the rewrites. The effectiveness of this two-step procedure was demonstrated through a Mechanical Turk experiment.

- **An approach to enable fitting of the multivariate Hawkes process in the partially interval-censored setting.** In Chapter 4 we developed the Partially Censored Multivariate Hawkes Process (PCMHP) to address the challenge of fitting the Hawkes process in the partially interval-censored setting, where we have event timestamps is some dimensions and aggregated event counts in the others. We demonstrated through synthetic tests that the PCMHP approximates the MHP parameters well and recovers the spectral radius of the process. We tested the PCMHP in the YouTube popularity prediction task and show that the PCMHP outperforms the fully interval-censored popularity estimation algorithm Hawkes Intensity Process (HIP). Lastly, we trained the PCMHP on a dataset of daily COVID-19 case counts and COVID-19-related news articles, revealing hidden interaction patterns between cases and news reporting and demonstrating the model's ability to uncover latent structure from real-world data.

5.2 Future Work

Lastly, we explore potential extensions of the work presented in this thesis and discuss the current research directions we are pursuing.

- **Effect of complex intervention strategies on opinion dynamics.** In the counterfactual analysis with the OMM in Section 2.9, we used a step function to represent our intervention to study the effect of media coverage. It would be interesting to explore the effects of other functional shapes of intervention, check whether multiple interventions working together is more effective, and to determine whether an optimal intervention profile exists to minimize far-right opinion.

- **Working with recent data challenges.** In February 2023, Twitter API was put behind a paywall, affecting data collection on the platform. As a consequence, social media data collection is shifting to other platforms like Facebook (via Crowdtangle), which introduces new data challenges. While Twitter data consist of timestamp-based data streams, Facebook data comprises interval-censored aggregated interaction counts. Our current research focuses on developing an equivalent of the BMH model compatible with interval-censored data. Instead of the continuous-time Hawkes process as the base model, we are exploring the discrete-time Hawkes process [14] and the mean-behavior Poisson process [102], which are suitable for interval-censored data.

- **Further exploration on modeling partially interval-censored data.** Future work on modeling partially interval-censored data can be split into three areas. First, a theoretical treatment of the approximation error from PCMHP model mismatch (*i.e.* PCMHP approximating the MHP) should be developed, given that we only performed an empirical evaluation in Chapter 4. Second, alternative methods to estimate the intensity of the PCMHP should be explored, since the discrete convolution approximation we presented is computationally challenging. Lastly, given that the PCMHP is only exciting on a subset of dimensions, it remains an open question whether one can formulate a stochastic process compatible for the partially interval-censored setting that maintains the self- and cross-exciting properties in all dimensions.

APPENDIX A

APPENDIX TO 'OPINION MARKET MODEL: STEMMING FAR-RIGHT OPINION SPREAD USING POSITIVE INTERVENTIONS'

A.1 Full Table of Notation

Table A.1 shows the full table of notations for the OMM.

Notation	Interpretation
P	number of social media platforms
M	number of opinion types
K	number of positive interventions
T	terminal time
Variable	
$S(t)$	input signal accounting for the volume of exogenous events
$X_k(t)$	input signal corresponding to the k^{th} positive intervention
$s_i^p(t)$	market share of opinion i on platform p at time t
$\lambda^p(t)$	conditional intensity of attention volume (Opinion Volume Model)
$\lambda^p(t\|i)$	conditional intensity of opinion i, assuming independence of opinions
$\lambda_i^p(t)$	conditional intensity of opinion i (Opinion Share Model)
$N^p(t)$	total attention volume on platform p at time t, based on OMM
$N_i^p(t)$	number of posts with opinion i on platform p at time t, based on OMM
$e(s_i^p(t), \lambda^q(t\|j))$	opinion share model elasticity w.r.t. endogenous dynamics
$e(s_i^p(t), X_k(t))$	opinion share model elasticity w.r.t. intervention
Data	
n_t^p	number of posts on platform p at time t
$n_{i,t}^p$	number of posts on platform p with opinion i at time t
$s_{i,t}^p$	fraction of posts on platform p with opinion i at time t
Parameter	
μ_j^p	exogenous scaling term for opinion j on platform p, μ_j^p
μ^p	exogenous scaling term for platform p, given by $\mu^p = \sum_{j=1}^M \mu_j^p$
α^{pq}	excitation parameter for intra-platform ($p = q$) and inter-platform (for $p \neq q$) dynamics
θ	memory parameter, describing how fast an event is forgotten, $\theta \in [0, 1]$
γ_{ik}^p	measure of the direct effect of the k^{th} intervention on the market share of opinion i on platform p
β_{ij}^{pq}	measure of the direct effect that opinion j on platform q has on the market share of opinion i on platform p.

Table A.1: Full table of notation in Chapter 2.

A.2 Model Likelihood, Estimation, Simulation and Gradients

In this section we provide technical details for Section 2.4. We first go over the derivation of the model likelihoods, followed by the estimation and simulation algorithms for the two-tier OMM model. Lastly, we derive the model gradients.

A.2.1 Likelihood Formulation

Likelihood function $\mathscr{L}_1(\mathbf{\Theta}_1|\{n_t^p\}_{p,t})$**,** where $\mathbf{\Theta}_1 = \{\mu^P, \alpha^{pq}, \theta\}$. The log-likelihood function can be derived by

$$
\begin{aligned}
&\mathscr{L}_1(\mathbf{\Theta}_1|\{n_t^p\}_{p,t}) \\
&= \log \mathbb{P}\left\{\bigcup_{t=1}^{T}\bigcup_{p=1}^{P}[N^p(t) = n_t^p]\right\} \\
&= \sum_{t=1}^{T}\sum_{p=1}^{P}\log \mathbb{P}\left\{N^p(t) = n_t^p\right\} \\
&= \sum_{t=1}^{T}\sum_{p=1}^{P}\log\left[\frac{e^{-\lambda^p(t)}\lambda^p(t)^{n_t^p}}{n_t^p!}\right] \\
&\propto \sum_{t=1}^{T}\sum_{p=1}^{P}\left[n_t^p \log\lambda^p(t) - \lambda^p(t)\right] \qquad \text{(A.1)}
\end{aligned}
$$

Likelihood function $\mathscr{L}_2(\mathbf{\Theta}_2|\mathbf{\Theta}_1, \{n_{i,t}^p\}_{i,p,t})$**,** where $\mathbf{\Theta}_2 = \{\mu_j^p, \gamma_{ik}^p, \beta_{ij}^{pq}\}$. Instead of estimating the parameters $\mu_j^p \in \mathbb{R}$, we can estimate the normalized parametes $\hat{\mu}_j^p \in [0,1]$, where $\mu_j^p = \mu^p \cdot \hat{\mu}_j^p$. Given that the magnitudes of γ_{ik}^p and β_{ij}^{pq} are typically less than one, estimating normalized parameters $\hat{\mu}_j^p$ instead of μ_j^p avoids scaling problems. Hence, we optimize for $\mathbf{\Theta}_2 = \{\hat{\mu}_j^p, \gamma_{ik}^p, \beta_{ij}^{pq}\}$.

$$
\begin{aligned}
&\mathscr{L}_2(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_1, \{n_{i,t}^p\}_{i,p,t}) \\
&= \log \mathbb{P}\left\{ \bigcup_{t=1}^{T} \bigcup_{p=1}^{P} \bigcup_{i=1}^{M} [N_i^p(t) = n_{i,t}^p] \right\} \\
&= \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \log \mathbb{P}\left\{ N_i^p(t) = n_{i,t}^p \right\} \\
&= \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \log \left[\frac{e^{-\lambda_i^p(t)} \lambda_i^p(t)^{n_{i,t}^p}}{n_{i,t}^p!} \right] \\
&\propto \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \left[n_{i,t}^p \log \lambda_i^p(t) - \lambda_i^p(t) \right] \\
&= \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \left[n_{i,t}^p \log(\lambda^p(t) \cdot s_i^p(t)) - (\lambda^p(t) \cdot s_i^p(t)) \right] \\
&= \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \left[n_{i,t}^p (\log \lambda^p(t) + \log s_i^p(t)) - (\lambda^p(t) \cdot s_i^p(t)) \right] \qquad \text{(A.2)}
\end{aligned}
$$

A.2.2 Estimation Algorithm

We estimate the parameters of OMM with the following two-step formula:

1. Given $\{n_t^p\}_{p,t}$, find $\hat{\boldsymbol{\Theta}}_1 = \boldsymbol{\Theta}_1$ that maximizes

$$\mathscr{L}_1(\boldsymbol{\Theta}_1|\{n_t^p\}_{p,t}) = \sum_{p,t} \left[n_t^p \log \lambda^p(t) - \lambda^p(t) \right]. \qquad \text{(A.3)}$$

2. Given $\{n_{i,t}^p\}_{i,p,t}$ and $\hat{\boldsymbol{\Theta}}_1$, find $\hat{\boldsymbol{\Theta}}_2 = \boldsymbol{\Theta}_2$ that maximizes

$$\mathscr{L}_2(\boldsymbol{\Theta}_2|\hat{\boldsymbol{\Theta}}_1, \{n_{i,t}^p\}_{i,p,t}) = \sum_{i,p,t} [n_{i,t}^p (\log \lambda^p(t) \cdot s_i^p(t)) - (\lambda^p(t) \cdot s_i^p(t))]. \qquad \text{(A.4)}$$

Due to the non-convexity of $\mathscr{L}_1(\cdot)$ and $\mathscr{L}_2(\cdot)$, we avoid local maxima by running the algorithm for multiple starting points and selecting the combination with the largest likelihood.

A.2.3 Sampling Algorithm

We generate samples from OMM by looping the following steps over $t \in \{1, \ldots, T\}$ and each platform p and opinion i.

1. Compute $\lambda_i^p(t) = \lambda^p(t|\cup_{q,s<t}\{n_s^q\}) \cdot s_i^p(t|\cup_{q,j,s<t}\{n_{j,s}^q\})$.
2. Draw a sample $n_{i,t}^p \sim \text{Poi}(\lambda_i^p(t))$.

A.2.4 Gradient Computations

Gradient $\partial_{\boldsymbol{\Theta}_1}\mathscr{L}_1(\boldsymbol{\Theta}_1|\{n_t^p\}_{p,t})$. Differentiating Eq. (A.3), we get

$$\partial_{\boldsymbol{\Theta}_1}\mathscr{L}_1(\boldsymbol{\Theta}_1|\{n_t^p\}_{p,t}) = \sum_{t=1}^{T}\sum_{p=1}^{P}\frac{\partial_{\boldsymbol{\Theta}_1}\lambda^p(t)}{\lambda^p(t)}\cdot\left[n_t^p - \lambda^p(t)\right], \tag{A.5}$$

where

$$\partial_{\mu^q}\lambda^p(t) = \delta_{pq}\cdot S(t) \tag{A.6}$$

$$\partial_{\alpha^{qr}}\lambda^p(t) = \delta_{pq}\cdot\sum_{s<t} f(t-s)\cdot N^r(s) \tag{A.7}$$

$$\partial_\theta\lambda^p(t) = \sum_{q=1}^{P}\sum_{s<t}\alpha^{pq}\cdot\partial_\theta f(t-s)\cdot N^q(s) \tag{A.8}$$

$$\partial_\theta f(t) = (1-\theta)^{t-2}\left[1-\theta t\right]. \tag{A.9}$$

Gradient $\partial_{\boldsymbol{\Theta}_2}\mathscr{L}_2(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_1,\{n_{i,t}^p\}_{i,p,t})$. Differentiating Eq. (A.4), we get

$$\partial_{\boldsymbol{\Theta}_2}\mathscr{L}_2(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_1,\{n_{i,t}^p\}_{i,p,t}) = \sum_{t=1}^{T}\sum_{p=1}^{P}\sum_{i=1}^{M}\left[\frac{\partial_{\boldsymbol{\Theta}_2}\lambda^p(t)}{\lambda^p(t)} + \frac{\partial_{\boldsymbol{\Theta}_2}s_i^p(t)}{s_i^p(t)}\right]\cdot\left[n_{i,t}^p - \lambda^p(t)\cdot s_i^p(t)\right], \tag{A.10}$$

where upon differentiating Eq. (2.7) and Eq. (2.8) we have

$$\partial_{\boldsymbol{\Theta}_2}s_i^p(t) = \frac{\left[\sum_j \mathscr{A}_j^p(t)\right]\partial_{\boldsymbol{\Theta}_2}\mathscr{A}_i^p(t) - \mathscr{A}_i^p(t)\left[\sum_j \partial_{\boldsymbol{\Theta}_2}\mathscr{A}_j^p(t)\right]}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2}, \tag{A.11}$$

and

$$\partial_{\boldsymbol{\Theta}_2}\mathscr{A}_i^p(t) = \mathscr{A}_i^p(t)\cdot\partial_{\boldsymbol{\Theta}_2}\mathscr{T}_i^p(t). \tag{A.12}$$

Plugging in Eq. (A.12) into Eq. (A.11), we get

$$\begin{aligned}
\partial_{\boldsymbol{\Theta}_2} s_i^p(t) &= \frac{\left[\sum_j \mathscr{A}_j^p(t)\right] \mathscr{A}_i^p(t) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_i^p(t)}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2} \\
&\quad - \frac{\mathscr{A}_i^p(t)\left[\sum_j \mathscr{A}_i^p(t) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t)\right]}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2} \\
&= \frac{\mathscr{A}_i^p(t)\left[\sum_j \mathscr{A}_j^p(t) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_i^p(t)\right]}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2} \\
&\quad - \frac{\left[\sum_j \mathscr{A}_i^p(t) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t)\right]}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2} \\
&= \frac{\mathscr{A}_i^p(t)}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2} \cdot \sum_j \mathscr{A}_j^p(t) \cdot \left[\partial_{\boldsymbol{\Theta}_2} \mathscr{T}_i^p(t) - \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t)\right] \\
&= s_i^p(t) \cdot \sum_j s_j^p(t) \cdot \left[\partial_{\boldsymbol{\Theta}_2} \mathscr{T}_i^p(t) - \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t)\right],
\end{aligned}$$

and so

$$\begin{aligned}
\frac{\partial_{\boldsymbol{\Theta}_2} s_i^p(t)}{s_i^p(t)} &= \sum_j s_j^p(t) \cdot \left[\partial_{\boldsymbol{\Theta}_2} \mathscr{T}_i^p(t) - \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t)\right] \\
&= \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_i^p(t) - \sum_j s_j^p(t) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t) \\
&= \sum_j (\delta_{ij} - s_j^p(t)) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t) \qquad \text{(A.13)}
\end{aligned}$$

Plugging in Eq. (A.13) into Eq. (A.10), we have

$$\begin{aligned}
\partial_{\boldsymbol{\Theta}_2} \mathscr{L}_2(\boldsymbol{\Theta}_2 | \boldsymbol{\Theta}_1, \{n_{i,t}^p\}_{i,p,t}) = \sum_{t=1}^{T} \sum_{p=1}^{P} \sum_{i=1}^{M} \Bigg[& \frac{\partial_{\boldsymbol{\Theta}_2} \lambda^p(t)}{\lambda^p(t)} \\
& + \sum_{j=1}^{M} (\delta_{ij} - s_j^p(t)) \cdot \partial_{\boldsymbol{\Theta}_2} \mathscr{T}_j^p(t) \Bigg] \cdot \left[n_{i,t}^p - \lambda^p(t) \cdot s_i^p(t) \right], \qquad \text{(A.14)}
\end{aligned}$$

where

$$\partial_{\mu_j^q} \lambda^p(t) = \delta_{pq} \cdot S(t) \tag{A.15}$$

$$\partial_{\gamma_{ik}^q} \lambda^p(t) = 0 \tag{A.16}$$

$$\partial_{\beta_{ij}^{pq}} \lambda^p(t) = 0 \tag{A.17}$$

$$\partial_{\mu_j^q} \mathcal{T}_i^p(t) = \beta_{ij}^{pq} \tag{A.18}$$

$$\partial_{\gamma_{jk}^q} \mathcal{T}_i^p(t) = \delta_{ij}\delta_{qp} \cdot \sum_{s<t} f(t-s) \cdot X_k(s) \tag{A.19}$$

$$\partial_{\beta_{jk}^{qr}} \mathcal{T}_i^p(t) = \delta_{ij}\delta_{qp} \cdot \lambda^r(t|k) \tag{A.20}$$

A.2.5 Fitting on Multiple Samples

Suppose that we are given $n_{samples}$ samples to fit the OMM.

Let $\mathcal{S} = \left\{ \{n_{i,t}^p\}_{i,p,t}^s | s \in \{1, \ldots, n_{samples}\} \right\}$. One can define the joint likelihood over $\mathcal{S}$ as the average likelihood over the $n_{samples}$ samples. That is,

$$\mathcal{L}_1(\boldsymbol{\Theta}_1|\mathcal{S}) = \frac{1}{n_{samples}} \sum_{s=1}^{n_{samples}} \mathcal{L}_1(\boldsymbol{\Theta}_1|\{n_t^p\}_{p,t}^s)$$

$$\mathcal{L}_2(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_1, \mathcal{S}) = \frac{1}{n_{samples}} \sum_{s=1}^{n_{samples}} \mathcal{L}_2(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_1, \{n_{i,t}^p\}_{i,p,t}^s).$$

where $\mathcal{L}_1(\cdot)$ and $\mathcal{L}_2(\cdot)$ are defined in Eq. (A.3) and Eq. (A.4), respectively. Parameter optimization proceeds in the same setup as the two-step procedure detailed in Section 3.

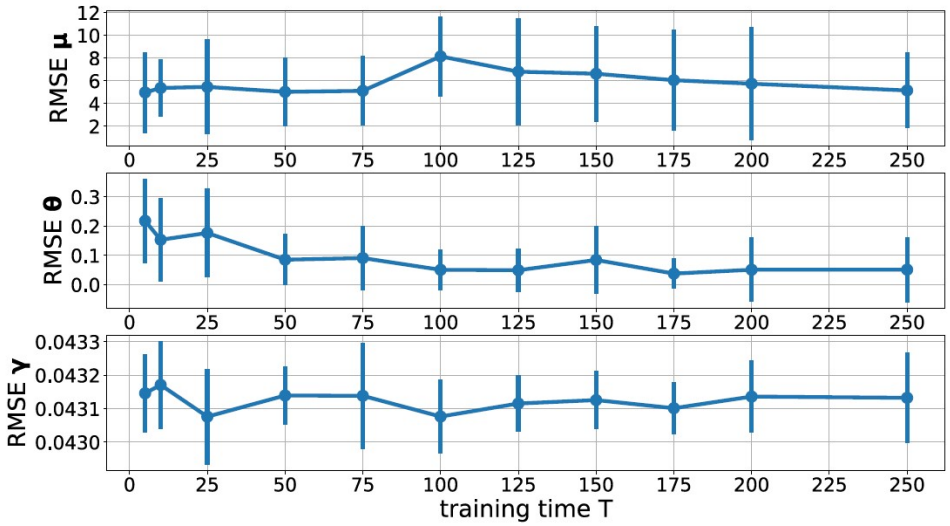


Figure A.1: Additional results on synthetic data. We show the convergence of the RMSE of the $\boldsymbol{\mu}$, θ, $\boldsymbol{\beta}$ as we increase the training time T.

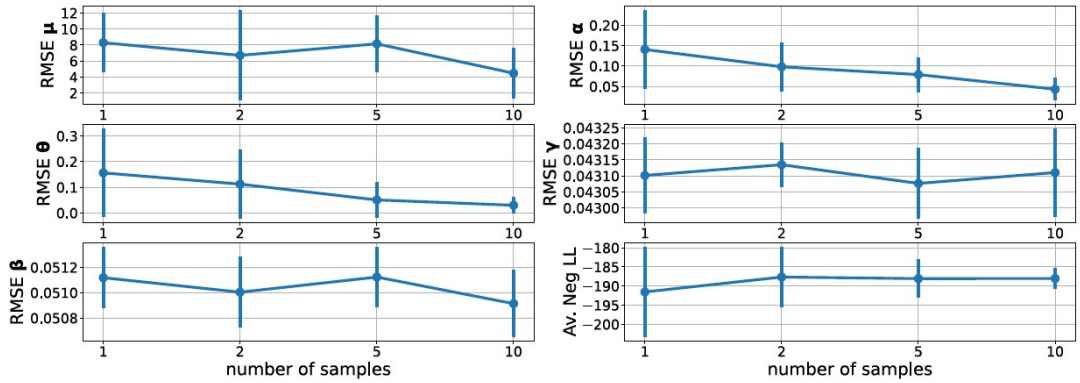


Figure A.2: Additional results on synthetic data. We show the behavior of the RMSE of our parameter set and the average negative log likelihood as we vary the number of samples in the joint fit.

A.3 Additional Results for Synthetic Data

In Fig. A.1 we show the behavior of the RMSE for $\boldsymbol{\mu}$, θ, $\boldsymbol{\beta}$ as we increase the training time T. Error stabilises and the model converges as we increase T. In Fig. A.2 we show behavior of the RMSE of our parameters as we vary the number of samples in the joint fit $n_{samples}$. Increasing the number of samples improves performance on the first-tier parameters $\boldsymbol{\mu}$, θ and $\boldsymbol{\alpha}$, but does not have a strong improvement on the second-tier parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Increasing $n_{samples}$ stabilizes the likelihood of the fit.

A.4 Additional Model Details

A.4.1 Stability of the Softmax Function

In Eq. (2.8), the tendency $\mathscr{T}_i^p(t)$ is unconstrained, and it can take both really large or really small numbers, which leads to numerical overflow and underflow in Eq. (2.7). To remedy this, instead of Eq. (2.8) we use

$$\tilde{\mathscr{A}}_i^p(t) = \exp\left[\mathscr{T}_i^p(t) - \max_{k\in\{1,\dots,M\}} \mathscr{T}_k^p(t)\right],$$

which does not affect market share calculations since

$$s_i^p(t) = \frac{\mathscr{A}_i^p(t)}{\sum_{j=1}^M \mathscr{A}_j^p(t)} = \frac{\tilde{\mathscr{A}}_i^p(t)}{\sum_{j=1}^M \tilde{\mathscr{A}}_j^p(t)}.$$

Gradient and elasticity calculations are unaffected when we use $\tilde{\mathscr{A}}_i^p(t)$ instead of $\mathscr{A}_i^p(t)$.

A.4.2 Regularizing the Bushfire Opinion Share Model

Fitting the opinion share model to data involves estimation of $\mathbf{\Theta}_2 = \{\hat{\mu}_j^p, \gamma_{ik}^p, \beta_{ij}^{pq}\}$, a total of $P\times M + P\times M\times K + P^2\times M^2$ parameters. Given the high dimensionality of this space, for the bushfire case study we opted to reduce the space of solutions by imposing platform-dependent structure on γ_{ik}^p via regularization.

Let $\mathbf{M}^p$ be the mask matrices given by

$$M_{ik}^{FB} = \begin{cases} 0, & \left(i \le \lfloor\frac{K}{2}\rfloor \wedge k \le \lfloor\frac{K}{2}\rfloor\right) \vee \left(i > \lfloor\frac{K}{2}\rfloor \wedge k > \lfloor\frac{K}{2}\rfloor\right) \\ 1, & \text{otherwise} \end{cases}$$

$$M_{ik}^{TW} = \begin{cases} 0, & (i=k) \vee \left(i = k - \lfloor\frac{K}{2}\rfloor\right) \vee \left(k = i - \lfloor\frac{K}{2}\rfloor\right) \\ 1, & \text{otherwise} \end{cases}$$

Instead of Eq. (A.4), we solve

$$\hat{\mathbf{\Theta}}_2 = \arg\min_{\mathbf{\Theta}_2}\left[-\mathscr{L}_2(\mathbf{\Theta}_2|\hat{\mathbf{\Theta}}_1, \{n_{i,t}^p\}_{i,p,t}) + \lambda \sum_{p,i,k}\left|\gamma_{ik}^p \cdot M_{ik}^p\right|\right],$$

where λ is a regularization parameter we set to 0.1.

Intuitively, the regularization encodes the echo chamber effect observed in Facebook far-right groups: far-right sympathizers interact mostly with news from controversial outlets, with limited interaction with reputable outlets. Similarly, far-right opponents interact mostly

with reputable news, with limited interaction with controversial outlets. Given the more dialog-heavy nature of Twitter where exchanges between sympathizers and opponents are more common, we assume news from reputable and controversial outlets penetrate both far-right sympathizers and opponents, though we assume that sympathizers and opponents of a given opinion are only concerned with (and influenced by) news of the same opinion.

A.4.3 Transformations on $\lambda^q(t|j)$ and $\bar{X}_k(s)$ in $\mathscr{T}_i^p(t)$

We perform two transformations on $\lambda^q(t|j)$ and $\bar{X}_k(s)$ in Eq. (2.11) to improve model fit.

First, it was observed that $\lambda^q(t|j)$ has a skewed distribution over time. The skewness is problematic since we estimate a time-independent linear parameter β_{ij}^{pq} for the direct effect of $\lambda^q(\cdot|j)$ on $\mathscr{T}_i^p(t)$. To reduce the skewness of $\lambda^q(\cdot|j)$, we transform $\lambda^q(t|j)$ to $\log\left[\lambda^q(t|j)+1\right]$, where we add 1 to avoid taking the logarithm of 0. Second, since $\mathscr{T}_i^p(t)$ is a linear combination of $\lambda^q(t|j)$ and $\bar{X}_k(t)$ terms, which could have totally different scales, we standardize these terms to bring them to a normalized scale. Let

$$\tilde{\lambda}^q(t|j) = \frac{\log\left[\lambda^q(t|j)+1\right] - \text{mean}_s \log\left[n_{j,s}^q+1\right]}{\text{std}_s \log\left[n_{j,s}^q+1\right]},$$

$$\tilde{X}_k(t) = \frac{\bar{X}_k(t) - \text{mean}_s\left[\bar{X}_k(s)\right]}{\text{std}_s\left[\bar{X}_k(s)\right]}$$

where the mean and standard deviation are computed over the training period.

Instead of Eq. (2.11), we use the following form of the tendency:

$$\mathscr{T}_i^p(t) = \sum_{k=1}^{K} \gamma_k^p \cdot \tilde{X}_k(t) + \sum_{q=1}^{P}\sum_{j=1}^{M} \beta_{ij}^{pq} \cdot \tilde{\lambda}^q(t|j). \tag{A.21}$$

A.4.4 Adjusting for Multiple Exogenous Signals $\{S_i(t)\}$

In the VEVO case study, we consider a different exogenous signal per artist i, given by the Google Trends time series $\{S_i(t)\}$. This leads to changes in Eq. (2.5) and Eq. (2.10), since these equations are formulated with a single artist-independent $S(t)$.

Adapting Eq. (2.5) to the case of multiple exogenous signals $\{S_i(t)\}$, we have

$$\lambda^p(t) = \sum_i \mu_i^p \cdot S_i(t) + \sum_{q=1}^{P}\sum_{s<t} \alpha^{pq} \cdot f(t-s) \cdot N^q(s).$$

Adapting Eq. (2.10), we have

$$\lambda^p(t|j) = \mu_j^p \cdot S_j(t) + \sum_{q=1}^{P}\sum_{s<t} \alpha^{pq} \cdot f(t-s) \cdot N_j^q(s).$$

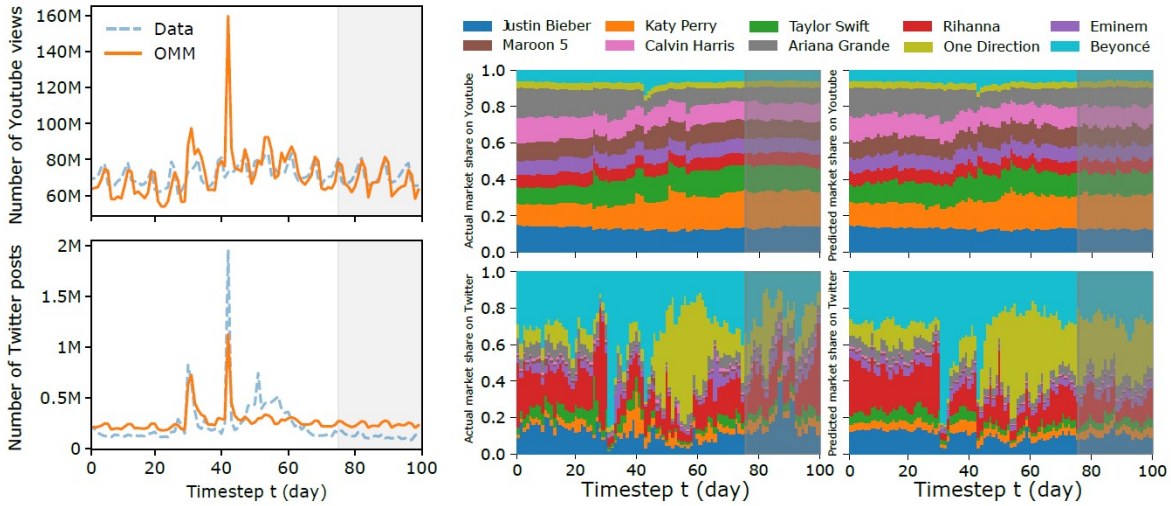
These modifications lead to changes in the structure of the two-tier optimization developed in Section 2.4, since the first-tier parameter set $\mathbf{\Theta}_1$ (originally $\{\mu^p, \alpha^{pq}, \theta\}$) now has to include artist-specific parameters $\{\mu_i^p\}$ due to the new form of $\lambda^p(t)$ above. Our new first-tier parameter set $\mathbf{\Theta}_1$ becomes $\{\mu_i^p, \alpha^{pq}, \theta\}$, while the new second-tier parameter set $\mathbf{\Theta}_2$ becomes $\{\gamma_{ik}^p, \beta_{ij}^{pq}\}$. We add an L2 regularizer on the first-tier optimization and an L1 regularizer on the second-tier optimization to prevent overfitting; we set the regularization parameters to 100.

Furthermore, our gradients also change. For the first-tier likelihood gradients, Eqs. (A.5) and (A.7) to (A.9) are still valid, and we replace Eq. (A.6) with

$$\partial_{\mu_i^q} \lambda^p(t) = \delta_{pq} \cdot S_i(t),$$

since we now estimate $\{\mu_i^p\}$ instead of $\{\mu^p\}$ in the first-tier optimization.

For the second-tier likelihood gradients, Eqs. (A.14) and (A.16) to (A.20) are still valid, but we do not anymore use Eq. (A.15) since we do not estimate $\{\mu_i^p\}$ in the second tier.



((a)) ((b))

Figure A.3: Fitting and predicting with OMM on the VEVO 2017 Top 10 dataset. We train OMM on the first 75 days and predict on days 76 to 100 (shaded area). We show results for Youtube and Twitter, respectively. (a) Actual (dashed blue lines) vs. fitted/predicted (orange lines) volumes; (b) Actual (left panels) and fitted/predicted (right panels) opinion market shares on Youtube (top panels) and Twitter (bottom panels)

A.5 OMM Fits and Predictions on VEVO 2017 Top 10

Fig. A.3 shows the fit and prediction of OMM on the VEVO 2017 Top 10 dataset for the first tier (attention volumes) and second tier (opinion market shares).

A.6 Model Elasticities

A.6.1 Intervention Elasticities $e(s_i^p(t), \bar{X}_k(t))$

Applying Eq. (2.4) on Eq. (2.7) and Eq. (A.21), we obtain the intervention elasticities $e(s_i^p(t), \bar{X}_k(t))$ as follows.

$$
\begin{aligned}
& e(s_i^p(t), \bar{X}_k(t)) \\
& = \partial_{\bar{X}_k(t)} s_i^p(t) \cdot \frac{\bar{X}_k(t)}{s_i^p(t)} \\
& = \left\{ -\frac{\mathscr{A}_i^p(t)}{\left[\sum_j \mathscr{A}_j^p(t)\right]^2} \partial_{\bar{X}_k(t)} \sum_j \mathscr{A}_j^p(t) + \frac{\partial_{\bar{X}_k(t)} \mathscr{A}_i^p(t)}{\sum_j \mathscr{A}_j^p(t)} \right\} \cdot \frac{\bar{X}_k(t)}{s_i^p(t)} \\
& = \left\{ -\frac{s_i^p(t)}{\sum_j \mathscr{A}_j^p(t)} \sum_j \mathscr{A}_j^p(t) \frac{\gamma_{jk}^p}{\sigma_{X,k}} + \frac{\mathscr{A}_i^p(t)}{\sum_j \mathscr{A}_j^p(t)} \cdot \frac{\gamma_{ik}^p}{\sigma_{X,k}} \right\} \cdot \frac{\bar{X}_k(t)}{s_i^p(t)} \\
& = \left\{ -\frac{s_i^p(t)}{\sum_j \mathscr{A}_j^p(t)} \sum_j \mathscr{A}_j^p(t) \frac{\gamma_{jk}^p}{\sigma_{X,k}} + s_i^p(t) \cdot \frac{\gamma_{ik}^p}{\sigma_{X,k}} \right\} \cdot \frac{\bar{X}_k(t)}{s_i^p(t)} \\
& = \left\{ -s_i^p(t) \sum_j \left[\frac{\mathscr{A}_j^p(t)}{\sum_l \mathscr{A}_l^p(t)} \right] \frac{\gamma_{jk}^p}{\sigma_{X,k}} + s_i^p(t) \cdot \frac{\gamma_{ik}^p}{\sigma_{X,k}} \right\} \cdot \frac{\bar{X}_k(t)}{s_i^p(t)} \\
& = \left\{ -s_i^p(t) \sum_j s_j^p(t) \frac{\gamma_{jk}^p}{\sigma_{X,k}} + s_i^p(t) \cdot \frac{\gamma_{ik}^p}{\sigma_{X,k}} \right\} \cdot \frac{\bar{X}_k(t)}{s_i^p(t)} \\
& = \left\{ -\sum_j s_j^p(t) \frac{\gamma_{jk}^p}{\sigma_{X,k}} + \frac{\gamma_{ik}^p}{\sigma_{X,k}} \right\} \cdot \bar{X}_k(t) \\
& = \frac{\bar{X}_k(t)}{\sigma_{X,k}} \cdot \sum_j \left[\delta_{ij} - s_j^p(t) \right] \gamma_{jk}^p
\end{aligned}
$$

Time-averaged intervention elasticities for the bushfire case study are shown in Fig. A.4.

A.6.2 Endogenous Elasticities $e(s_i^p(t), \lambda^q(t|j))$

Applying Eq. (2.4) on Eq. (2.7) and Eq. (A.21), we obtain the endogenous elasticities $e(s_i^p(t), \lambda^q(t|j))$ as follows. Let

$$
\phi_j^q(t) = \frac{1}{\mathrm{std}_s \log\left[N_j^q(s)+1\right]} \cdot \frac{1}{\lambda^q(t|j)+1}.
$$

We have:

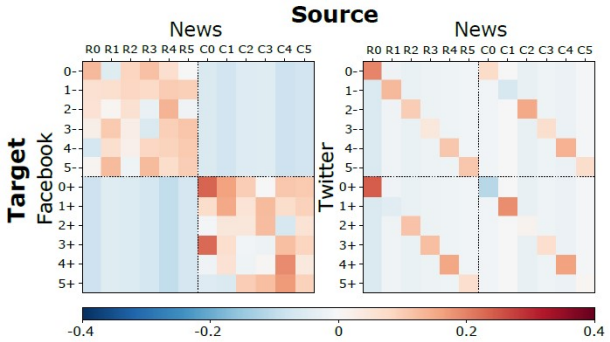


Figure A.4: Time-averaged intervention elasticities $e(s_i^p(t), \bar{X}_k(t))$ for the bushfire case study. Elasticities have direction and should be read from column (source) to row (target). The matrix on the left (right) corresponds to influences from reputable (R) and controversial (C) news for each opinion (in $\{0,1,2,3,4,5\}$) on the different stanced opinions on Facebook (Twitter).

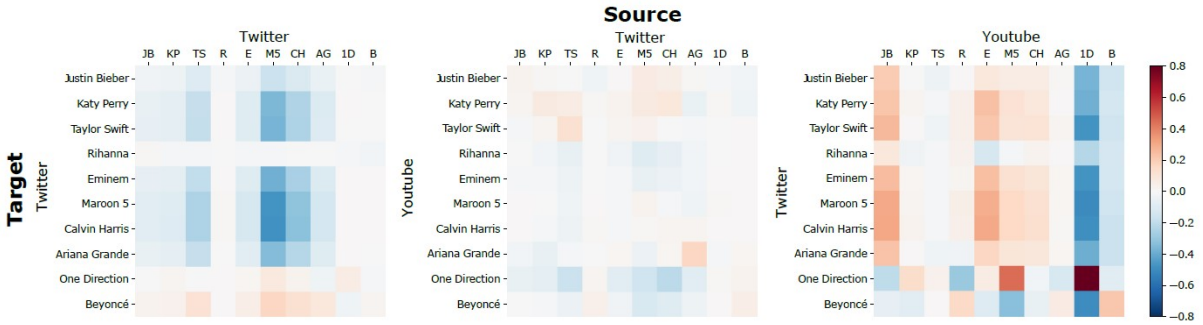


Figure A.5: Time-averaged endogenous elasticities $e(s_i^p(t), \lambda^q(t|j))$ of OMM in the VEVO case study. (Left) Twitter-to-Twitter elasticities. (Middle) Twitter-to-Youtube elasticities. (Right) Youtube-to-Twitter elasticities. Elasticities have direction and should be read from column (source) to row (target), both for the platform and within each color matrix.

$$
\begin{aligned}
& e(s_i^p(t), \lambda^q(t|j)) \\
&= \partial_{\lambda^q(t|j)} s_i^p(t) \cdot \frac{\lambda^q(t|j)}{s_i^p(t)} \\
&= \left\{ -\frac{\mathscr{A}_i^p(t)}{\left[\sum_k \mathscr{A}_k^p(t)\right]^2} \partial_{\lambda^q(t|j)} \sum_k \mathscr{A}_k^p(t) + \frac{\partial_{\lambda^q(t|j)} \mathscr{A}_i^p(t)}{\sum_k \mathscr{A}_k^p(t)} \right\} \cdot \frac{\lambda^q(t|j)}{s_i^p(t)} \\
&= \left\{ -\frac{s_i^p(t)}{\sum_k \mathscr{A}_k^p(t)} \sum_k \mathscr{A}_k^p(t) \frac{\beta_{kj}^{pq}}{\phi_j^q(t)} + \frac{\mathscr{A}_i^p(t)}{\sum_k \mathscr{A}_k^p(t)} \cdot \frac{\beta_{ij}^{pq}}{\phi_j^q(t)} \right\} \cdot \frac{\lambda^q(t|j)}{s_i^p(t)} \\
&= \left\{ -\frac{s_i^p(t)}{\sum_k \mathscr{A}_k^p(t)} \sum_k \mathscr{A}_k^p(t) \frac{\beta_{kj}^{pq}}{\phi_j^q(t)} + s_i^p(t) \cdot \frac{\beta_{ij}^{pq}}{\phi_j^q(t)} \right\} \cdot \frac{\lambda^q(t|j)}{s_i^p(t)} \\
&= \left\{ -s_i^p(t) \sum_k \left[\frac{\mathscr{A}_k^p(t)}{\sum_l \mathscr{A}_l^p(t)} \right] \frac{\beta_{kj}^{pq}}{\phi_j^q(t)} + s_i^p(t) \cdot \frac{\beta_{ij}^{pq}}{\phi_j^q(t)} \right\} \cdot \frac{\lambda^q(t|j)}{s_i^p(t)} \\
&= \left\{ -s_i^p(t) \sum_k s_k^p(t) \frac{\beta_{kj}^{pq}}{\phi_j^q(t)} + s_i^p(t) \cdot \frac{\beta_{ij}^{pq}}{\phi_j^q(t)} \right\} \cdot \frac{\lambda^q(t|j)}{s_i^p(t)} \\
&= \left\{ -\sum_k s_k^p(t) \frac{\beta_{kj}^{pq}}{\phi_j^q(t)} + \frac{\beta_{ij}^{pq}}{\phi_j^q(t)} \right\} \cdot \lambda^q(t|j) \\
&= \frac{\lambda^q(t|j)}{\phi_j^q(t)} \cdot \sum_k \left[\delta_{ki} - s_k^p(t) \right] \beta_{kj}^{pq}
\end{aligned}
$$

Time-averaged endogenous Twitter-to-Twitter and cross-platform elasticities for the VEVO case study are shown in Fig. A.5.

A.7 Bushfire Opinions Dataset Construction

The *Bushfire Opinions dataset* consists of Twitter posts and Facebook posts & comments from Australian user accounts and pages expressing problematic opinions on climate change and the 2019-2020 Australian bushfire season during the 90-day period of November 1, 2019 to January 29, 2020.

The Bushfire Opinions dataset derives from the *SocialSense dataset* introduced in [61], which consists of user posts and comments from three major online social media platforms: Facebook, Twitter and Youtube. Postings included in the SocialSense were on two general topics – first, the Australian bushfires and climate change, and second, Covid-19 and vaccination – and expressed problematic opinions. In this work, we focus on Facebook/ Twitter and the Australian bushfires/ climate change topic. Postings were collected using Crowdtangle focused on a set of far-right Australian Facebook groups identified with a digital ethnographic study (for Facebook), the Twitter commercial API (for Twitter), and the Youtube API (for Youtube) using the following keywords as input: *bushfire, australian fires, arson, scottyfrommarketing, liarfromtheshiar, australiaburns, australiaburning, itsthegreensfault, backburning, back burning, climate change, climate emergency, climate hoax, climate crisis, climate action now*. It is important to point out that the Facebook sample is sourced predemoninantly from far-right groups, whereas the Twitter and Youtube are general scrapes. Two sets of augmentations were added to the postings: the *topic* and the *opinion* of the post, obtained using a set of topic and opinion classifiers trained in [61]. The set of opinions were constructed via a qualitative study.

A limitation of the original SocialSense dataset is that the Twitter dataset for the Australian bushfires/ climate change topic was scraped only from December 2019 to February 2020, which did not capture early opinion during the start of the bushfire crisis. To that end, we decided to rescrape the Twitter dataset from November 1, 2019 to January 29, 2020 using the Twitter Academic v2 API and the same set of keywords. Since the Twitter Academic API does not allow querying based on user account location, we utilized AWS's Amazon Location Service to geocode users based on their free-text location and description fields and filtered only for tweets from Australian users. Finally, we applied the same set of topic and opinion classifiers to augment the Twitter data.

Once we aligned the Facebook dataset from SocialSense and the rescraped Twitter dataset on the target timeframe, we observed that 10 (out of 34) opinions account for most of the Twitter (95%) and Facebook (81%) postings. To limit the set of opinions in our analysis, we focus on six *opinions of interest* constructed by merging subsets of the 10 opinions, after

which we then filter the Twitter and Facebook datasets on this set of opinions. We index the six opinions we consider as $\{0,1,2,3,4,5\}$ and are shown below:

0. Greens influence and policy are the cause of the 2019-2020 Australian bushfires./ I am opposed to the policies of Greens political parties.
1. Mainstream media cannot be trusted.
2. Climate change crisis isn't real/ Climate change is a UN hoax/ Climate change is a scam to generate profit for the wealthy and powerful.
3. 2019-2020 Australian bushfires and climate change not related.
4. 2019-2020 Australian bushfires were caused by random arsonists.
5. Changes in the earth's climate are a natural, normal phenomenon/Bush fires are a normal summer occurrence for Australia.

Lastly, keeping in mind our goal of uncovering the interactions between sympathisers and opponents of the aforementioned problematic opinions, we furthermore differentiate whether the expressed opinion shows a *far-right* or *moderate* stance, which effectively splits our set of 6 opinions into 12 *stanced* opinions. For instance, the anti-Greens opinion (labeled 0) splits as far-right (labeled 0+) and moderate (labeled 0+). We represent our set of opinions as $\{(i-,i+)|i \in \{0,\ldots,5\}\}$. We leverage the far-right stance detector introduced by [96] and apply it on each post of the aligned Facebook and Twitter dataset.

In summary, the *Bushfire Opinions dataset* consists of posts on $P = 2$ platforms: 474,461 on Twitter and 27,974 on Facebook, exhibiting $M = 12$ stanced opinions. For compatibility with our discrete-time model, we aggregate post volumes on Facebook and Twitter into hourly counts, yielding $T = 2,160$ time points over the 90-day period of November 1, 2019 to January 29, 2020.

APPENDIX B

APPENDIX TO 'WHAT DRIVES ONLINE POPULARITY: AUTHOR, CONTENT OR SHARERS? ESTIMATING SPREAD DYNAMICS WITH BAYESIAN MIXTURE HAWKES'

B.1 Background Material

B.1.1 Hawkes Process

B.1.1.1 Inference

Given a cascade $\mathcal{H}$ of length N, i.e. an ordered collection of time stamps $\{t_i\}_{i=1}^N$ observed until some terminal time $T \geq t_N$, we can estimate the parameters of the Hawkes process $(\mu^*, \alpha^*, \boldsymbol{\Theta}^*)$ that generated the data by maximizing the log-likelihood function,

$$\mathscr{L}(\mu, \alpha, \boldsymbol{\Theta} | \mathcal{H} = \{t_j\}_{j=1}^N) = \sum_{j=1}^{N} \log \lambda(t_j; \mu, \alpha, \boldsymbol{\Theta}) - \int_0^T \lambda(s; \mu, \alpha, \boldsymbol{\Theta}) \mathrm{d}s. \tag{B.1}$$

This approach can be extended to the case of a collection of cascades $\mathbb{H} = \{\mathcal{H}_i\}$, where the best-fitting Hawkes process is obtained by maximizing the sum of the log-likelihood functions,

$$\mathscr{L}(\mu, \alpha, \boldsymbol{\Theta} | \mathbb{H}) = \sum_{\mathcal{H}_i \in \mathbb{H}} \mathscr{L}(\mu, \alpha, \boldsymbol{\Theta} | \mathcal{H}_i). \tag{B.2}$$

B.1.1.2 Prediction

The fitted Hawkes process can be leveraged to predict the cascade size $\hat{N}$ of a new cascade $\mathcal{H} \in \mathbb{H}$:

$$\hat{N} = \mathbb{E}[N|\alpha] = \frac{1}{1-\alpha}. \tag{B.3}$$

B.1.2 Dual Mixture Model

B.1.2.1 Inference

Given the pre-defined number of components K, we obtain the Borel mixture model M^B by maximizing the following log-likelihood function,

$$\mathscr{L}_{BMM} = \sum_{\mathcal{H}_i \in \mathbb{H}} \log \sum_{k=1}^{K} p_k^B \mathbb{B}(N_i | \alpha_k^*). \tag{B.4}$$

Note that the DMM is not formulated as a Bayesian model in [63] and the Expectation-Maximization (EM) algorithm [117] is employed to maximize $\mathscr{L}_{BMM}$. Similarly, the kernel mixture model M^g is obtained by applying the EM algorithm to the kernel log-likelihood

$$\mathscr{L}_{KMM} = \sum_{\mathcal{H}_i \in \mathbb{H}} \log \sum_{k=1}^{K} p_k^g f^g(\mathcal{H}_i | \boldsymbol{\Theta}_k^*), \tag{B.5}$$

where $f(\mathcal{H}_i|\boldsymbol{\Theta}) = \prod_{t_j \in \mathcal{H}_i} \sum_{t_z < t_j} g(t_j - t_z|\boldsymbol{\Theta})$.

B.1.2.2 Cold Start Popularity Prediction

Assume that we are given a collection of related cascade groups $\{\mathbb{H}_a\}_{a\in\mathscr{A}}$. For instance, suppose $\mathscr{A}$ is a set of news articles from a common online publisher ρ and $\mathbb{H}_a$ is the set of retweet cascades discussing article a. Give a yet-to-be-published article $a^* \notin \mathscr{A}$, we wish to model its popularity $\hat{N}^{a^*}$ by learning from historical data $\{\mathbb{H}_a\}_{a\in\mathscr{A}}$.

To do this, we can construct a publisher-level popularity model M_ρ^B by fitting an independent BMM M_a^B (with K_a classes) to each $\mathbb{H}_a$ and then collecting these as a mixture M_ρ^B over $\mathscr{A}$, i.e.,

$$M_\rho^B = \bigcup_{a\in\mathscr{A}} M_a^B = \bigcup_{a\in\mathscr{A}} \left\{ \left(\alpha_1^a, \frac{p_1^{B,a}}{|\mathscr{A}|}\right), \ldots, \left(\alpha_{K_a}^a, \frac{p_{K_a}^{B,a}}{|\mathscr{A}|}\right) \right\}, \tag{B.6}$$

where $(\alpha_i^a, p_i^{B,a}) \in M_a^B$. We can estimate the cold-start popularity of a new article a^* as

$$\hat{N}^{a^*} = \hat{C}_\rho \cdot \mathbb{E}_{M_\rho^B}\left[\frac{1}{1-\alpha}\right] = \hat{C}_\rho \cdot \sum_{a\in\mathscr{A}} \sum_{i=1}^{K_a} \frac{1}{1-\alpha_i^a} \cdot \frac{p_i^{B,a}}{|\mathscr{A}|}, \tag{B.7}$$

where $\hat{C}_\rho$ is an estimate of the cascade count of article a^*, which we can take as the average cascade count of articles in $\mathscr{A}$.

B.2 Additional Material for BMH Formulation

B.2.1 Complete Table of Notation

In Table B.1 we show the full set of notation, BMH model parameters, their interpretation and real-world mapping.

B.2.2 BMH-P Model

B.2.2.1 Assumptions

In Eqs. (3.3) and (3.4) we assume that the item-level features $\vec{y}^a$ influence the location (i.e. mean) and membership probability of each popularity class k, while the cascade-level features $\vec{x}^{ac}$ influence only the membership probability. As a concrete example, if we have two popularity classes (popular and unpopular), $\mathscr{A}$ being a set of articles, $\vec{y}^a$ being the headline embedding vector of article a, and $\vec{x}^{ac}$ the follower count of the initiator of cascade c, our assumptions imply how large a cascade will turn out to be (Eq. (3.3)) is influenced only by article content $\vec{y}^a$, but whether a cascade will be popular or not is influenced by both article content $\vec{y}^a$ and follower count $\vec{x}^{ac}$.

B.2.2.2 Likelihood Function

The log-likelihood of $\mathscr{P}_\alpha$ given the set of cascade sizes $\{N_{ac}\}_{ac}$ can be derived as:

$$
\begin{aligned}
\mathscr{L}(\mathscr{P}_\alpha|\{N_{ac}\}_{ac}) &= \log \mathbb{P}(\{N_{ac}\}_{ac}|\mathscr{P}_\alpha) \\
&= \log \prod_{a\in\mathscr{A}} \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a} \mathbb{P}(N_{ac}|\mathscr{P}_\alpha) \\
&= \log \prod_{a\in\mathscr{A}} \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a} \mathbb{P}(N_{ac}|\alpha^{ac}) \\
&\overset{(a)}{=} \log \prod_{a\in\mathscr{A}} \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a} \mathbb{B}(N_{ac}|\alpha^{ac}) \\
&\overset{(b)}{=} \log \prod_{a\in\mathscr{A}} \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a} \sum_{k=1}^{K_\alpha} z^{ac}_{\alpha,k} \cdot \mathbb{B}(N_{ac}|\text{inv-logit}(\delta^a_{\alpha,k} + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^a)) \\
&= \sum_{\mathscr{H}^{ac}\in\mathbb{H}^a, a\in\mathscr{A}} \log \sum_{k=1}^{K_\alpha} z^{ac}_{\alpha,k} \cdot \mathbb{B}(N_{ac}|\text{inv-logit}(\delta^a_{\alpha,k} + \vec{\gamma}_{\alpha,k}\cdot\vec{y}^a)),
\end{aligned}
$$

where in (a) we use the fact that the cascade size of a Hawkes process is Borel-distributed with parameter α^{ac} and in (b) we note that the BMH-P model specifies a^{ac} as a mixture over the K_α classes, weighted by the membership probabilities $\{z^{ac}_{\alpha,k}\}$.

Table B.1: Full table of notation in Chapter 3.

Parameter	Interpretation	Real-World Mapping
Source-Level		
ρ	source of items	news publisher
$\mathscr{A}$	set of items produced by ρ	news articles from publisher ρ
$f_\rho(\cdot)$	follower count distribution	
$\hat{C}_\rho$	cascade count estimate	
Item-Level		
$a \in \mathscr{A}$	item produced by ρ	news article
$\mathbb{H}^a$	set of cascades related to item a	retweet cascades for article a
$\vec{y}^a$	item-level features	headline embedding for article a
N^a	item popularity	overall tweet count for article a
$\tau^a_{1/2}$	content half-life	
Cascade-Level		
$\mathscr{H}^{ac} \in \mathbb{H}^a$	cascade related to item a	retweet cascade for article a
$\vec{x}^{ac}$	cascade-level features	follower count of seed user
N^{ac}	cascade size	
$\tau^{ac}_{1/2}$	cascade half-life	
$\mathscr{T}^{ac}$	intereevent-time distribution	
α^{ac}	Hawkes branching factor	
$\mathbf{\Theta}^{ac}$	Hawkes kernel parameters	
BMH-P		
K_α	# of BMH-P mixture classes	
$\vec{\gamma}_{\alpha,k}$	effect of $\vec{y}^a$ on center of class k	
$\vec{\gamma}_{z_{\alpha,k}}$	effect of $\vec{y}^a$ on membership probability of class k	
$\delta^a_{\alpha,k}/\delta_{\alpha,k}$	item-/ publisher-level baseline value of logit(α) for class k	
$\vec{\beta}^a_{\alpha,k}/\vec{\beta}_{\alpha,k}$	effect of $\vec{x}^{ac}$ on center of class k	
$z^{ac}_{\alpha,k}$	mem. probability for class k	
$\delta^a_{z_{\alpha,k}}/\delta_{z_{\alpha,k}}$	item-/ publisher-level mem. prob. softmax baseline for class k	
$\vec{\beta}^a_{z_{\alpha,k}}/\vec{\beta}_{z_{\alpha,k}}$	effect of $\vec{x}^{ac}$ on membership probability for class k	
$\vec{p}^a_\alpha/\vec{p}_\alpha$	item-/ pub.-level parameter vector	
$\Sigma_\alpha/\Omega_\alpha$	cov./ corr. matrix for $\vec{p}_\alpha$	
BMH-K		
$K_{\mathbf{\Theta}}$	# of BMH-K mixture classes	
$\vec{\gamma}_{\theta,k}$	effect of $\vec{y}^a$ on center of class k	
$\vec{\gamma}_{z_{\mathbf{\Theta},k}}$	effect of $\vec{y}^a$ on membership probability of class k	
$\delta^a_{\theta,k}/\delta_{\theta,k}$	item-/ publisher-level baseline value of log(θ) for class k	
$\vec{\beta}^a_{\theta,k}/\vec{\beta}_{\theta,k}$	effect of $\vec{x}^{ac}$ on center of class k	
$z^{ac}_{\mathbf{\Theta},k}$	mem. probability for class k	
$\delta^a_{z_{\mathbf{\Theta},k}}/\delta_{z_{\mathbf{\Theta},k}}$	item-/ publisher-level mem. prob. softmax baseline for class k	
$\vec{\beta}^a_{z_{\mathbf{\Theta},k}}/\vec{\beta}_{z_{\mathbf{\Theta},k}}$	effect of $\vec{x}^{ac}$ on membership probability for class k	
$\vec{p}^a_{\mathbf{\Theta},k}/\vec{p}_{\mathbf{\Theta},k}$	item-/ pub.-level kernel parameter baseline values for class k	
$\vec{p}^a_{z_{\mathbf{\Theta}}}/\vec{p}_{z_{\mathbf{\Theta}}}$	item-/ pub.-level membership probability parameters for class k	
$\Sigma_{\mathbf{\Theta},k}/\Omega_{\mathbf{\Theta},k}$	cov./ corr. matrix for $\vec{p}_{\mathbf{\Theta},k}$	
$\Sigma_{z_{\mathbf{\Theta}}}/\Omega_{z_{\mathbf{\Theta}}}$	cov./ corr. matrix for $\vec{p}_{z_{\mathbf{\Theta}}}$	

B.2.2.3 Cold-Start Popularity

First, from our dataset $\mathscr{A}$, compute $\hat{C}_\rho$ as the average cascade count for an article in $\mathscr{A}$ and $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^a)$ as the empirical probability density of the cascade feature vector $\vec{x}^{ac}$ given item feature vector $\vec{y}^a$.

Second, from Eq. (3.6) we draw the parameter set $\vec{p}_\alpha^{a*}$ for the out-of-sample item a^*. Consider an arbitrary cascade c of item a^* with feature vector $\vec{x}^{a^*c}$. The expected cascade size of c is given by the expectation of Eq. (3.3) over the K_α popularity classes $\mathbb{E}_{z_{\alpha,k}^{a^*c}}[\mathbb{E}[N^{a^*c}]|\vec{x}^{a^*c},\vec{y}^{a^*}]$. Since c is arbitrary, we need to average $\vec{x}^{a^*c}$ out. Hence, our expected cascade size is $\mathbb{E}_{\vec{x}^{a^*c}}\mathbb{E}_{z_{\alpha,k}^{a^*c}}\left[\mathbb{E}[N^{a^*c}]|\vec{x}^{a^*c},\vec{y}^{a^*}\right]$. Our estimate $\hat{N}^{a^*}$ of item a^*'s popularity is then given by

$$
\begin{aligned}
\hat{N}^{a^*} &= \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}}\mathbb{E}_{z_{\alpha,k}^{a^*c}}\left[\mathbb{E}[N^{a^*c}]|\vec{x}^{a^*c},\vec{y}^{a^*}\right] \\
&= \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}}\mathbb{E}_{z_{\alpha,k}^{a^*c}}\left[\frac{1}{1-\alpha^{a^*c}}\middle|\vec{x}^{a^*c},\vec{y}^{a^*}\right] \\
\overset{(a)}{=} \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}} &\sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \frac{1}{1-\text{inv-logit}\left(\delta_{\alpha,k}^{a^*}+\vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right)} \\
&= \hat{C}_\rho \cdot \mathbb{E}_{\vec{x}^{a^*c}} \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left[1+\exp\left(\delta_{\alpha,k}^{a^*}+\vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right)\right] \\
\overset{(b)}{=} \hat{C}_\rho \cdot &\int \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left[1+\exp\left(\delta_{\alpha,k}^{a^*}+\vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right)\right] \cdot f_\rho(\vec{x}^{a^*c}|\vec{y}^{a^*}) \cdot \mathrm{d}\vec{x}^{a^*c} \\
\overset{(c)}{\approx} \hat{C}_\rho \cdot &\int \sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left[1+\exp\left(\delta_{\alpha,k}^{a^*}+\vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right)\right] \cdot \hat{f}_\rho(\vec{x}^{a^*c}) \cdot \mathrm{d}\vec{x}^{a^*c},
\end{aligned}
$$

where in (a) we use the fact that the BMH-P model specifies a^{ac} as a mixture over the K_α classes, weighted by the membership probabilities $\{z_{\alpha,k}^{ac}\}$, in (b) we marginalize over the unobserved cascade-level features $\vec{x}^{a^*c}$ in a cold-start setup, and in (c) we use the simplification $f_\rho(\vec{x}^{ac}|\vec{y}^{a}) \approx \hat{f}_\rho(x)$ as detailed in Section 3.5.

To simplify this expression, we impose two additional assumptions on the feature vectors. First, assume our cascade feature vector is one-dimensional, discrete and nonnegative (for instance, this may be the follower count of the seed user). This simplifies our probability density $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^{a})$ into a probability mass function over $x \in \mathbb{N}\cup\{0\}$, converting the integral over $\vec{x}^{ac}$ into a sum. Second, in practice we usually will not have enough variance across $\vec{y}^{a}$ to build $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^{a})$ reliably, and so we assume that $\vec{x}^{ac}$ is independent of $\vec{y}^{a}$. These two assumptions allow us to write $\hat{f}_\rho(\vec{x}^{ac}|\vec{y}^{a}) \approx \hat{f}_\rho(x)$. Our expression simplifies to

$$
\hat{N}^{a^*} \approx \hat{C}_\rho \cdot \sum_{x=0}^{\infty}\sum_{k=1}^{K_\alpha} z_{\alpha,k}^{a^*,c} \cdot \left[1+\exp\left(\delta_{\alpha,k}^{a^*}+\vec{\gamma}_{\alpha,k}\cdot\vec{y}^{a^*}\right)\right] \cdot \hat{f}_\rho(x). \tag{B.8}
$$

B.2.3 BMH-K Model

B.2.3.1 Assumptions

In Eqs. (3.9) and (3.10), we assume that the item-level features $\vec{y}^a$ influence the location of θ^{ac} and not d^{ac}. We found that including influence of item-level features $\vec{y}^a$ on both parameters leads to identifiability issues in the BMH-K model. We assume $\vec{y}^a$ influence the location (i.e. mean) and membership probability of each popularity class k, while the cascade-level features $\vec{x}^{ac}$ influence only the membership probability. As a concrete example, if we have two kernel classes (slow and fast), $\mathscr{A}$ being a set of articles, $\vec{y}^a$ being the headline embedding vector of article a, and $\vec{x}^{ac}$ the follower count of the initiator of cascade c, our assumptions imply the speed at which a cascade will diffuse (Eq. (3.9)) is influenced only by article content $\vec{y}^a$, but whether a cascade will be slow or fast is influenced by both article content $\vec{y}^a$ and follower count $\vec{x}^{ac}$.

B.2.3.2 Likelihood Function

The log-likelihood of $\mathscr{P}_{\mathbf{\Theta}}$ given the set of interevent-time distributions $\{\mathscr{T}^{ac}\}_{ac}$ is

$$
\begin{aligned}
\mathscr{L}(\mathscr{P}_{\mathbf{\Theta}}|\{\mathscr{T}^{ac}\}_{ac}) &= \log \mathbb{P}(\{\mathscr{T}^{ac}\}_{ac}|\mathscr{P}_{\mathbf{\Theta}}) \\
&\overset{(a)}{=} \log \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a, a\in\mathscr{A}} \left[\prod_{t_j\in\mathscr{H}^{ac}, j\geq 1} \sum_{t_z<t_j} g(t_j - t_z|\theta^{ac}, d^{ac}) \right] \\
&\overset{(b)}{=} \log \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a, a\in\mathscr{A}} f(\mathscr{H}^{ac}|\theta^{ac}, d^{ac}) \\
&\overset{(c)}{=} \log \prod_{\mathscr{H}^{ac}\in\mathbb{H}^a, a\in\mathscr{A}} \left[\sum_{k=1}^{K_{\mathbf{\Theta}}} z^{ac}_{\mathbf{\Theta},k} \cdot f(\mathscr{H}^{ac}|e^{\delta^a_{\theta,k}+\vec{\gamma}_{\theta,k}\cdot\vec{y}^a}, e^{\delta^a_{d,k}}) \right] \\
&= \sum_{\mathscr{H}^{ac}\in\mathbb{H}^a, a\in\mathscr{A}} \log \sum_{k=1}^{K_{\mathbf{\Theta}}} z^{ac}_{\mathbf{\Theta},k} \cdot f(\mathscr{H}^{ac}|e^{\delta^a_{\theta,k}+\vec{\gamma}_{\theta,k}\cdot\vec{y}^a}, e^{\delta^a_{d,k}})
\end{aligned}
$$

where in (a) we make use of the likelihood for the interevent-time distribution for separable Hawkes processes as derived in [63], in (b) we set $f(\mathscr{H}|\theta,d) = \prod_{t_j\in\mathscr{H}} \sum_{t_z<t_j} g(t_j - t_z|\theta,d)$, and in (c) we use the fact that the BMH-K model specifies $\mathbf{\Theta}^{ac}$ as a mixture over the $K_{\mathbf{\Theta}}$ classes, weighted by the membership probabilities $\{z^{ac}_{\mathbf{\Theta},k}\}$.

B.2.3.3 Half-Life Prediction

Under the BMH-K model, the half-life of an out-of-sample item a^* can be expressed as

$$\begin{aligned}
\hat{\tau}_{1/2}^{a^*} &= \mathbb{E}_{\vec{x}^{a^*c}} \mathbb{E}_{z_{\mathbf{\Theta},k}^{a^*c}} \left[\tau_{1/2}^{a^*} | \vec{x}^{a^*c}, \vec{y}^{a^*} \right] \\
&\overset{(a)}{=} \mathbb{E}_{\vec{x}^{a^*c}} \mathbb{E}_{z_{\mathbf{\Theta},k}^{a^*c}} \left[d^{a^*c} \cdot (2^{\theta^{a^*c}} - 1) | \vec{x}^{a^*c}, \vec{y}^{a^*} \right] \\
&\overset{(b)}{=} \mathbb{E}_{\vec{x}^{a^*c}} \sum_{k=1}^{K_{\mathbf{\Theta}}} z_{\mathbf{\Theta},k}^{a^*,c} \cdot e^{\delta_{d,k}^{a^*}} \cdot \left[2^{\exp(\delta_{\theta,k}^{a^*} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^{a^*})} - 1 \right] \\
&\overset{(c)}{\approx} \sum_{x=0}^{\infty} \sum_{k=1}^{K_{\mathbf{\Theta}}} z_{\mathbf{\Theta},k}^{a^*,c} \cdot e^{\delta_{d,k}^{a^*}} \cdot \left[2^{\exp(\delta_{\theta,k}^{a^*} + \vec{\gamma}_{\theta,k} \cdot \vec{y}^{a^*})} - 1 \right] \cdot \hat{f}_{\rho}(x),
\end{aligned}$$

where in (a) we use the expression for the half-life of a Hawkes process under the power law $g(t) = \theta \cdot d^{\theta} \cdot (t+d)^{-(1+\theta)}$ (i.e. by solving $\tau_{1/2}$ such that $\int_0^{\tau_{1/2}} g(t)\,dt = \frac{1}{2}$), in (b) we use the fact that the BMH-K model specifies $\mathbf{\Theta}^{ac}$ as a mixture over the $K_{\mathbf{\Theta}}$ classes, weighted by the membership probabilities $\{z_{\mathbf{\Theta},k}^{ac}\}$, and in (c) we marginalize over the unobserved cascade-level features $\vec{x}^{a^*c}$ and use the simplification $f_{\rho}(\vec{x}^{ac} | \vec{y}^{a}) \approx \hat{f}_{\rho}(x)$ as detailed in Section 3.5.

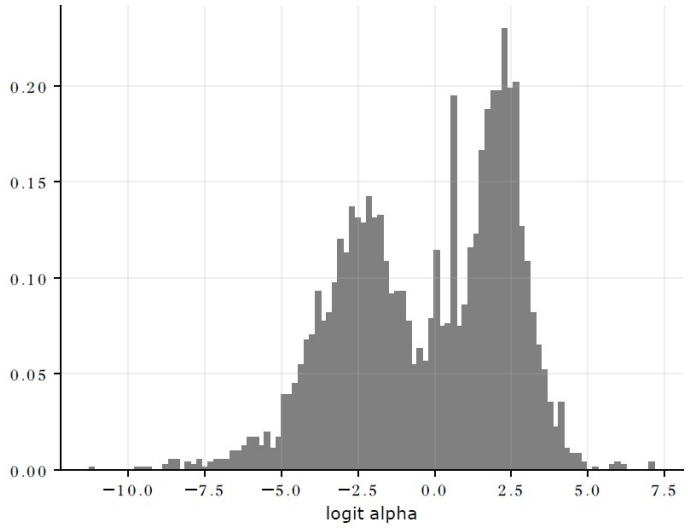


Figure B.1: Distribution of DMM-estimated logit(α) across $RNIX$ publishers. We note the bimodality of the distribution, with the modes corresponding to low and high cascade sizes. Based on this observation we set $K_\alpha = 2$ for the BMH-P model.

B.3 Additional Material for BMH Evaluation

B.3.1 Selection of K_α and $K_{\mathbf{\Theta}}$

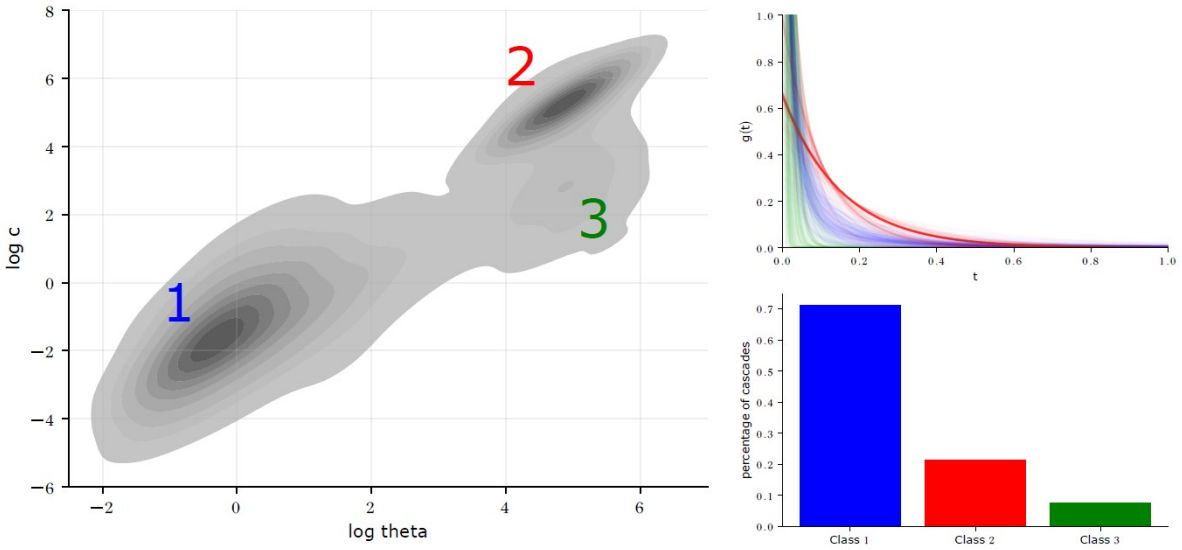
To guide the selection of the number of mixture components for the BMH-P (i.e. K_α) and BMH-K (i.e. $K_{\mathbf{\Theta}}$) models in Section 3.5, we fit the DMM [63] to each publisher in $RNIX$. Given that the EM algorithm is very sensitive to initial conditions, we use 10 random EM initializations and select the output that yields the highest log-likelihood.

We collect the distribution of parameter estimates for logit(α) across publishers in Fig. B.1. We see two modes for α, corresponding to cascade groups with low and high sizes, prompting us to set $K_\alpha = 2$ in Section 3.5.

We collect the distribution of parameter estimates for $(\log(c), \log(\theta))$ across publishers in the upper plot of Fig. B.2, where we see three modes for the kernel parameters. From the lower plot of Fig. B.2, we can interpret these modes as belonging to usual, fast and slow cascade groups, prompting us to set $K_{\mathbf{\Theta}} = 3$ in Section 3.5.

B.3.2 Prior Specification for the BMH-P Model

The full set of priors for the BMH-P model implementation is given below. Informative priors are set for $\delta_{\alpha,1}, \delta_{\alpha,2}, \delta_{z_{\alpha,2}}$ based on the observations in Appendix B.3.1. Weakly informative



((a)) ((b))

Figure B.2: (a) Distribution of DMM-estimated $(\log(c), \log(\theta))$ across $RNIX$ publishers. We observe the trimodality of the distribution, with the modes corresponding to usual (labeled 1), slow (labeled 2) and fast (labeled 3) cascades. Based on this observation we set $K_{\boldsymbol{\Theta}} = 3$ for the BMH-K model. (b) In the top plot, we show samples of the power law kernel g for the three classes. In the bottom plot, we show the distribution of cascades for each class.

priors are set for the other parameters. We use a Laplace prior on $\vec{\gamma}_{\alpha,1}, \vec{\gamma}_{\alpha,2}, \vec{\gamma}_{z_{\alpha,2}}$ to impose regularization given the high dimensionality of the article feature vector ($|\vec{y}^a| = 32$) we consider.

$$
\begin{aligned}
\delta_{\alpha,1} &\sim \mathcal{N}(-2, 0.5) \\
\delta_{\alpha,2} &\sim \mathcal{N}(2, 0.5) \\
\delta_{z_{\alpha,2}} &\sim \mathcal{N}(-1.39, 0.5) \\
\vec{\beta}_{z_{\alpha,2}} &\sim \mathcal{N}(0, 0.1) \\
\vec{\gamma}_{\alpha,1}, \vec{\gamma}_{\alpha,2}, \vec{\gamma}_{z_{\alpha,2}} &\sim \text{Laplace}(0, 0.01) \\
\Omega_\alpha &\sim \text{LKJCorr}(2) \\
\sigma_{\delta_{\alpha,1}}, \sigma_{\delta_{\alpha,2}}, \sigma_{\delta_{z_{\alpha,2}}} &\sim \mathcal{N}(0, 1) \\
\sigma_{\vec{\beta}_{z_{\alpha,2}}} &\sim \mathcal{N}(0, 0.1)
\end{aligned}
$$

B.3.3 Prior Specification for the BMH-K Model

The full set of priors for the BMH-K model implementation is given below. Informative priors are set for $\delta_{\theta,1}, \delta_{d,1}, \delta_{\theta,2}, \delta_{d,2}, \delta_{\theta,3}, \delta_{d,3}, \delta_{z_{\mathbf{\Theta},2}}, \delta_{z_{\mathbf{\Theta},3}}$ based on the observations in Appendix B.3.1. Weakly informative priors are set for the other parameters. We use a Laplace prior on $\vec{\gamma}_{\mathbf{\Theta},2}, \vec{\gamma}_{\mathbf{\Theta},3}, \vec{\gamma}_{z_{\mathbf{\Theta},2}}, \vec{\gamma}_{z_{\mathbf{\Theta},3}}$ to impose regularization given the high dimensionality of the article feature vector ($|\vec{y}^a| = 32$) we consider. For $\Omega_{\mathbf{\Theta},1}, \Omega_{\mathbf{\Theta},2}, \Omega_{\mathbf{\Theta},3}$, we set a LKJCorr(0.5) prior (i.e. higher weights on the tails of [0,1]) as (θ, d) for any given Hawkes fit are correlated.

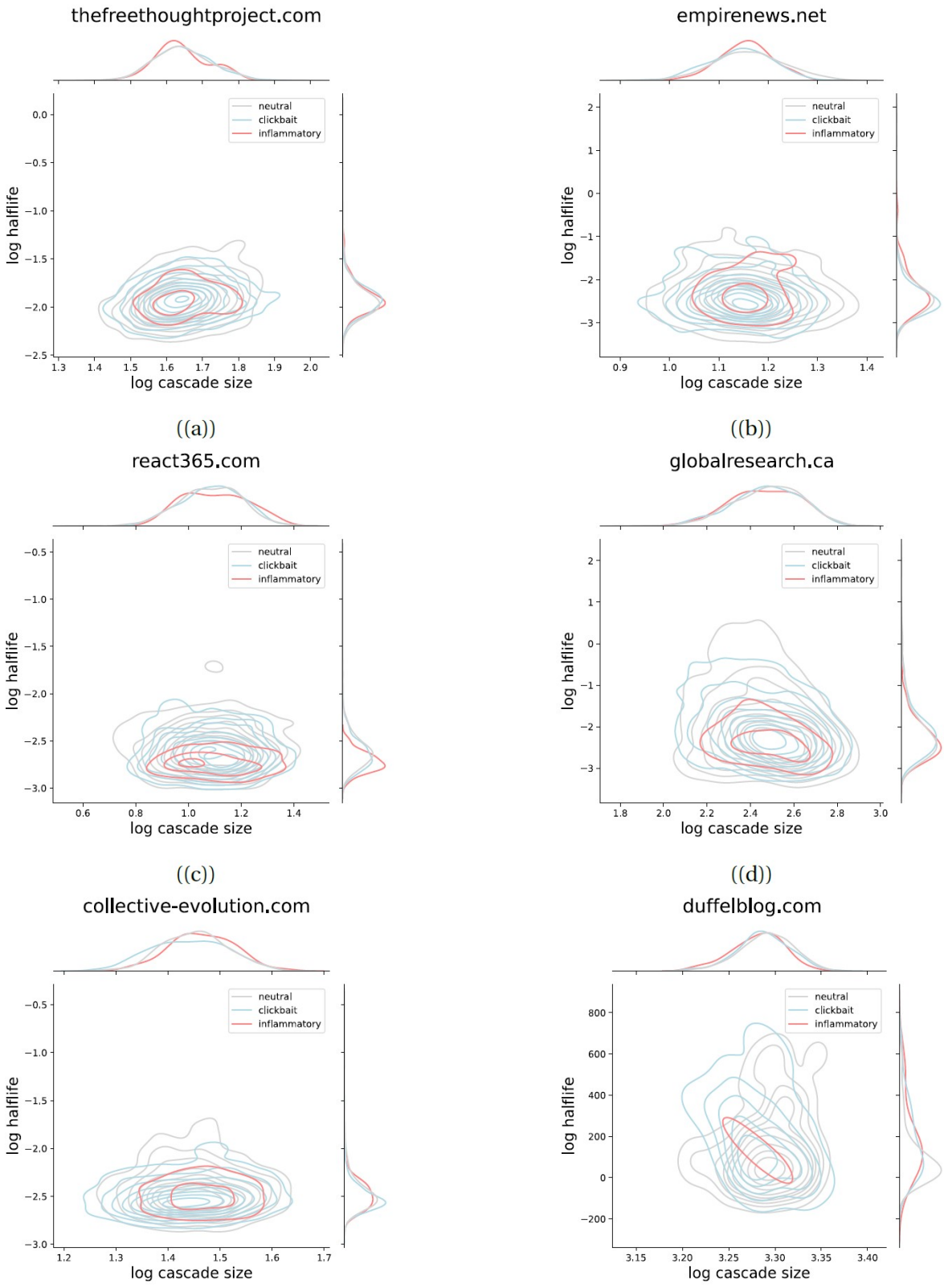
$$
\begin{aligned}
\delta_{\theta,1} &\sim \mathcal{N}(-0.41, 0.5) \\
\delta_{d,1} &\sim \mathcal{N}(-1.37, 1) \\
\delta_{\theta,2} &\sim \mathcal{N}(4, 0.5) \\
\delta_{d,2} &\sim \mathcal{N}(4.805, 0.5) \\
\delta_{\theta,3} &\sim \mathcal{N}(4, 0.5) \\
\delta_{d,3} &\sim \mathcal{N}(1, 0.5) \\
\delta_{z_{\mathbf{\Theta},2}}, \delta_{z_{\mathbf{\Theta},3}} &\sim \mathcal{N}(-2, 1) \\
\vec{\beta}_{z_{\mathbf{\Theta},2}}, \vec{\beta}_{z_{\mathbf{\Theta},3}} &\sim \mathcal{N}(0, 0.1) \\
\vec{\gamma}_{\mathbf{\Theta},2}, \vec{\gamma}_{\mathbf{\Theta},3}, \vec{\gamma}_{z_{\mathbf{\Theta},2}}, \vec{\gamma}_{z_{\mathbf{\Theta},3}} &\sim \text{Laplace}(0, 0.01) \\
\Omega_{\mathbf{\Theta},1}, \Omega_{\mathbf{\Theta},2}, \Omega_{\mathbf{\Theta},3} &\sim \text{LKJCorr}(0.5) \\
\Omega_{z_{\mathbf{\Theta}}} &\sim \text{LKJCorr}(2) \\
\sigma_{\delta_{\theta,1}}, \sigma_{\delta_{\theta,2}}, \sigma_{\delta_{\theta,3}}, \sigma_{\delta_{d,1}}, \sigma_{\delta_{d,2}}, \sigma_{\delta_{d,3}}, \sigma_{\delta_{z_{\mathbf{\Theta},2}}}, \sigma_{\delta_{z_{\mathbf{\Theta},3}}} &\sim \mathcal{N}(0, 1) \\
\sigma_{\vec{\beta}_{z_{\mathbf{\Theta},2}}}, \sigma_{\vec{\beta}_{z_{\mathbf{\Theta},3}}} &\sim \mathcal{N}(0, 0.1)
\end{aligned}
$$

B.3.4 Implementation Details

We use the Python implementation of Stan [15] to run both the BMH-P and BMH-K models. We run for 4 chains, adapt delta set to 0.9, 500 warmup iterations and 500 post-warmup iterations. To speed up convergence we implement non-centered parametrization [86] for each of the normally distributed priors.

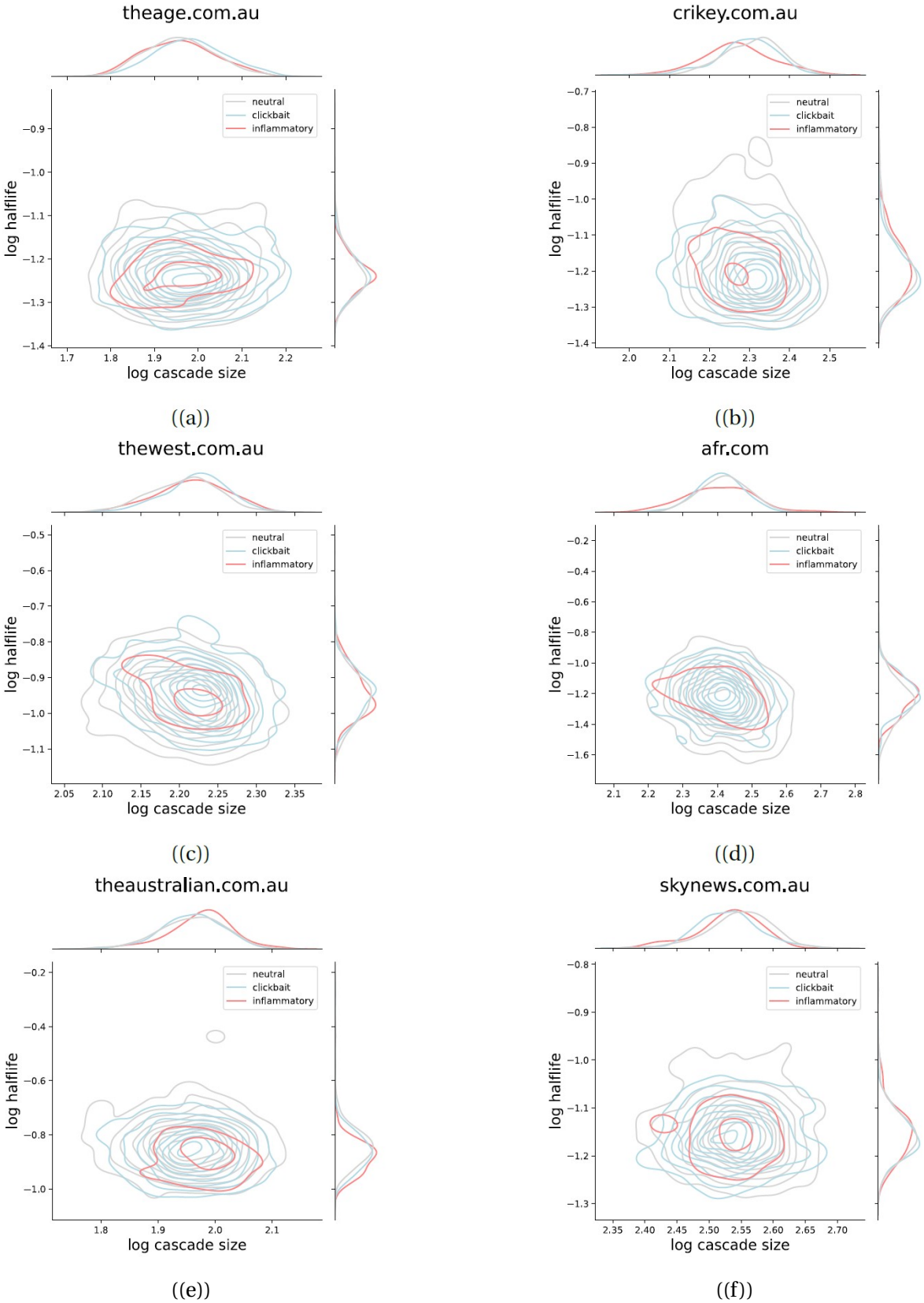
B.4 Performance Heatmaps for CNIX and RNIX

We show performance heatmaps for a selection of CNIX publishers in Fig. B.3 and RNIX publishers in Fig. B.4. Note the different patterns of which headline style works for each publisher, implying that the BMH model picks up subtle differences of what is effective across publishers.



((a)) ((b)) ((c)) ((d)) ((e)) ((f))

Figure B.3: Performance heatmaps for a selection of CNIX publishers.



((a)) ((b)) ((c)) ((d)) ((e)) ((f))

Figure B.4: Performance heatmaps for a selection of RNIX publishers.

APPENDIX C

APPENDIX TO 'LINKING ACROSS DATA GRANULARITY: FITTING MHP TO PARTIALLY INTERVAL-CENSORED DATA'

C.1 Background Material

C.1.1 Multivariate Hawkes Process

Alternative view of conditional intensity. The conditional intensity can be viewed as the mean number of events occurring in an infinitesimal interval, conditioned on the past. A simple multivariate extension of the result in [97] gives

$$\boldsymbol{\lambda}^{\star}(t)\mathrm{d}t = \mathbb{E}\left[\mathrm{d}\mathbf{N}(t)|\mathcal{H}_{t^-}^{D}\right]. \tag{C.1}$$

Compensator. By integrating the conditional intensity, we obtain another important measure: the compensator $\boldsymbol{\Lambda}(t)$ of the process.

Definition C.1. Given a temporal point process with conditional intensity $\boldsymbol{\lambda}^{\star}(t)$, the compensator $\boldsymbol{\Lambda}(t)$ is defined as

$$\boldsymbol{\Lambda}(t) = \int_0^t \boldsymbol{\lambda}^{\star}(\tau)\mathrm{d}\tau. \tag{C.2}$$

where $0 \leq s \leq t$.

Proposition C.1. *The compensator $\boldsymbol{\Lambda}(t)$ can be interpreted as the expected number of events over $[0,t)$ given $\mathcal{H}_{t^-}^{D}$. This follows by integrating Eq.* (C.1) *over* $[0,t)$,

By integrating $\boldsymbol{\lambda}^{\star}(t)$, we obtain an explicit form for the compensator $\boldsymbol{\Lambda}(t)$ of the d-dimensional Hawkes process:

$$\boldsymbol{\Lambda}(t) = \mathbf{M}(t) + \sum_{j=1}^{d} \sum_{t_k^j < t} \boldsymbol{\Phi}^j(t - t_k^j), \tag{C.3}$$

where

$$\mathbf{M}(t) = \int_0^t \boldsymbol{\mu}(s)\mathrm{d}s, \tag{C.4}$$

$$\boldsymbol{\Phi}(t) = \int_0^t \boldsymbol{\varphi}(s)\mathrm{d}s. \tag{C.5}$$

Regularity condition. A univariate Hawkes process is *subcritical* if the expected number of direct and indirect offsprings (*i.e.*, the progeny) spawned by a single parent is finite. In this case, the Hawkes process is expected to die out as $t \to \infty$. The intuition for the multivariate Hawkes process is similar, but in this case we have to consider that any event in one dimension is capable of producing events in any other dimension by cross-excitation.

A multivariate Hawkes process is subcritical if the progeny resulting from a single event of dimension j to dimension i is finite for every pair $(i, j) \in D \times D$.

Definition C.2. Let $\lambda^1, \cdots, \lambda^d$ be the eigenvalues of the branching matrix $\boldsymbol{\alpha}$. The spectral radius $\rho(\boldsymbol{\alpha})$ of $\boldsymbol{\alpha}$ is defined as

$$\rho(\boldsymbol{\alpha}) = \max\{|\lambda^1|, \ldots, |\lambda^d|\}. \tag{C.6}$$

For a one-dimensional Hawkes process, the spectral radius is exactly the branching factor, the expected number of secondary events triggered by a parent event. If the branching factor is less than one, the Hawkes process is **subcritical**. If the branching factor is greater than one, the process is **supercritical**, and the progeny of a single parent event is expected to have an infinite number of offspring events as $t \to \infty$. In this case the Hawkes process is also called **explosive**.

The following proposition is a standard result that characterizes the convergence of a geometric series of matrices. We use the following to obtain a closed-form expression of the total progeny produced by events in every dimension for the MHP.

Proposition C.2 ([51])**.** *If $\rho(\boldsymbol{\alpha}) < 1$, $\sum_{n=0}^{\infty} \boldsymbol{\alpha}^n$ converges and is equal to $(\boldsymbol{I} - \boldsymbol{\alpha})^{-1}$.*

The subcriticality condition for a multivariate Hawkes process is given by the following.

Theorem C.1. *A Hawkes process with branching matrix $\boldsymbol{\alpha}$ is subcritical if $\rho(\boldsymbol{\alpha}) < 1$.*

Proof. Let $\rho(\boldsymbol{\alpha}) < 1$. Suppose we have one parent event in dimension $j \in D$. Let us consider the offsprings of this parent event.

The expected number of direct (*i.e.*, first-generation) offsprings in dimension i is α^{ij}. The expected number of second-generation offsprings in dimension i is $\sum_{k=1}^{d} \alpha^{ik}\alpha^{kj} = (\boldsymbol{\alpha}^2)^{ij}$, which is intuitively the dimension i offsprings of the first-generation offsprings of the dimension j parent event. By the same argument, the number of m^{th} generation offsprings would then be $(\boldsymbol{\alpha}^m)^{ij}$. Thus, it follows that the (i, j) element of $\sum_{n=1}^{m} \boldsymbol{\alpha}^n$ tracks the total number of dimension i offsprings up to the m^{th} generation produced a single parent event in dimension j. In the limit $m \to \infty$, we can conclude by Proposition C.2 that the Hawkes process is subcritical. Furthermore, the expected number of dimension i offsprings of a dimension j parent event is given by the (i, j) element of $(\mathbf{I} - \boldsymbol{\alpha})^{-1} - \mathbf{I}$. ■

Parameter estimation. Suppose that we are given a set of observed events $\mathcal{H}_{T^-}^D$ up until some maximum time $T > 0$. Our task is to find the parameter set $\boldsymbol{\Theta}$ that best fits this given set

of observations. The standard approach is maximum likelihood estimation (MLE), where we find $\boldsymbol{\Theta}$ that maximizes the probability of observing the data given the point process model. Equivalently, we can minimize the negative log-likelihood function $\mathscr{L}\left(\boldsymbol{\Theta}; \bigcup_{j=1}^{d} \mathscr{H}_{T^-}^{j}\right)$ of the parameter set $\boldsymbol{\Theta}$.

For the d-dimensional Hawkes process (and in general, for a d-dimensional point process with intensity $\boldsymbol{\lambda}^{\star}(t)$), the negative log-likelihood function is given by

$$\begin{aligned} \mathscr{L}_{\text{PP-LL}}\left(\boldsymbol{\Theta}; \mathscr{H}_{T^-}^{D}\right) &:= -\log \mathbb{P}\left\{\mathscr{H}_{T^-}^{D} \middle| \boldsymbol{\Theta}\right\} \\ &= -\sum_{j=1}^{d} \left[\sum_{t_k^j \in \mathscr{H}_{T^-}^{j}} \log \lambda^j(t_k^j; \boldsymbol{\Theta}) - \Lambda^j(T; \boldsymbol{\Theta}) \right]. \end{aligned} \tag{C.7}$$

We add the subscript PP-LL (Point-Process Log-Likelihood) to emphasize that the likelihood is evaluated with respect to event timestamps.

Sampling. Given an MHP, the standard approach to sample event sequences is via the thinning algorithm discussed in [82]. This techniques converts the task of sampling a Hawkes process into the significantly simpler task of sampling a homogeneous Poisson process. The rate of this Poisson process is obtained as an upper bound to the Hawkes conditional intensity and is recomputed every time a new event is accepted. Proposed events from the procedure are 'thinned' out with rejection sampling using the Hawkes conditional intensity. Algorithm 1 shows how to sample event sequences from a d-dimensional Hawkes process given a constant background intensity.

C.1.2 Mean Behavior Poisson Process

Regularity condition. The sufficient condition for the subcriticality of the MBP process is $\alpha < 1$. This condition ensures that the infinite sum $\sum_{n=1}^{\infty} \varphi^{\otimes n}(t)$ in the MBP intensity $\xi(t)$ converges to zero as $t \to \infty$.

Parameter estimation in interval-censored settings. 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 maximum likelihood estimation.

Suppose instead of observing the sequence of events $\mathscr{H}_{T^-}$, we observe interval-censored counts over a given partition of $[0, T)$, which we denote as $\mathscr{P}[0, T)$. Furthermore, assume that the partition is subdivided into m subintervals, so that $\mathscr{P}[0, T) = \bigcup_{k=1}^{m} [o_{k-1}, o_k)$, where

Algorithm 1: Simulating a d-dimensional Hawkes Process on $[0, T)$ with Thinning [82]

Input: kernel matrix $\boldsymbol{\varphi}(t)$, background intensity $\boldsymbol{\mu}$, time horizon $T > 0$
Output: $\mathscr{H}_T^j = \{t_k^j\}$ for $j = 1 : d$
initialize $t = 0; \mathscr{H}_T^1 = \ldots = \mathscr{H}_T^d = \emptyset$;
while $t < T$ **do**
  $\bar{\lambda} = \sum_{i=1}^d \lambda^m(t^+) = \sum_{i=1}^d \left[\mu^i + \sum_{j=1}^d \sum_{t_k^j \le t} \varphi^{ij}(t - t_k^j)\right]$;
  $u \sim$ `uniform(0,1)`;
  $w = -\log \frac{u}{\bar{\lambda}}$;
  $t = t + w$;
  $U \sim$ `uniform(0,1)`;
  if $U\bar{\lambda} \le \sum_{i=1}^d \lambda^m(t)$ **then**
    $j = 1$;
    while $U\bar{\lambda} \le \sum_{i=1}^j \lambda^m(t)$ **do**
      $j = j + 1$;
    end
    $t_k^j = t$;
    $\mathscr{H}_T^j = \mathscr{H}_T^j \cup \{t_k^j\}$;
  end
end
if $t_{k^j}^j < T$ **then**
  return $\mathscr{H}_T^j$ for $j = 1 : d$;
else
  return $\mathscr{H}_T^1, \ldots, \mathscr{H}_T^j \setminus \{t_k^j\}, \ldots, \mathscr{H}_T^d$;
end

$o_0 = 0$ and $o_m = T$. For each subinterval $[o_{k-1}, o_k)$, we are given the count C_k of events that occur. In this setting and given the MBP process, the negative log-likelihood function $\mathscr{L}(\boldsymbol{\Theta}; \{\mathsf{C}_k\}_{k=1}^m)$ can be obtained with the following result.

Proposition C.3 ([102])**.** *Suppose we are given interval-censored counts $\{\mathsf{C}_k\}_{k=1}^m$ over the partition $\mathscr{P}[0, T) = \bigcup_{k=1}^m [o_{k-1}, o_k)$, where $o_0 = 0$ and $o_m = T$. The negative log-likelihood function of an MBP process with intensity $\xi(t)$ and compensator $\Xi(t)$ is given by*

$$\begin{aligned} \mathscr{L}_{\text{IC-LL}}(\boldsymbol{\Theta}; \{\mathsf{C}_k\}_{k=1}^m) &= -\log \mathbb{P}\left\{\bigcup_{k=1}^m \{\Xi(o_{k-1}, o_k) = \mathsf{C}_k\} | \boldsymbol{\Theta}\right\} \\ &\propto \sum_{k=1}^m \left[\Xi(o_{k-1}, o_k; \boldsymbol{\Theta}) - \mathsf{C}_k \log \Xi(o_{k-1}, o_k; \boldsymbol{\Theta})\right], \end{aligned} \tag{C.8}$$

where we add the subscript IC-LL (Interval-Censored Log Likelihood) to make explicit the fact that we are calculating the likelihood with respect to interval-cen

C.2 Interpretation of $\mathbf{h}_E$

The function $\mathbf{h}_E(t) = \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(t)$ appearing in Eq. (4.5) can be expressed term-by-term as

$$\mathbf{h}_E(t) = \boldsymbol{\varphi}_E(t) + \boldsymbol{\varphi}_E^{\otimes 2}(t) + \boldsymbol{\varphi}_E^{\otimes 3}(t) + \cdots \tag{C.9}$$

Given a Hawkes kernel of the form $\boldsymbol{\varphi}(t) = \boldsymbol{\alpha} \odot \mathbf{f}(t)$, where $\odot$ denotes elementwise multiplication, Eq. (C.9) can be written as

$$\mathbf{h}_E(t) = \boldsymbol{\alpha}_{\boldsymbol{E}} \odot \mathbf{f}_E(t) + \boldsymbol{\alpha}_{\boldsymbol{E}}{}^2 \odot \mathbf{f}_E^{\otimes 2}(t) + \boldsymbol{\alpha}_E^3 \odot \mathbf{f}_E^{\otimes 3}(t) + \cdots \tag{C.10}$$

Consider the i, j entry of the n^{th} term of the sum:

$$(\boldsymbol{\alpha}_E^n \odot \mathbf{f}_E^{\otimes n}(t))^{ij} = (\boldsymbol{\alpha}_E^n)^{ij} \odot (\mathbf{f}_E^{\otimes n}(t))^{ij}$$

This expression can be interpreted as follows:

- $(\boldsymbol{\alpha}_E^n)^{ij}$ is the expected number of n^{th} generation offspring events of type i produced by a single parent of type j in a d-dimensional branching process where only the E dimensions can produce offsprings.
- $(\mathbf{f}_E^{\otimes n}(t))^{ij}$ is the density of n^{th} generation type i offspring events at time t produced by a single parent of type j at time 0.
- The product $(\boldsymbol{\alpha}_E^n)^{ij} \odot (\mathbf{f}_E^{\otimes n}(t))^{ij}$ can be intepreted as the expected intensity contribution at time t from n^{th} generation type i offspring events produced by a single parent of type j.

Thus $h_E^{ij}(t)$, given by

$$\mathbf{h}_E^{ij}(t) = \underbrace{(\boldsymbol{\alpha}_{\boldsymbol{E}} \odot \mathbf{f}_E(t))^{ij}}_{\text{first generation}} + \underbrace{(\boldsymbol{\alpha}_{\boldsymbol{E}}{}^2 \odot \mathbf{f}_E^{\otimes 2}(t))^{ij}}_{\text{second generation}} + \underbrace{(\boldsymbol{\alpha}_E^3 \odot \mathbf{f}_E^{\otimes 3}(t))^{ij}}_{\text{third generation}} + \cdots,$$

is the expected type-i intensity at time t from a single parent event of type j, over the entire progeny of offsprings.

Fig. C.1 shows the nonzero entries of $\mathbf{h}_E(t)$ for a PCMHP(3,2) process with parameter set $\boldsymbol{\theta} = [1, 0.1, 0, 1, 1, 0, 1, 1, 0]$, $\boldsymbol{\alpha} = [0.2, 0.2, 0, 0.2, 0.2, 0, 0.2, 0.2, 0]$. In the plot we show the contributions of the first to the fifth generation to $\mathbf{h}_E(t)$. As we can see, the contributions of succeeding generations become increasingly smaller as every $\alpha^{ij} < 1$. The contributions all go to zero asymptotically as $t \to \infty$. In addition, we see that the mode of each generation's contribution is shifted to the right as the generation index increases. Intuitively, a delay exists because offspring events in generation $n+1$ are produced by generation n events. The extent of the delay is controlled by $\boldsymbol{\theta}$.

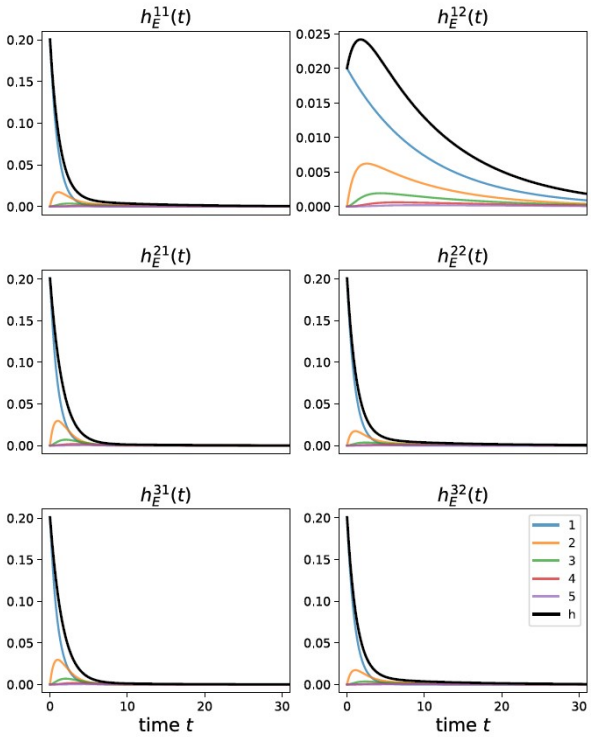


Figure C.1: Nonzero entries of $\mathbf{h}_E(t)$ for a PCMHP(3,2) process with parameter set $\boldsymbol{\theta} = [1, 0.1, 0, 1, 1, 0, 1, 1, 0]$, $\boldsymbol{\alpha} = [0.2, 0.2, 0, 0.2, 0.2, 0, 0.2, 0.2, 0]$. Colored lines correspond to the contribution of the first up to the fifth generation offsprings to $\mathbf{h}_E(t)$. The black line ($\mathbf{h}_E(t)$) is the total contribution of the progeny.

C.3 Closed Form $\boldsymbol{\xi}_E(t)$ for the PCMHP$(2,1)$ Process

Given a PCMHP(2,1) process, the conditional intensity function $\boldsymbol{\xi}_E(t)$ is given by

$$\xi_1^1(t \mid \mathscr{H}_t^2) = \mu^1 + \sum_{t_k^2 < t} \varphi^{12}(t - t_k^2) + (\varphi^{11} * \xi_1^1)(t)$$
$$\xi_1^2(t \mid \mathscr{H}_t^2) = \mu^2 + \sum_{t_k^2 < t} \varphi^{22}(t - t_k^2) + (\varphi^{21} * \xi_1^1)(t).$$

Given a fixed event sequence for dimension 2 $\{t_1^2 < t_2^2 < \ldots < t_N^2\}$ prior to time t, the intensity function given by $\xi_1^1(t \mid \mathscr{H}_t^1)$ can be interpreted as a univariate MBP process. Assuming an exponential kernel, the intensity function in this case can be expressed in closed form using the impulse response function.

Suppose that $\varphi^{ij}(x) = \alpha^{ij}\theta^{ij}\exp(-\theta^{ij}x)$. The impulse response for $\xi_1^1(t \mid \mathscr{H}_t^2)$ is given by:

$$E_1(t) = \delta(t) + h(t), \tag{C.11}$$

where

$$h(t) = \alpha^{11}\theta^{11}\exp((\alpha^{11} - 1)\theta^{11}t) \cdot [\![t \geq 1]\!]. \tag{C.12}$$

Setting $\hat{s}(t) = \left[\mu^1 + \sum_{t_k^2 < t} \alpha^{12}\theta^{12}\exp(-\theta^{12}(t - t_k^2))\right]$,

$$\xi_1^1(t \mid \mathscr{H}_t^2) = (E_1 * \hat{s})(t) = \hat{s}(t) + (h * \hat{s})(t). \tag{C.13}$$

$$
\begin{aligned}
&(h * \hat{s})(t) \\
&= \int_1^t h(t-\tau) \cdot \hat{s}(\tau) \mathrm{d}\tau \\
&= \int_1^t \alpha^{11}\theta^{11} \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \cdot \hat{s}(\tau) \mathrm{d}\tau \\
&= \int_1^t \alpha^{11}\theta^{11} \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \cdot \left[\mu^1 + \sum_{t_k^2 < t} \alpha^{12}\theta^{12} \exp(-\theta^{12}(\tau - t_k^2)) \right] \mathrm{d}\tau \\
&= \mu^1 \int_1^t \alpha^{11}\theta^{11} \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \mathrm{d}\tau \\
&\qquad + \int_1^t \alpha^{11}\theta^{11} \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \cdot \left[\sum_{t_k^2 < \tau} \alpha^{12}\theta^{12} \exp(-\theta^{12}(\tau - t_k^2)) \right] \mathrm{d}\tau \\
&= \mu^1 \alpha^{11}\theta^{11} \int_1^t \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \mathrm{d}\tau \\
&\qquad + \alpha^{11}\theta^{11}\alpha^{12}\theta^{12} \int_1^t \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \cdot \left[\sum_{t_k^2 < \tau} \exp(-\theta^{12}(\tau - t_k^2)) \right] \mathrm{d}\tau.
\end{aligned}
$$

We calculate each of the integral terms separately:

$$
\begin{aligned}
&\int_1^t \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \mathrm{d}\tau \\
&= -\frac{1}{(\alpha^{11}-1)\theta^{11}} \exp((\alpha^{11}-1)\theta^{11}(t-\tau)) \Big]_{\tau=1}^{\tau=t} \\
&= \frac{1}{(\alpha^{11}-1)\theta^{11}} \left[\exp((\alpha^{11}-1)\theta^{11} t) - 1 \right].
\end{aligned}
$$

And,

$$
\begin{aligned}
&\int_1^t \exp((\alpha^{11}-1)\theta^{11}(t-\tau))\cdot\left[\sum_{t_k^2<\tau}\exp(-\theta^{12}(\tau-t_k^2))\right]\mathrm{d}\tau \\
&=\sum_{k=1}^{N}\int_{t_k^2}^{t}\exp((\alpha^{11}-1)\theta^{11}(t-\tau))\cdot\exp(-\theta^{12}(\tau-t_k^2))\mathrm{d}\tau \\
&=\sum_{k=1}^{N}\int_{t_k^2}^{t}\exp((\alpha^{11}-1)\theta^{11}t)\cdot\exp(-(\alpha^{11}-1)\theta^{11}\tau)\cdot\exp(-\theta^{12}\tau)\cdot\exp(\theta^{12}t_k^2)\mathrm{d}\tau \\
&=\sum_{i=1}^{N}\exp((\alpha^{11}-1)\theta^{11}t)\cdot\exp(\theta^{12}t_k^2)\cdot\int_{t_k^2}^{t}\exp(-(\alpha^{11}-1)\theta^{11}\tau)\cdot\exp(-\theta^{12}\tau)\mathrm{d}\tau \\
&=\sum_{i=1}^{N}\exp((\alpha^{11}-1)\theta^{11}t)\cdot\exp(\theta^{12}t_k^2)\cdot\int_{t_k^2}^{t}\exp(-[(\alpha^{11}-1)\theta^{11}+\theta^{12}]\tau)\mathrm{d}\tau \\
&=\sum_{i=1}^{N}\frac{\exp((\alpha^{11}-1)\theta^{11}t)\cdot\exp(\theta^{12}t_k^2)}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \\
&\qquad\cdot\left[\exp(-[(\alpha^{11}-1)\theta^{11}+\theta^{12}]t_k^2)-\exp(-[(\alpha^{11}-1)\theta^{11}+\theta^{12}]t)\right] \\
&=\sum_{i=1}^{N}\frac{\exp((\alpha^{11}-1)\theta^{11}(t-t_k^2))-\exp(-\theta^{12}(t-t_k^2))}{(\alpha^{11}-1)\theta^{11}+\theta^{12}}
\end{aligned}
$$

Together we have,

$$
\begin{aligned}
(h*\hat{s})(t) = &\frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\left[\exp((\alpha^{11}-1)\theta^{11}t)-1\right] \\
&+\sum_{k=1}^{N}\frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}}\left[\exp((\alpha^{11}-1)\theta^{11}(t-t_k^2))-\exp(-\theta^{12}(t-t_k^2))\right].
\end{aligned}
$$

This gives the MBP intensity function for dimension 1 as

$$
\begin{aligned}
\xi_1^1(t\mid\mathscr{H}_t^2) = &\mu^1+\sum_{t_k^2<t}\alpha^{12}\theta^{12}\exp(-\theta^{12}(t-t_k^2))+\frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\left[\exp((\alpha^{11}-1)\theta^{11}t)-1\right] \\
&+\sum_{t_k^2<t}\frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}}\left[\exp((\alpha^{11}-1)\theta^{11}(t-t_k^2))-\exp(-\theta^{12}(t-t_k^2))\right].
\end{aligned}
\tag{C.14}
$$

To calculate the MBP intensity function ξ_1^2 for dimension 2, we would need to calculate

$$
(\varphi^{21}*\xi_1^1)(t)=\alpha^{21}\theta^{21}\int_1^t\exp(-\theta^{21}(t-\tau))\xi_1^1(\tau\mid\mathscr{H}_\tau^2)\mathrm{d}\tau.
$$

We expand the integrand above using the expression for ξ_1^1 in Eq. (C.15) and calculate the integral term-by-term.

$$\begin{aligned}\int_1^t \exp(-\theta^{21}(t-\tau))\mu^1 \mathrm{d}\tau &= \mu^1 \int_1^t \exp(-\theta^{21}(t-\tau))\mathrm{d}\tau \\ &= -\frac{\mu^1}{\theta^{21}} \exp(-\theta^{21}(t-\tau))\Big]_{\tau=1}^{\tau=t} \\ &= \frac{\mu^1}{\theta^{21}}\left[2-\exp(-\theta^{21}(t))\right].\end{aligned}$$

$$\begin{aligned}&\int_2^t \exp(-\theta^{21}(t-\tau))\frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\left[\exp((\alpha^{11}-1)\theta^{11}\tau)-1\right]\mathrm{d}\tau \\ &= \frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\int_1^t \exp(-\theta^{21}(t-\tau))\left[\exp((\alpha^{11}-1)\theta^{11}\tau)-1\right]\mathrm{d}\tau \\ &= \frac{\mu^2\alpha^{11}}{\alpha^{11}-1}\exp(-\theta^{21}t)\int_1^t \exp(\theta^{21}\tau))\left[\exp((\alpha^{11}-1)\theta^{11}\tau)-1\right]\mathrm{d}\tau \\ &= \frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\exp(-\theta^{21}t)\left[\int_1^t \exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})\tau)\mathrm{d}\tau - \int_1^t \exp(\theta^{21}\tau))\mathrm{d}\tau\right] \\ &= \frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\exp(-\theta^{21}t)\left[\frac{\exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})t)-1}{\theta^{21}+(\alpha^{11}-1)\theta^{11}} - \frac{\exp(\theta^{21}t)-1}{\theta^{21}}\right].\end{aligned}$$

For the remaining two terms in the integral, we need to consider two cases: $\theta^{21}=\theta^{12}$ and $\theta^{21}\neq\theta^{12}$.

Case 1: $\theta^{21}=\theta^{12}$.

$$\begin{aligned}&\int_1^t \exp(-\theta^{21}(t-\tau))\sum_{t_k^2<t}\alpha^{12}\theta^{12}\exp(-\theta^{12}(\tau-t_k^2))\mathrm{d}\tau \\ &= \alpha^{12}\theta^{12}\int_1^t \exp(-\theta^{21}(t-\tau))\sum_{t_k^2<t}\exp(-\theta^{12}(\tau-t_k^2))\mathrm{d}\tau \\ &= \alpha^{12}\theta^{12}\sum_{k=1}^{N}\int_{t_k^2}^t \exp(-\theta^{21}(t-\tau))\exp(-\theta^{12}(\tau-t_k^2))\mathrm{d}\tau \\ &= \alpha^{12}\theta^{12}\sum_{k=1}^{N}\exp(\theta^{12}t_k^2-\theta^{21}t)\int_{t_k^2}^t \exp((\theta^{21}-\theta^{12})\tau)\mathrm{d}\tau \\ &= \alpha^{12}\theta^{12}\sum_{k=1}^{N}\exp(\theta^{12}t_k^2-\theta^{21}t)\int_{t_k^2}^t \mathrm{d}\tau \\ &= \alpha^{12}\theta^{12}\sum_{k=1}^{N}\exp(\theta^{12}t_k^2-\theta^{21}t)(t-t_k^2)\end{aligned}$$

$$
\begin{aligned}
&\int_1^t \exp(-\theta^{21}(t-\tau)) \sum_{t_k^2<\tau} \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \\
&\qquad \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \int_1^t \exp(-\theta^{21}(t-\tau)) \\
&\qquad \sum_{t_k^2<\tau} \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \int_1^t \sum_{t_k^2<\tau} \exp(-\theta^{21}(t-\tau)) \\
&\qquad \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \int_{t_k^2}^t \exp(\theta^{21}\tau) \\
&\qquad \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \\
&\qquad \left[\exp(-(\alpha^{11}-1)\theta^{11}t_k^2) \int_{t_k^2}^t \exp(\theta^{21}\tau)\exp((\alpha^{11}-1)\theta^{11}\tau)\mathrm{d}\tau \right. \\
&\qquad\quad \left. - \exp(\theta^{12}t_k^2) \int_{t_k^2}^t \exp(\theta^{21}\tau)\exp(-\theta^{12}\tau)\mathrm{d}\tau \right] \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \\
&\qquad \left[\exp(-(\alpha^{11}-1)\theta^{11}t_k^2) \int_{t_k^2}^t \exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})\tau)\mathrm{d}\tau \right. \\
&\qquad\quad \left. - \exp(\theta^{12}t_k^2) \int_{t_k^2}^t \exp((\theta^{21}-\theta^{12})\tau)\mathrm{d}\tau \right] \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \\
&\qquad \left[\exp(-(\alpha^{11}-1)\theta^{11}t_k^2) \int_{t_k^2}^t \exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})\tau)\mathrm{d}\tau - \exp(\theta^{12}t_k^2) \int_{t_k^2}^t \mathrm{d}\tau \right] \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \left[\frac{\exp((\alpha^{11}-1)\theta^{11}(t-t_k^2))}{\theta^{21}+(\alpha^{11}-1)\theta^{11}} - \frac{\exp(-\theta^{21}(t-t_k^2))}{\theta^{21}+(\alpha^{11}-1)\theta^{11}} \right. \\
&\qquad \left. - \exp(\theta^{12}t_k^2 - \theta^{21}t)(t-t_k^2)\right].
\end{aligned}
$$

Case 2: $\theta^{21} \neq \theta^{12}$.

$$
\begin{aligned}
&\int_1^t \exp(-\theta^{21}(t-\tau)) \sum_{t_k^2<t} \alpha^{12}\theta^{12} \exp(-\theta^{12}(\tau - t_k^2))\mathrm{d}\tau \\
&= \alpha^{12}\theta^{12} \int_1^t \exp(-\theta^{21}(t-\tau)) \sum_{t_k^2<t} \exp(-\theta^{12}(\tau - t_k^2))\mathrm{d}\tau \\
&= \alpha^{12}\theta^{12} \sum_{k=1}^{N} \int_{t_k^2}^t \exp(-\theta^{21}(t-\tau)) \exp(-\theta^{12}(\tau - t_k^2))\mathrm{d}\tau \\
&= \alpha^{12}\theta^{12} \sum_{k=1}^{N} \exp(\theta^{12} t_k^2 - \theta^{21} t) \int_{t_k^2}^t \exp((\theta^{21} - \theta^{12})\tau)\mathrm{d}\tau \\
&= \frac{\alpha^{12}\theta^{12}}{\theta^{21} - \theta^{12}} \sum_{k=1}^{N} \exp(\theta^{12} t_k^2 - \theta^{21} t) \left[\exp((\theta^{21} - \theta^{12})\, t) - \exp((\theta^{21} - \theta^{12})\, t_k^2)\right] \\
&= \frac{\alpha^{12}\theta^{12}}{\theta^{21} - \theta^{12}} \sum_{k=1}^{N} \left[\exp(-\theta^{12}(t - t_k^2)) - \exp(-\theta^{21}(t - t_k^2))\right].
\end{aligned}
$$

$$
\begin{aligned}
&\int_1^t \exp(-\theta^{21}(t-\tau)) \sum_{t_k^2<\tau} \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \int_1^t \exp(-\theta^{21}(t-\tau)) \sum_{t_k^2<\tau} \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \int_1^t \sum_{t_k^2<\tau} \exp(-\theta^{21}(t-\tau)) \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \int_{t_i^1}^t \exp(-\theta^{21}(t-\tau)) \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \int_{t_k^2}^t \exp(\theta^{21}\tau) \left[\exp((\alpha^{11}-1)\theta^{11}(\tau-t_k^2)) - \exp(-\theta^{12}(\tau-t_k^2))\right] \mathrm{d}\tau \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \left[\exp(-(\alpha^{11}-1)\theta^{11}t_k^2) \int_{t_k^2}^t \exp(\theta^{21}\tau)\exp((\alpha^{11}-1)\theta^{11}\tau)\mathrm{d}\tau \right. \\
&\qquad \left. - \exp(\theta^{12}t_k^2) \int_{t_k^2}^t \exp(\theta^{21}\tau)\exp(-\theta^{12}\tau)\mathrm{d}\tau \right] \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \left[\exp(-(\alpha^{11}-1)\theta^{11}t_k^2) \int_{t_k^2}^t \exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})\tau)\mathrm{d}\tau \right. \\
&\qquad \left. - \exp(\theta^{12}t_k^2) \int_{t_k^2}^t \exp((\theta^{21}-\theta^{12})\tau)\mathrm{d}\tau \right] \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \exp(-\theta^{21}t) \left[\frac{\exp(-(\alpha^{11}-1)\theta^{11}t_k^2)}{\theta^{21}+(\alpha^{11}-1)\theta^{11}} \left[\exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})\,t) \right.\right. \\
&\qquad \left.\left. - \exp((\theta^{21}+(\alpha^{11}-1)\theta^{11})\,t_k^2)\right] - \frac{\exp(\theta^{12}t_k^2)}{\theta^{21}-\theta^{12}} \left[\exp((\theta^{21}-\theta^{12})\,t) - \exp((\theta^{21}-\theta^{12})\,t_k^2)\right] \right] \\
&= \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11}+\theta^{12}} \sum_{k=1}^{N} \left[\frac{\exp((\alpha^{11}-1)\theta^{11}(t-t_k^2))}{\theta^{21}+(\alpha^{11}-1)\theta^{11}} - \frac{\exp(-\theta^{21}(t-t_k^2))}{\theta^{21}+(\alpha^{11}-1)\theta^{11}} - \frac{\exp(-\theta^{12}(t-t_k^2))}{\theta^{21}-\theta^{12}} \right. \\
&\qquad \left. + \frac{\exp(-\theta^{21}(t-t_k^2))}{\theta^{21}-\theta^{12}} \right].
\end{aligned}
$$

Therefore, if $\theta^{21} = \theta^{12}$,

$$\begin{aligned}\xi_1^2(t \mid \mathscr{H}_t^2) &= \mu^2 + \sum_{k=1}^{N} \alpha^{22}\theta^{22} \exp(-\theta^{22}(t - t_k^2)) + \mu^1\alpha^{21}\left[2 - \exp(-\theta^{21}(t))\right] \\ &\quad + \alpha^{12}\theta^{12}\alpha^{21}\theta^{21} \sum_{k=1}^{N} \exp(\theta^{12}t_k^2 - \theta^{21}t)(t - t_k^2) \\ &\quad + \frac{\mu^1\alpha^{11}\alpha^{21}\theta^{21}}{\alpha^{11}-1}\left[\frac{\exp(((\alpha^{11}-1)\theta^{11})t) - \exp(-\theta^{21}t)}{\theta^{21} + (\alpha^{11}-1)\theta^{11}} - \frac{2 - \exp(-\theta^{21}t)}{\theta^{21}}\right] \\ &\quad + \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}\alpha^{21}\theta^{21}}{(\alpha^{11}-1)\theta^{11} + \theta^{12}} \sum_{k=1}^{N}\left[\frac{\exp((\alpha^{11}-1)\theta^{11}(t - t_k^2))}{\theta^{21} + (\alpha^{11}-1)\theta^{11}}\right. \\ &\qquad \left. - \frac{\exp(-\theta^{21}(t - t_k^2))}{\theta^{21} + (\alpha^{11}-1)\theta^{11}} - \exp(\theta^{12}t_k^2 - \theta^{21}t)(t - t_k^2)\right].\end{aligned}$$

If $\theta^{21} \neq \theta^{12}$,

$$\begin{aligned}\xi_1^2(t \mid \mathscr{H}_t^2) &= \mu^2 + \sum_{k=1}^{N} \alpha^{22}\theta^{22} \exp(-\theta^{22}(t - t_k^2)) + \mu^1\alpha^{21}\left[2 - \exp(-\theta^{21}(t))\right] \\ &\quad + \frac{\alpha^{12}\theta^{12}\alpha^{21}\theta^{21}}{\theta^{21} - \theta^{12}} \sum_{k=1}^{N}\left[\exp(-\theta^{12}(t - t_k^2)) - \exp(-\theta^{21}(t - t_k^2))\right] \\ &\quad + \frac{\mu^1\alpha^{11}\alpha^{21}\theta^{21}}{\alpha^{11}-1}\left[\frac{\exp(((\alpha^{11}-1)\theta^{11})t) - \exp(-\theta^{21}t)}{\theta^{21} + (\alpha^{11}-1)\theta^{11}} - \frac{2 - \exp(-\theta^{21}t)}{\theta^{21}}\right] \\ &\quad + \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}\alpha^{21}\theta^{21}}{(\alpha^{11}-1)\theta^{11} + \theta^{12}} \sum_{k=1}^{N}\left[\frac{\exp((\alpha^{11}-1)\theta^{11}(t - t_k^2))}{\theta^{21} + (\alpha^{11}-1)\theta^{11}}\right. \\ &\qquad \left. - \frac{\exp(-\theta^{21}(t - t_k^2))}{\theta^{21} + (\alpha^{11}-1)\theta^{11}} - \frac{\exp(-\theta^{12}(t - t_k^2))}{\theta^{21} - \theta^{12}} + \frac{\exp(-\theta^{21}(t - t_k^2))}{\theta^{21} - \theta^{12}}\right].\end{aligned}$$

Integrating $\xi_1^1(t \mid \mathscr{H}_t^2)$ and $\xi_1^2(t \mid \mathscr{H}_t^2)$ over the interval $[1, T]$, we obtain the compensator as (Take note that integrals over the exponentials are from t_2^k to t)

$$\begin{aligned}\Xi_1^1(t \mid \mathscr{H}_t^2) &= \mu^1 t + \sum_{t_k^2 < t} \alpha^{12}(2 - \exp(-\theta^{12}(t - t_k^2))) \\ &\quad + \frac{\mu^1\alpha^{11}}{\alpha^{11}-1}\left[\frac{1}{(\alpha^{11}-1)\theta^{11}}(\exp((\alpha^{11}-1)\theta^{11}t) - 1) - t\right] \\ &\quad + \sum_{t_k^2 < t} \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}}{(\alpha^{11}-1)\theta^{11} + \theta^{12}}\left[\frac{\exp((\alpha^{11}-1)\theta^{11}(t - t_k^2)) - 1}{(\alpha^{11}-1)\theta^{11}} - \frac{2 - \exp(-\theta^{12}(t - t_2^k))}{\theta^{12}}\right].\end{aligned} \tag{C.15}$$

If $\theta^{21} = \theta^{12}$,

$$\begin{aligned}
\Xi_1^2(t \mid \mathscr{H}_t^2) &= \mu^2 t + \sum_{k=1}^{N} \alpha^{22}(2 - \exp(-\theta^{22}(t - t_2^k))) + \mu^1 \alpha^{21} \left[t - \frac{1}{\theta^{21}}(2 - \exp(-\theta^{21} t)) \right] \\
&+ \frac{\alpha^{12}\theta^{12}\alpha^{21}}{\theta^{21}} \sum_{k=1}^{N} \exp(\theta^{12} t_k^2 - \theta^{21}(t_k^2 + t)) \left[\exp(\theta^{21} t_k^2) \left(\theta^{21} t_k^2 - \theta^{21} t - 1 \right) + \exp(\theta^{21} t) \right] \\
&+ \frac{\mu^1 \alpha^{11} \alpha^{21} \theta^{21}}{\alpha^{11} - 1} \left[\frac{1}{\theta^{21} + (\alpha^{11} - 1)\theta^{11}} \left(\frac{\exp(((\alpha^{11} - 1)\theta^{11})\, t) - 2}{(\alpha^{11} - 1)\theta^{11}} + \frac{\exp(-\theta^{21} t) - 1}{\theta^{21}} \right) \right. \\
&\quad \left. - \left(\frac{t}{\theta^{21}} + \frac{\exp(-\theta^{21} t) - 1}{(\theta^{21})^2} \right) \right] \\
&+ \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}\alpha^{21}\theta^{21}}{(\alpha^{11} - 1)\theta^{11} + \theta^{12}} \sum_{k=1}^{N} \left[\frac{\exp((\alpha^{11} - 1)\theta^{11}(t - t_2^k)) - 2}{(\theta^{21} + (\alpha^{11} - 1)\theta^{11})(\alpha^{11} - 1)\theta^{11}} \right. \\
&\qquad + \frac{\exp(-\theta^{21}(t - t_k^2)) - 2}{(\theta^{21} + (\alpha^{11} - 1)\theta^{11})\theta^{21}} - \frac{1}{(\theta^{21})^2} \exp(\theta^{12} t_k^2 - \theta^{21}(t_k^2 + t)) \\
&\quad \left. \left[\exp(\theta^{21} t_k^2) \left(\theta^{21} t_k^2 - \theta^{21} t - 1 \right) + \exp(\theta^{21} t) \right] \right].
\end{aligned}$$

If $\theta^{21} \neq \theta^{12}$,

$$\begin{aligned}
\Xi_1^2(t \mid \mathscr{H}_t^2) &= \mu^2 t + \sum_{k=1}^{N} \alpha^{22}(2 - \exp(-\theta^{22}(t - t_2^k))) + \mu^1 \alpha^{21} \left[t - \frac{1}{\theta^{21}}(2 - \exp(-\theta^{21} t)) \right] \\
&\quad + \frac{\alpha^{12}\theta^{12}\alpha^{21}\theta^{21}}{\theta^{21} - \theta^{12}} \sum_{k=1}^{N} \left[\frac{2 - \exp(-\theta^{12}(t - t_2^k))}{\theta^{12}} - \frac{2 - \exp(-\theta^{21}(t - t_2^k))}{\theta^{21}} \right] \\
&\quad + \frac{\mu^1 \alpha^{11} \alpha^{21} \theta^{21}}{\alpha^{11} - 1} \left[\frac{1}{\theta^{21} + (\alpha^{11} - 1)\theta^{11}} \left(\frac{\exp(((\alpha^{11} - 1)\theta^{11})\, t) - 2}{(\alpha^{11} - 1)\theta^{11}} \right. \right. \\
&\qquad \left. + \frac{\exp(-\theta^{21} t) - 1}{\theta^{21}} \right) - \left(\frac{t}{\theta^{21}} + \frac{\exp(-\theta^{21} t) - 1}{(\theta^{21})^2} \right) \Bigg] \\
&\quad + \frac{\alpha^{11}\theta^{11}\alpha^{12}\theta^{12}\alpha^{21}\theta^{21}}{(\alpha^{11} - 1)\theta^{11} + \theta^{12}} \sum_{k=1}^{N} \left[\frac{\exp((\alpha^{11} - 1)\theta^{11}(t - t_k^2)) - 2}{(\theta^{21} + (\alpha^{11} - 1)\theta^{11})(\alpha^{11} - 1)\theta^{11}} \right. \\
&\qquad - \frac{2 - \exp(-\theta^{21}(t - t_k^2))}{(\theta^{21} + (\alpha^{11} - 1)\theta^{11})\theta^{21}} - \frac{1}{\theta^{21} - \theta^{12}} \left[\frac{2 - \exp(-\theta^{12}(t - t_2^k))}{\theta^{12}} \right. \\
&\qquad \left. \left. - \frac{2 - \exp(-\theta^{21}(t - t_2^k))}{\theta^{21}} \right] \right].
\end{aligned}$$

C.4 Additional Results and Proofs for PCMHP Formulation

C.4.1 Convolutional Formula

Lemma C.1. *Consider* PCMHP(d,e) *with conditional intensity* $\boldsymbol{\xi}_E(t)$ *and* $\mathbf{N}(t)$ *be the counting process of the corresponding* d*-dimensional Hawkes process. The following holds:*

$$\boldsymbol{\xi}_E(s)\mathrm{d}s = \mathbb{E}_{\mathcal{H}_{s^-}^E}\left[\mathrm{d}\mathbf{N}(s)\middle|\mathcal{H}_{s^-}^{E^c}\right]. \tag{C.16}$$

Proof. Taking the conditional expectation over $\mathcal{H}_{s^-}^E$ of both sides of Eq. (C.1), we get

$$\mathbb{E}_{\mathcal{H}_{s^-}^E}\left[\boldsymbol{\lambda}^\star(s)\mathrm{d}s\middle|\mathcal{H}_{s^-}^{E^c}\right] = \mathbb{E}_{\mathcal{H}_{s^-}^E}\left[\mathbb{E}\left[\mathrm{d}\mathbf{N}(s)|\mathcal{H}_{s^-}^{D}\right]\middle|\mathcal{H}_{s^-}^{E^c}\right]$$

Pulling out the infinitesimal $\mathrm{d}s$ out of the expectation on the left-hand side, we get

$$\mathbb{E}_{\mathcal{H}_{s^-}^E}\left[\boldsymbol{\lambda}^\star(s)\middle|\mathcal{H}_{s^-}^{E^c}\right]\mathrm{d}s = \mathbb{E}_{\mathcal{H}_{s^-}^E}\left[\mathbb{E}\left[\mathrm{d}\mathbf{N}(s)|\mathcal{H}_{s^-}^{D}\right]\middle|\mathcal{H}_{s^-}^{E^c}\right]$$

Notice that the left-hand side is exactly $\boldsymbol{\xi}_E(t)$, as defined in Eq. (4.4). For the right-hand side, note that $\mathcal{H}_{s^-}^D = (\mathcal{H}_{s^-}^E) \cup (\mathcal{H}_{s^-}^{E^c})$. Applying the tower property of conditional expectation, we average out the inner conditioning over $\mathcal{H}_{s^-}^E$ and arrive at the desired result:

$$\boldsymbol{\xi}_E(s)\mathrm{d}s = \mathbb{E}_{\mathcal{H}_{s^-}^E}\left[\mathrm{d}\mathbf{N}(s)\middle|\mathcal{H}_{s^-}^{E^c}\right].$$

■

Proposition C.4. *Lemma C.1 extends the result in Eq.* (C.1) *for the* d*-dimensional Hawkes process, to the partial multivariate case considered by the* PCMHP(d,e) *process. To go from Eq.* (C.1) *to Eq.* (C.16)*, we simply replace the total expectation with the conditional expectation over event histories in the E dimensions.*

The following theorem provides an expression for the conditional intensity $\boldsymbol{\xi}_E(t)$ of a PCMHP(d,e) process in terms of the Hawkes kernel $\boldsymbol{\varphi}(t)$ and the background intensity $\boldsymbol{\mu}(t)$.

Theorem C.2. *Given the Hawkes process with the kernel* $\boldsymbol{\varphi}(t)$ *and the background intensity* $\boldsymbol{\mu}(t)$*, the conditional intensity of its corresponding* PCMHP(d,e) *process is given by*

$$\boldsymbol{\xi}_E(t) = \boldsymbol{\mu}(t) + (\boldsymbol{\varphi}_E * \boldsymbol{\xi}_E)(t) + \sum_{j\in E^c}\sum_{t_k^j<t}\boldsymbol{\varphi}^j(t-t_k^j). \tag{C.17}$$

Proof. Starting from the PCMHP definition Eq. (4.5), we expand $\boldsymbol{\lambda}^\star(t)$ as shown below.

$$\boldsymbol{\xi}_E(t) = \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\boldsymbol{\mu}(t) + \int_0^{t^-} \boldsymbol{\varphi}(t-s)\cdot \mathrm{d}\mathbf{N}(s)\middle| \mathscr{H}^{E^c}_{t^-}\right].$$

The exogenous term $\boldsymbol{\mu}(t)$ is independent of the conditioning and may be taken out. We then use the decomposition of $\boldsymbol{\varphi}$ as $\boldsymbol{\varphi}_E + \boldsymbol{\varphi}_{E^c}$. We then have

$$\boldsymbol{\xi}_E(t) = \boldsymbol{\mu}(t) + \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\int_0^{t^-} \boldsymbol{\varphi}_E(t-s)\cdot \mathrm{d}\mathbf{N}(s) + \int_0^{t^-} \boldsymbol{\varphi}_{E^c}(t-s)\cdot \mathrm{d}\mathbf{N}(s)\middle| \mathscr{H}^{E^c}_{t^-}\right].$$

Given that the conditioning assumes that events in $\mathscr{H}^{E^c}_{t^-}$ are observed, we can write the integral involving $\boldsymbol{\varphi}_{E^c}$ as a sum of events in the E^c dimensions. We then have the following sequence of calculations.

$$\begin{aligned}
\boldsymbol{\xi}_E(t) &= \boldsymbol{\mu}(t) + \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\int_0^{t^-} \boldsymbol{\varphi}_E(t-s)\cdot \mathrm{d}\mathbf{N}(s) + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}_{E^c}(t-t_k^j)\middle| \mathscr{H}^{E^c}_{t^-}\right] \\
&\overset{(a)}{=} \boldsymbol{\mu}(t) + \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\int_0^{t^-} \boldsymbol{\varphi}_E(t-s)\cdot \mathrm{d}\mathbf{N}(s)\middle| \mathscr{H}^{E^c}_{t^-}\right] + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}_{E^c}(t-t_k^j) \\
&\overset{(b)}{=} \boldsymbol{\mu}(t) + \int_0^{t^-} \boldsymbol{\varphi}_E(t-s)\cdot \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\mathrm{d}\mathbf{N}(s)\middle| \mathscr{H}^{E^c}_{t^-}\right] + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}_{E^c}(t-t_k^j) \\
&\overset{(c)}{=} \boldsymbol{\mu}(t) + \int_0^{t^-} \boldsymbol{\varphi}_E(t-s)\cdot \boldsymbol{\xi}_E(s)\mathrm{d}s + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}_{E^c}(t-t_k^j) \\
&= \boldsymbol{\mu}(t) + (\boldsymbol{\varphi}_E * \boldsymbol{\xi}_E)(t) + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}_{E^c}(t-t_k^j).
\end{aligned}$$

In (a), we take out the event intensity contributed by events in the dimensions in E^c, as these are observed in $\mathscr{H}^{E^c}_{t^-}$. In (b), we reverse the order of the integral over time and the conditional expectation. In (c), we use Lemma C.1. ■

Proof (of Theorem 4.1.) First, note that the input-output map $\boldsymbol{\mu}(t) \overset{\diamond}{\mapsto} \boldsymbol{\xi}_E(t)$ does not correspond to an LTI system (see the succeeding proof). However, if we define the effective input as

$$\mathbf{s}(t) := \boldsymbol{\mu}(t) + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}^j_{E^c}(t-t_k^j), \tag{C.18}$$

then the map $\mathbf{s}(t) \overset{\heartsuit}{\mapsto} \boldsymbol{\xi}_E(t)$ given by

$$\boldsymbol{\xi}_E(t) = \mathbf{s}(t) + (\boldsymbol{\varphi}_E * \boldsymbol{\xi}_E)(t) \tag{C.19}$$

corresponds to an LTI system.

Let $i \in D$ and $\boldsymbol{\xi}_E(t)$ be the corresponding impulse response under $\mathbf{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$. Suppose that $\mathbf{E}^i(t)$ is the system's response to $\boldsymbol{\delta}^i(t)$, the unit impulse in dimension i given by the i^{th} column of the diagonal matrix $\boldsymbol{\delta}(t)$.

Applying Eq. (C.19) on the input-output pair $(\boldsymbol{\delta}^i(t), \mathbf{E}^i(t))$,

$$\mathbf{E}^i(t) = \boldsymbol{\delta}^i(t) + (\boldsymbol{\varphi}_E * \mathbf{E}^i)(t). \tag{C.20}$$

Observe that Eq. (C.20) is a recursive equation in $\mathbf{E}^i(t)$. Substituting $\mathbf{E}^i(t)$ back into itself, we get

$$\begin{aligned}
\mathbf{E}^i(t) &= \boldsymbol{\delta}^i(t) + \boldsymbol{\varphi}_E(t) * [\boldsymbol{\delta}^i(t) + (\boldsymbol{\varphi}_E * \mathbf{E}^i)(t)] \\
&= \boldsymbol{\delta}^i(t) + \boldsymbol{\varphi}_E^i(t) + (\boldsymbol{\varphi}_E^{\otimes 2} * \mathbf{E}^i)(t) \\
&= \boldsymbol{\delta}^i(t) + \boldsymbol{\varphi}_E^i(t) + \left[\boldsymbol{\varphi}_E^{\otimes 2}\right]^i(t) + (\boldsymbol{\varphi}_E^{\otimes 3} * \mathbf{E}^i)(t) \\
&\vdots \\
&= \boldsymbol{\delta}^i(t) + \sum_{n=1}^{\infty} \left[\boldsymbol{\varphi}_E^{\otimes n}\right]^i(t) + \lim_{n\to\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \mathbf{E}^i)(t).
\end{aligned} \tag{C.21}$$

Note that $\mathbf{s}(t)$ can be expressed as

$$\mathbf{s}(t) = \sum_{i=1}^{d} (\boldsymbol{\delta}^i * s^i)(t). \tag{C.22}$$

Since $\mathbf{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ is LTI, we have

$$\boldsymbol{\xi}_E(t) = \sum_{i=1}^{d} (\mathbf{E}^i * s^i)(t). \tag{C.23}$$

Substituting Eq. (C.21) into Eq. (C.23), we then have

$$\boldsymbol{\xi}_E(t) = \sum_{i=1}^{d} \left[\boldsymbol{\delta}^i(t) + \sum_{n=1}^{\infty} \left[\boldsymbol{\varphi}_E^{\otimes n}\right]^i(t) + \lim_{n\to\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \mathbf{E}^i)(t) \right] * s^i(t). \tag{C.24}$$

We consider each convolution of the term in the right-hand side parenthesis separately.

The first term is precisely $\mathbf{s}(t)$ in Eq. (C.22).

With some algebra (see succeeding proof), the second term can be written as

$$\sum_{i=1}^{d} \sum_{n=1}^{\infty} \left[\boldsymbol{\varphi}_E^{\otimes n}\right]^i(t) * s^i(t) = (\mathbf{h}_E * \mathbf{s})(t). \tag{C.25}$$

By commutativity of convolution, the third term can be written as

$$\sum_{i=1}^{d} \lim_{n\to\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \mathbf{E}^i)(t) * s^i(t) = \lim_{n\to\infty} \boldsymbol{\varphi}_E^{\otimes n}(t) * \left[\sum_{i=1}^{d} (\mathbf{E}^i * s^i)(t) \right] = \lim_{n\to\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \boldsymbol{\xi}_E)(t),$$

where we made use of Eq. (C.23) to arrive at the last equality.

By combining the three terms, Eq. (C.24) can be written as

$$\boldsymbol{\xi}_E(t) = \mathbf{s}(t) + (\mathbf{h}_E * \mathbf{s})(t) + \lim_{n\to\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \boldsymbol{\xi}_E)(t).$$

Imposing on the previous equation the assumption $\lim_{n\to\infty} \boldsymbol{\varphi}_E^{\otimes n}(t) \to 0$, we have

$$\boldsymbol{\xi}_E(t) = \mathbf{s}(t) + \left(\sum_{i=1}^{d} \boldsymbol{\varphi}_E^{\otimes n} * \mathbf{s} \right)(t). \tag{C.26}$$

Lastly, applying the specific form of $\mathbf{s}(t)$ in Eq. (C.18), we obtain the desired formula. ■

Proof. (in Proof of Theorem 4.1, that $\boldsymbol{\mu}(t) \overset{\diamond}{\Longmapsto} \boldsymbol{\xi}_E(t)$ is not LTI, but $\mathbf{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ is.) Let $c \in \mathbb{R}$. Define $\boldsymbol{\mu}'(t) = c\boldsymbol{\mu}(t)$ and $\boldsymbol{\xi}'_E(t) = c\boldsymbol{\xi}_E(t)$.

If the system were LTI, then linearity $\boldsymbol{\mu}'(t) \overset{\diamond}{\Longmapsto} \boldsymbol{\xi}'_E(t)$ must hold. That is, Eq. (C.17) tells us

$$\boldsymbol{\xi}'_E(t) = \boldsymbol{\mu}'(t) + (\boldsymbol{\varphi}_E * \boldsymbol{\xi}'_E)(t) + \sum_{j\in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}^j(t - t_k^j), \tag{C.27}$$

However, if we multiply both sides of Eq. (C.17) by c, we see that

$$\begin{aligned} c\boldsymbol{\xi}_E(t) &= c\boldsymbol{\mu}(t) + c(\boldsymbol{\varphi}_E * \boldsymbol{\xi}_E)(t) + c \sum_{j\in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}^j(t - t_k^j) \\ \Leftrightarrow \boldsymbol{\xi}'_E(t) &= \boldsymbol{\mu}'(t) + (\boldsymbol{\varphi}_E * \boldsymbol{\xi}'_E)(t) + c \sum_{j\in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}^j(t - t_k^j) \end{aligned} \tag{C.28}$$

Unless $c = 1$ (trivial case) or $d = e$, Eq. (C.27) and Eq. (C.28) cannot hold simultaneously because of the summation over E^c.

Thus, $\boldsymbol{\mu}(t) \overset{\diamond}{\Longmapsto} \boldsymbol{\xi}_E(t)$ is not LTI unless $e = d$, which is the special case of an MBP process.

Now consider the system $\mathbf{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ as defined in Eq. (C.18). The proof of this system being LTI is a straightforward multivariate extension of the proof of Theorem 2 in [102], but we present it here for completeness.

(Linearity) This follows simply by multiplying both sides of Eq. (C.17) by a constant c.

(Time invariance) The response of $\mathbf{s}(t-t_0)$ under $\mathbf{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ is

$$\begin{aligned}
&\mathbf{s}(t-t_0) + \int_0^t \boldsymbol{\varphi}_E(t-t_0-s)\cdot\boldsymbol{\xi}_E(s)\mathrm{d}s \\
&= \mathbf{s}(t-t_0) + \int_0^{t'+t_0} \boldsymbol{\varphi}_E(t'-s)\cdot\boldsymbol{\xi}_E(s)\mathrm{d}s \\
&\overset{(a)}{=} \mathbf{s}(t-t_0) + \int_0^{t'} \boldsymbol{\varphi}_E(t'-s)\cdot\boldsymbol{\xi}_E(s)\mathrm{d}s + \int_{t'}^{t'+t_0} \boldsymbol{\varphi}_E(t'-s)\cdot\boldsymbol{\xi}_E(s)\mathrm{d}s \\
&= \mathbf{s}(t-t_0) + (\boldsymbol{\varphi}_E * \boldsymbol{\xi}_E)(t-t_0) \\
&= \boldsymbol{\xi}_E(t-t_0),
\end{aligned}$$

where in (a) we used the fact that $\boldsymbol{\varphi}_E(s) = 0$ if $s < 0$. Thus $\mathbf{s}(t) \to \boldsymbol{\xi}_E(t)$ is LTI. ■

Proof. (in Proof of Theorem 4.1, that $\sum_{i=1}^d \sum_{n=1}^\infty \left[\boldsymbol{\varphi}_E^{\otimes n}\right]^i(t) * s^i(t) = \left(\sum_{i=1}^d \boldsymbol{\varphi}_E^{\otimes n}(t) * \mathbf{s}\right)(t)$)

$$\begin{aligned}
\sum_{i=1}^d \sum_{n=1}^\infty \left[\boldsymbol{\varphi}_E^{\otimes n}\right]^i(t) * s^i(t) &= \sum_{n=1}^\infty \sum_{i=1}^d (\left[\boldsymbol{\varphi}_E^{\otimes n}\right]^i * s^i)(t) \\
&= \sum_{n=1}^\infty \left[\begin{pmatrix} \left[\varphi_E^{\otimes n}\right]^{11}(t) \\ \vdots \\ \left[\varphi_E^{\otimes n}\right]^{d1}(t) \end{pmatrix} * s^1(t) + \cdots + \begin{pmatrix} \left[\varphi_E^{\otimes n}\right]^{1d}(t) \\ \vdots \\ \left[\varphi_E^{\otimes n}\right]^{dd}(t) \end{pmatrix} * s^d(t) \right] \\
&= \sum_{n=1}^\infty \begin{pmatrix} \left[\varphi_E^{\otimes n}\right]^{11}(t) * s^1(t) & \cdots & \left[\varphi_E^{\otimes n}\right]^{1d}(t) * s^d(t) \\ \vdots & \vdots & \vdots \\ \left[\varphi_E^{\otimes n}\right]^{d1}(t) * s^1(t) & \cdots & \left[\varphi_E^{\otimes n}\right]^{dd}(t) * s^d(t) \end{pmatrix} \\
&= \sum_{n=1}^\infty \begin{pmatrix} \left[\varphi_E^{\otimes n}\right]^{11}(t) & \cdots & \left[\varphi_E^{\otimes n}\right]^{1d}(t) \\ \vdots & \vdots & \vdots \\ \left[\varphi_E^{\otimes n}\right]^{d1}(t) & \cdots & \left[\varphi_E^{\otimes n}\right]^{dd}(t) \end{pmatrix} * \begin{pmatrix} s^1(t) \\ \vdots \\ s^d(t) \end{pmatrix} \\
&= \sum_{n=1}^\infty (\boldsymbol{\varphi}_E^{\otimes n} * \mathbf{s})(t) \\
&= \left(\sum_{n=1}^\infty \boldsymbol{\varphi}_E^{\otimes n} * \mathbf{s}\right)(t).
\end{aligned}$$

■

The additional assumption $\lim_{n\to\infty} \boldsymbol{\varphi}_E^{\otimes n}(t) = 0$ introduced in Theorem 4.1 is a necessary condition for the convergence of $\sum_{n=1}^\infty \boldsymbol{\varphi}_E^{\otimes n}(t)$. Intuitively, what this entails is that the intensity contribution of late-generation offsprings has to asymptotically go to zero to achieve convergence of the infinite sum. This is discussed further in Appendix C.2.

Remark C.1 (Nonlinear Hawkes)**.** *Consider the nonlinear Hawkes process [12] with conditional intensity*

$$\boldsymbol{\lambda}^{\star}{}_{NL}(t) := \boldsymbol{\phi}\left(\int_0^{t^-} \boldsymbol{\varphi}(t-s)\cdot \mathrm{d}\mathbf{N}(s)\right), \tag{C.29}$$

where $\boldsymbol{\phi}:\mathbb{R}^d \to (\mathbb{R}^+)^d$*. Following Definition 4.1, the nonlinear* PCMHP *can be defined as the process with intensity*

$$\boldsymbol{\xi}_{NL,E}(t) := \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\boldsymbol{\lambda}^{\star}{}_{NL}(t)\middle|\mathscr{H}^{E^c}_{t^-}\right]. \tag{C.30}$$

The convolution formula in Theorem C.2 does not hold exactly for the nonlinear PCMHP*. If* $\boldsymbol{\phi}$ *is convex (concave) in all dimensions, we have an upper (lower) bound for* $\boldsymbol{\xi}_{NL,E}(t)$ *(see Appendix C.4). In either case, we can make the approximation*

$$\boldsymbol{\xi}_{NL,E}(t) \approx \boldsymbol{\phi}\left((\boldsymbol{\varphi}_E * \boldsymbol{\xi}_{NL,E})(t) + \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}_{E^c}(t-t_k^j)\right), \tag{C.31}$$

which becomes exact for linear $\boldsymbol{\phi}$ (i.e. *Theorem C.2).*

Due to the nonlinearity introduced by $\boldsymbol{\phi}$*, the linear time-invariant solution concept is inapplicable, and alternative methods must be used to solve Eq.* (C.31) *for* $\boldsymbol{\xi}_{NL,E}(t)$*. Eq.* (C.31) *can be classified as a nonlinear Volterra functional integral equation (VFIE) of the second kind [55]. VFIEs have been solved numerically using collocation methods [128] and cubic B-spline scaling functions [73], which may be applicable to solve Eq.* (C.31)*. Once* $\boldsymbol{\xi}_{NL,E}(t)$ *is obtained, the compensator* $\boldsymbol{\Xi}_{NL,E}(t)$ *can be approximated by numerical integration.*

The techniques we introduce to approximate $\boldsymbol{\xi}_E(t)$ *(Appendix C.7) and sample* PCMHP *(Appendix C.12) are not applicable for the nonlinear* PCMHP *as they rely on the impulse response solution for* $\boldsymbol{\xi}_E(t)$*. Fitting the the nonlinear* PCMHP *to data can still be done via maximum likelihood estimation using the likelihood function we introduce in Section 4.4.2.*

A more in-depth study of the nonlinear PCMHP *process, including an analysis of the tightness of Eq.* (C.31)*, regularity conditions, and efficient numerical schemes to solve Eq.* (C.31)*, is left for future work.*

Proof (in Remark C.1, that Theorem C.2 does not apply for nonlinear PCMHP) Consider the nonlinear Hawkes process and the nonlinear PCMHP with their intensities defined in Eq. (C.29) and Eq. (C.30), respectively. Plugging in Eq. (C.29) into Eq. (C.30), we have

$$\boldsymbol{\xi}_{NL,E}(t) = \mathbb{E}_{\mathscr{H}^E_{t^-}}\left[\boldsymbol{\phi}\left(\int_0^{t^-} \boldsymbol{\varphi}(t-s)\cdot \mathrm{d}\mathbf{N}(s)\right)\middle|\mathscr{H}^{E^c}_{t^-}\right]. \tag{C.32}$$

Without loss of generality, assume two cases for $\boldsymbol{\phi}$. If $\boldsymbol{\phi}$ is convex in all dimensions, then Jensen's inequality applied to each dimension yields

$$\begin{aligned}\boldsymbol{\xi}_{NL,E}(t) &\leq \boldsymbol{\phi}\left(\mathbb{E}_{\mathscr{H}_{t^-}^E}\left[\int_0^{t^-}\boldsymbol{\varphi}(t-s)\cdot \mathrm{d}\mathbf{N}(s)\middle|\mathscr{H}_{t^-}^{E^c}\right]\right)\\ &\overset{(a)}{=} \boldsymbol{\phi}\left((\boldsymbol{\varphi}_E * \boldsymbol{\xi}_{NL,E})(t) + \sum_{j\in E^c}\sum_{t_k^j<t}\boldsymbol{\varphi}_{E^c}(t-t_k^j)\right)\end{aligned} \tag{C.33}$$

where in (a) we follow the proof of Theorem C.2 to simplify the term inside $\boldsymbol{\phi}(\cdot)$

Similarly if $\boldsymbol{\phi}$ is concave in all dimensions,

$$\boldsymbol{\xi}_{NL,E}(t) \geq \boldsymbol{\phi}\left((\boldsymbol{\varphi}_E * \boldsymbol{\xi}_{NL,E})(t) + \sum_{j\in E^c}\sum_{t_k^j<t}\boldsymbol{\varphi}_{E^c}(t-t_k^j)\right). \tag{C.34}$$

We see that equality does not hold except when $\boldsymbol{\phi}$ is both convex and concave (*i.e.* linear), allowing us to combine Eq. (C.33) and Eq. (C.34) to obtain

$$\boldsymbol{\xi}_{NL,E}(t) = \boldsymbol{\phi}\left((\boldsymbol{\varphi}_E * \boldsymbol{\xi}_{NL,E})(t) + \sum_{j\in E^c}\sum_{t_k^j<t}\boldsymbol{\varphi}_{E^c}(t-t_k^j)\right), \tag{C.35}$$

which would be the corresponding convolution formula for the nonlinear PCMHP analogous to Theorem C.2.

Since Theorem 4.1 relies on expressing $\boldsymbol{\xi}_{NL,E}(t)$ as the solution of a linear time-invariant (LTI) system, the nonlinearity induced by $\boldsymbol{\phi}$ makes the LTI approach inapplicable.

■

Remark C.2. *Under a given effective input $\boldsymbol{s}(t)$, the system $\boldsymbol{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ returns the resulting intensity of the* PCMHP *process averaged over the stochastic history of the E dimensions. Given that the effective input $\boldsymbol{s}(t)$ treats the intensity contribution of events in the E^c dimensions as exogenous, only events in the E dimensions are self- and cross-exciting in $\boldsymbol{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$. In other words, the associated branching process for $\boldsymbol{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ considers the scenario where only events in the E dimensions produce offsprings.*

Under $\boldsymbol{s}(t) = \boldsymbol{\delta}^j(0)$, Eq. (C.26) *tells us that the response of the LTI system $\boldsymbol{s}(t) \overset{\heartsuit}{\Longmapsto} \boldsymbol{\xi}_E(t)$ in the i^{th} dimension is*

$$\xi_E^i(\tau) = h_E^{ij}(\tau), \tag{C.36}$$

for $\tau > 0$.

Integrating both sides of Eq. (C.36) *over $[0,t)$, we get*

$$\Xi_E^i(t) = \int_0^t h_E^{ij}(\tau)\mathrm{d}\tau.$$

Taking the limit of both sides as $t \to \infty$,

$$\Xi_E^i(\infty) = \int_0^\infty h_E^{ij}(t).$$

We see that if the integral of the right-hand side diverges to ∞, *then the expected number of events, in view of Proposition C.1, explodes as* $t \to \infty$.

In order to have a finite expected number of events in the branching process over the E dimensions, the integral of $\mathbf{h}_E(t)$ *over* $[0,\infty)$ *must be finite. Stated differently, we require* $\mathbf{h}_E \in \mathscr{L}^1(\mathbb{R}^+)^{d\times d}$.

Corollary C.1. *The* PCMHP(d,e) *compensator* $\boldsymbol{\Xi}_E(t)$ *is given by*

$$\boldsymbol{\Xi}_E(t) = [\boldsymbol{\delta}(t) + \mathbf{h}_E(t)] * \left[\mathbf{M}(t) + \sum_{j\in E^c} \sum_{t_k^j < t} \boldsymbol{\Phi}_{E^c}^j (t - t_k^j) \right], \tag{C.37}$$

where $\mathbf{h}_E(t) = \sum_{n=1}^\infty \boldsymbol{\varphi}_E^{\otimes n}(t)$ *and* $\mathbf{M}(t)$ *and* $\boldsymbol{\Phi}(t)$ *are the integral of the background intensity* $\boldsymbol{\mu}(t)$ *and the integral of the Hawkes kernel* $\boldsymbol{\varphi}(t)$ *as defined in Eq.* (C.4) *and Eq.* (C.5), *respectively.* $\boldsymbol{\Phi}_{E^c}(t)$ *is defined similar to* $\boldsymbol{\varphi}_{E^c}(t)$, *where we zero out the columns corresponding to the* E^c *dimensions.*

Proof. Integrating both sides of Eq. (4.4) over $[0, t)$, we get

$$\boldsymbol{\Xi}_E(t) = \int_0^t \left[[\boldsymbol{\delta}(\tau) + \mathbf{h}_E(\tau)] * \left[\boldsymbol{\mu}(\tau) + \sum_{j\in E^c} \sum_{t_k^j < \tau} \boldsymbol{\varphi}_{E^c}^j (\tau - t_k^j) \right] \right] \mathrm{d}\tau.$$

For brevity, set $\mathbf{s}(t)$ as in Eq. (C.18), and set $\mathbf{S}(t) = \int_0^t \boldsymbol{s}(\tau)\mathrm{d}\tau$. Now, we focus on the i^{th}

entry of both sides. Expanding the convolution, we get

$$
\begin{aligned}
\Xi_E^i(t) &= \int_0^t \left[s^i(\tau) + \sum_{j=1}^d (h_E^{ij} * s^j)(\tau) \right] \mathrm{d}\tau \\
&\overset{(a)}{=} \int_0^t s^i(\tau)\mathrm{d}\tau + \sum_{j=1}^d \int_0^t \left[(h_E^{ij} * s^j)(\tau) \right] \mathrm{d}\tau \\
&\overset{(b)}{=} S^i(t) + \sum_{j=1}^d \int_0^t \int_0^\tau \left[h_E^{ij}(v) \cdot s^j(\tau - v) \right] \mathrm{d}v\mathrm{d}\tau \\
&\overset{(c)}{=} S^i(t) + \sum_{j=1}^d \int_0^t \int_v^t \left[h_E^{ij}(v) \cdot s^j(\tau - v) \right] \mathrm{d}\tau\mathrm{d}v \\
&= S^i(t) + \sum_{j=1}^d \int_0^t h_E^{ij}(v) \int_v^t s^i(\tau - v)\mathrm{d}\tau\mathrm{d}v \\
&\overset{(d)}{=} S^i(t) + \sum_{j=1}^d \int_0^t h_E^{ij}(v) \int_0^{t-v} s^i(u)\mathrm{d}u\mathrm{d}v \\
&= S^i(t) + \sum_{j=1}^d \int_0^t h_E^{ij}(v) \cdot S^i(t - v)\mathrm{d}v \\
&= S^i(t) + \sum_{j=1}^d (h_E^{ij} * S^i)(t),
\end{aligned}
$$

where in (a) we applied the linearity of integration, (b) we used the definition of convolution, (c) reversed the order of integration, and (d) made a change of variable $u = \tau - v$.

Collecting the results for every dimension i, we arrive at the desired formula. ■

C.4.2 Regularity Conditions

The following theorem presents the condition on the branching matrix $\boldsymbol{\alpha}$ that ensures $\mathscr{L}^1$-convergence of $\mathbf{h}_E(t) = \sum_{n=1}^\infty \boldsymbol{\varphi}_E^{\otimes n}(t)$ appearing in the PCMHP intensity in Theorem 4.1. As discussed in Remark C.2, $\mathscr{L}^1$-convergence of $\mathbf{h}_E(t)$ guarantees that the dynamics under the branching process on the E dimensions is nonexplosive and that we have a finite expected number of events.

Theorem C.3. *If the branching submatrix $\boldsymbol{\alpha}^{EE}$ satisfies $\rho(\boldsymbol{\alpha}^{EE}) < 1$, then $\mathbf{h}_E \in \mathscr{L}^1(\mathbb{R}^+)^{d\times d}$.*

To prove Theorem C.3, we need the following result on convergence in $\mathscr{L}^p$ spaces.

Theorem C.4 ([35])**.** *Let $\mathscr{L}^p(\mathbb{R}^+)$ is the space of functions $f : \mathbb{R}^+ \to \mathbb{R}$ that have finite p-norm. If $1 \le p < \infty$, every absolutely convergent series in $\mathscr{L}^p(\mathbb{R}^+)$ converges.*

The following preliminary result is an upper bound on the 1-norm of each entry of $\boldsymbol{\varphi}^{\otimes n}(t)$ expressed in terms of the n^{th} power of the branching matrix $\boldsymbol{\alpha}$. We use this upper bound in the proof of Theorem C.3, which identifies the condition on the branching matrix $\boldsymbol{\alpha}$ for a convergent $\mathbf{h}_E(t)$.

Lemma C.2. *Let $\boldsymbol{\alpha} = (\alpha^{ij}) \in (\mathbb{R}^+)^{e\times e}$ and $\boldsymbol{\varphi}(t) = (\varphi^{ij}(t)) \in (\mathbb{R}^+)^{e\times e}$ be a matrix satisfying $\varphi^{ij}(t) = \alpha^{ij} f^{ij}(t)$, $f^{ij}(t) \geq 0$ and $\int_0^\infty f^{ij}(t)\mathrm{d}t = 1$. Then for $n \geq 1$,*

$$\mathbf{b}_n \leq \boldsymbol{\alpha}^n,$$

where $\mathbf{b}_n = (b_n^{ij}) \in (\mathbb{R}^+)^{e\times e}$ and $b_n^{ij} = \|(\varphi^{\otimes n})^{ij}\|_1$.

Proof. We proceed by induction. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\varphi}(t)$ be as stated. Let $(i, j) \in E \times E$.

Suppose $n = 1$.

$$\begin{aligned} b_1^{ij} &= \|(\boldsymbol{\varphi}^{\otimes 1})^{ij}\|_1 \\ &= \|\varphi^{ij}\|_1 \\ &\overset{(a)}{=} \|\alpha^{ij} f^{ij}\|_1 \\ &\overset{(b)}{=} \alpha^{ij}\|f^{ij}\|_1 \\ &\overset{(c)}{=} \alpha^{ij}, \end{aligned}$$

where (a) and (c) follow from the definitions of φ^{ij} and f^{ij}, and (b) follows from α^{ij} being a constant. Thus, $\mathbf{b}_n \leq \boldsymbol{\alpha}^n$ holds for $n = 1$.

Suppose that the relation holds for $n = k \in \mathbb{N}$. That is, $\mathbf{b}_k \leq \boldsymbol{\alpha}^k$. This means that for every $(p, q) \in E \times E$, the following holds:

$$\|(\boldsymbol{\varphi}^{\otimes k})^{pq}\|_1 \leq (\boldsymbol{\alpha}^k)^{pq}. \tag{C.38}$$

Now, we show that the relation holds for $n = k+1$. We observe that

$$\begin{aligned}
b_{k+1}^{ij} &= \|(\boldsymbol{\varphi}^{\otimes k+1})^{ij}\|_1 \\
&= \|(\boldsymbol{\varphi}^{\otimes k} * \boldsymbol{\varphi})^{ij}\|_1 \\
&\overset{(a)}{=} \left\| \sum_l (\boldsymbol{\varphi}^{\otimes k})^{il} * \boldsymbol{\varphi}^{lj} \right\|_1 \\
&\overset{(b)}{\leq} \sum_l \left\| (\boldsymbol{\varphi}^{\otimes k})^{il} * \boldsymbol{\varphi}^{lj} \right\|_1 \\
&\overset{(c)}{\leq} \sum_l \left\| (\boldsymbol{\varphi}^{\otimes k})^{il} \right\|_1 \left\| \boldsymbol{\varphi}^{lj} \right\|_1 \\
&= \sum_l \left\| (\boldsymbol{\varphi}^{\otimes k})^{il} \right\|_1 \left\| \alpha^{lj} f^{lj} \right\|_1 \\
&= \sum_l \left\| (\boldsymbol{\varphi}^{\otimes k})^{il} \right\|_1 \alpha^{lj} \left\| f^{lj} \right\|_1 \\
&= \sum_l \left\| (\boldsymbol{\varphi}^{\otimes k})^{il} \right\|_1 \alpha^{lj} \\
&\overset{(d)}{\leq} \sum_l (\boldsymbol{\alpha}^k)^{il} \alpha^{lj} \\
&= (\boldsymbol{\alpha}^{k+1})^{ij}
\end{aligned}$$

In (a), the definition of the matrix product was used. In (b) and (c), the Minkowski inequality and Young's convolution inequality were used, respectively. In (d), our induction hypothesis Eq. (C.38) was used.

By induction, $\mathbf{b}_n \leq \boldsymbol{\alpha}^n$ holds for $n \in \mathbb{N}$. ■

We are now ready to prove Theorem C.3.

Proof. (of Theorem C.3). Let $\rho(\boldsymbol{\alpha}^{EE}) < 1$.

If we are able to show that the series $\mathbf{h}_E = \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}$ is absolutely convergent in $\mathscr{L}^1(\mathbb{R}^+)^{d\times d}$, then Theorem C.4 tells us that $\mathbf{h}_E \in \mathscr{L}^1(\mathbb{R}^+)^{d\times d}$.

Our task is then to prove absolute convergence of every entry in $\mathbf{h}_E$. If we denote $\mathbf{h}_E^k$ to be the k^{th} partial sum of the series $\mathbf{h}_E$, *i.e.*

$$\mathbf{h}_E^k(t) = \sum_{n=1}^{k} \boldsymbol{\varphi}_E^{\otimes n}(t),$$

then what we need to show is

$$\lim_{k\to\infty} \|(h_E^k)^{ij}\|_1 < \infty,$$

for $(i, j) \in D \times D$.

Define $\boldsymbol{\varphi}^{EE} \in \mathscr{L}^1(\mathbb{R}^+)^{e\times e}$ and $\boldsymbol{\varphi}^{E^cE} \in \mathscr{L}^1(\mathbb{R}^+)^{(d-e)\times e}$ as the submatrices of $\boldsymbol{\varphi}$ given by

$$\boldsymbol{\varphi}^{EE} = \begin{pmatrix} \alpha^{11} f^{11} & \dots & \alpha^{1e} f^{1e} \\ \vdots & \vdots & \vdots \\ \alpha^{e1} f^{e1} & \dots & \alpha^{ee} f^{ee} \end{pmatrix},$$

$$\boldsymbol{\varphi}^{E^cE} = \begin{pmatrix} \alpha^{e+1,1} f^{e+1,1} & \dots & \alpha^{e+1,e} f^{e+1,e} \\ \vdots & \vdots & \vdots \\ \alpha^{d1} f^{d1} & \dots & \alpha^{de} f^{de} \end{pmatrix}.$$

$\boldsymbol{\varphi}_E$ may be written in block matrix form as

$$\boldsymbol{\varphi}_E = \left[\begin{array}{c|c} \boldsymbol{\varphi}^{EE} & \mathbf{0} \\ \hline \boldsymbol{\varphi}^{E^cE} & \mathbf{0} \end{array}\right].$$

Convolving $\boldsymbol{\varphi}_E$ with itself, we see that $\boldsymbol{\varphi}_E^{\otimes 2}$ can be written as

$$\boldsymbol{\varphi}_E^{\otimes 2} = \left[\begin{array}{c|c} (\boldsymbol{\varphi}^{EE})^{\otimes 2} & \mathbf{0} \\ \hline \boldsymbol{\varphi}^{E^cE} * \boldsymbol{\varphi}^{EE} & \mathbf{0} \end{array}\right].$$

Convolving n times, we arrive at

$$\boldsymbol{\varphi}_E^{\otimes n} = \left[\begin{array}{c|c} (\boldsymbol{\varphi}^{EE})^{\otimes n} & \mathbf{0} \\ \hline \boldsymbol{\varphi}^{E^cE} * (\boldsymbol{\varphi}^{EE})^{\otimes n-1} & \mathbf{0} \end{array}\right]. \tag{C.39}$$

For $(i,j) \in E \times E$, Eq. (C.39) states that

$$(\boldsymbol{\varphi}_E^{\otimes n})^{ij} = ((\boldsymbol{\varphi}^{EE})^{\otimes n})^{ij} \tag{C.40}$$

Taking the 1-norm of both sides, we see that

$$\|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1 = \|((\boldsymbol{\varphi}^{EE})^{\otimes n})^{ij}\|_1 \leq ((\boldsymbol{\alpha}^{EE})^n)^{ij}, \tag{C.41}$$

where the last inequality is due to Lemma C.2.

Similarly, for $(i,j) \in E^c \times E$,

$$\begin{aligned}
\|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1 &= \|(\boldsymbol{\varphi}^{E^cE} * (\boldsymbol{\varphi}^{EE})^{\otimes n-1})^{i-e,j}\|_1 \\
&\overset{(a)}{=} \|(\boldsymbol{\varphi}^{E^cE} * (\boldsymbol{\varphi}^{EE})^{\otimes n-1})^{i'j}\|_1 \\
&\overset{(b)}{=} \left\| \sum_l (\boldsymbol{\varphi}^{E^cE})^{i'l} * ((\boldsymbol{\varphi}^{EE})^{\otimes n-1})^{lj} \right\|_1 \\
&\overset{(c)}{\leq} \sum_l \left\| (\boldsymbol{\varphi}^{E^cE})^{i'l} * ((\boldsymbol{\varphi}^{EE})^{\otimes n-1})^{lj} \right\|_1 \\
&\overset{(d)}{\leq} \sum_l \left\| (\boldsymbol{\varphi}^{E^cE})^{i'l} \right\|_1 \left\| ((\boldsymbol{\varphi}^{EE})^{\otimes n-1})^{lj} \right\|_1 \\
&\overset{(e)}{\leq} \sum_l \left\| (\boldsymbol{\varphi}^{E^cE})^{i'l} \right\|_1 ((\boldsymbol{\alpha}^{EE})^{n-1})^{lj} \\
&\overset{(f)}{=} \sum_l (\boldsymbol{\alpha}^{E^cE})^{i'l} ((\boldsymbol{\alpha}^{EE})^{n-1})^{lj} \\
&= (\boldsymbol{\alpha}^{E^cE} (\boldsymbol{\alpha}^{EE})^{n-1})^{i'j} \qquad \text{(C.42)} \\
&= (\boldsymbol{\alpha}^{E^cE} (\boldsymbol{\alpha}^{EE})^{n-1})^{i-e,j} \qquad \text{(C.43)}
\end{aligned}$$

In (a), we reindex $i' = i - e$ to start our index at 1. Lines (b), (c), (d) are applications of the definition of matrix product, Minkowski's inequality, and Young's convolution inequality, respectively. In (e), we apply Lemma C.2, and in (f), we apply the definition of $\boldsymbol{\varphi}^{E^cE}$.

Putting in Eq. (C.41) and Eq. (C.43) into Eq. (C.39), we see that

$$(\|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1) \leq \left[\begin{array}{c|c} (\boldsymbol{\alpha}^{EE})^n & \mathbf{0} \\ \hline \boldsymbol{\alpha}^{E^cE} (\boldsymbol{\alpha}^{EE})^{n-1} & \mathbf{0} \end{array}\right]. \qquad \text{(C.44)}$$

Taking $\sum_{n=1}^{\infty}$ of both sides, we get

$$\begin{aligned}
\sum_{n=1}^{\infty} (\|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1) &\leq \left[\begin{array}{c|c} \sum_{n=1}^{\infty} (\boldsymbol{\alpha}^{EE})^n & \mathbf{0} \\ \hline \boldsymbol{\alpha}^{E^cE} \sum_{n=1}^{\infty} (\boldsymbol{\alpha}^{EE})^{n-1} & \mathbf{0} \end{array}\right] \\
&= \left[\begin{array}{c|c} (\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} - \mathbf{I} & \mathbf{0} \\ \hline \boldsymbol{\alpha}^{E^cE} (\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} & \mathbf{0} \end{array}\right]. \qquad \text{(C.45)}
\end{aligned}$$

where the previous equation follows from $\rho(\boldsymbol{\alpha}^{EE}) < 1$ and Proposition C.2.

Inspecting every block of Eq. (C.45), it is clear that that for all $(i,j) \in D \times D$,

$$\sum_{n=1}^{\infty} \|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1 < \infty. \qquad \text{(C.46)}$$

Given that

$$\|(h_E^k)^{ij}\|_1 \leq \sum_{n=1}^{k} \|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1 \xrightarrow[k\to\infty]{} \sum_{n=1}^{\infty} \|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1 < \infty, \tag{C.47}$$

we have proven that every entry in $\mathbf{h}_E$ is absolutely convergent in $\mathscr{L}^1(\mathbb{R}^+)$. ■

Proof (of Theorem 4.2.) Consider a PCMHP(d,e) process with conditional intensity $\boldsymbol{\xi}_E(t)$ given by Eq. (4.5). If $\rho(\boldsymbol{\alpha}^{EE}) < 1$, $\sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}$ is a convergent function by Theorem C.3, so $\boldsymbol{\xi}_E(t)$ is well-defined.

We proceed by noting that the intensity $\boldsymbol{\xi}_E(t)$ splits as the sum of three distinct terms:

1. an inhomogeneous Poisson process rate:

$$\boldsymbol{\mu}(t) + \sum_{n=1}^{\infty} (\varphi_E^{\otimes n} * \boldsymbol{\mu})(t),$$

2. a multivariate Hawkes intensity:

$$\sum_{j\in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}^j(t - t_k^j),$$

3. a convolution term involving the Hawkes intensity:

$$\sum_{n=1}^{\infty} \varphi_E^{\otimes n} * \sum_{j\in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}^j(t - t_k^j).$$

We inspect each of these intensities separately.

For the process with intensity $\boldsymbol{\xi}_E(t)$ to be subcritical, the processes corresponding to each of these three intensities necessarily have to be subcritical.

Intensity (A). A Poisson process is independent of the arrival of new events, so the process is subcritical as long as $\boldsymbol{\mu}(t)$ is a bounded function and $\sum_{n=1}^{\infty} \varphi_E^{\otimes n}$ is $\mathscr{L}^1$-convergent.

Intensity (B). The multivariate Hawkes process corresponding to intensity (B) may be viewed as a process with branching matrix

$$\boldsymbol{\alpha}_{E^c} = \left(\begin{array}{c|c} \mathbf{0} & \boldsymbol{\alpha}^{EE^c} \\ \hline \mathbf{0} & \boldsymbol{\alpha}^{E^cE^c} \end{array}\right).$$

By Theorem C.1, this process is subcritical if $\rho(\boldsymbol{\alpha}_{E^c}) < 1$. The eigenvalues of $\boldsymbol{\alpha}_{E^c}$ are the solutions λ of

$$\det(\boldsymbol{\alpha}_{E^c} - \lambda \mathbf{I}) = 0. \tag{C.48}$$

Expanding the left-hand side of Eq. (C.48),

$$\begin{aligned}\det(\boldsymbol{\alpha}_{E^c} - \lambda \mathbf{I}) &= \left[\begin{array}{c|c} \mathbf{0} - \lambda \mathbf{I} & \boldsymbol{\alpha}^{EE^c} \\ \hline \mathbf{0} & \boldsymbol{\alpha}^{E^cE^c} - \lambda \mathbf{I} \end{array}\right] = \left[\begin{array}{c|c} -\lambda \mathbf{I} & \boldsymbol{\alpha}^{EE^c} \\ \hline \mathbf{0} & \boldsymbol{\alpha}^{E^cE^c} - \lambda \mathbf{I} \end{array}\right] \\ &= \det(-\lambda \mathbf{I}) \det(\boldsymbol{\alpha}^{E^cE^c} - \lambda \mathbf{I}) = (-1)^{|E|} \lambda^{|E|} \det(\boldsymbol{\alpha}^{E^cE^c} - \lambda \mathbf{I}),\end{aligned}$$

we see that the eigenvalues of $\boldsymbol{\alpha}_{E^c}$ are precisely 0 and the eigenvalues of $\boldsymbol{\alpha}^{E^cE^c}$. Thus, $\rho(\boldsymbol{\alpha}_{E^c}) = \rho(\boldsymbol{\alpha}^{E^cE^c})$

By Theorem C.1 and given $\rho(\boldsymbol{\alpha}^{E^cE^c}) < 1$, the multivariate Hawkes process with intensity (B) is subcritical.

Intensity (C). Let $(i, j) \in D \times D$. We can see that intensity (C) is intensity (B) (with branching matrix $\boldsymbol{\alpha}_{E^c}$) convolved with the infinite sum $\sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}$.

First, consider the matrix

$$\boldsymbol{\Omega} := \left[\begin{array}{c|c} (\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} - \mathbf{I} & \mathbf{0} \\ \hline \boldsymbol{\alpha}^{E^cE} (\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} & \mathbf{0} \end{array}\right]. \tag{C.49}$$

In Eq. (C.45), we showed that

$$\sum_{n=1}^{\infty} (\|(\boldsymbol{\varphi}_E^{\otimes n})^{ij}\|_1) \leq \boldsymbol{\Omega}. \tag{C.50}$$

Now, consider the (i, j) branching factor of the process with intensity (C), which can be expressed as

$$\left\| \sum_k \left(\sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n} \right)^{ik} * \boldsymbol{\varphi}^{kj} \right\|_1. \tag{C.51}$$

Applying (a) Young's convolution inequality, (b) Minkowski's inequality, and (c) the matrix

upper bound in Eq. (C.50), we see that Eq. (C.51) is upper bounded by:

$$
\begin{aligned}
&\left\| \sum_k \left(\sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n} \right)^{ik} * \boldsymbol{\varphi}^{kj} \right\|_1 \\
&\overset{(a)}{\leq} \sum_k \left\| \left(\sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n} \right)^{ik} \right\|_1 \left\| \boldsymbol{\varphi}^{kj} \right\|_1 = \sum_k \left\| \left(\sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n} \right)^{ik} \right\|_1 \alpha^{kj} = \sum_k \left\| \sum_{n=1}^{\infty} \left(\boldsymbol{\varphi}_E^{\otimes n} \right)^{ik} \right\|_1 \alpha^{kj} \\
&\overset{(b)}{\leq} \sum_k \sum_{n=1}^{\infty} \left\| \left(\boldsymbol{\varphi}_E^{\otimes n} \right)^{ik} \right\|_1 \alpha^{kj} \\
&\overset{(c)}{\leq} \sum_k \Omega^{ik} \alpha^{kj} = (\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c})^{ij}.
\end{aligned}
$$

This tells us that the (i, j) branching factor of the process with intensity (C) is upper-bounded by the (i, j) branching factor of a multivariate Hawkes process with branching matrix $\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}$. Thus, if we are able to show that the multivariate Hawkes process with branching matrix $\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}$ is subcritical, *i.e.*, $\rho(\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}) < 1$, the process defined by intensity (C) is necessarily subcritical as well.

$\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}$ in block form may be written as:

$$
\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c} = \left[\begin{array}{c|c} (\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} - \mathbf{I} & \mathbf{0} \\ \hline \boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} & \mathbf{0} \end{array}\right] \left[\begin{array}{c|c} \mathbf{0} & \boldsymbol{\alpha}^{EE^c} \\ \hline \mathbf{0} & \boldsymbol{\alpha}^{E^cE^c} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{0} & [(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} - \mathbf{I}]\boldsymbol{\alpha}^{EE^c} \\ \hline \mathbf{0} & \boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{E^cE^c} \end{array}\right]
$$

The eigenvalues of $\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}$ are the solutions λ of

$$
\det(\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c} - \lambda\mathbf{I}) = 0. \tag{C.52}
$$

Expanding the left-hand side of Eq. (C.52), we see that

$$
\begin{aligned}
\det(\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c} - \lambda\mathbf{I}) &= \det \left[\begin{array}{c|c} \mathbf{0} - \lambda\mathbf{I} & [(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1} - \mathbf{I}]\boldsymbol{\alpha}^{EE^c} \\ \hline \mathbf{0} & \boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{E^cE^c} - \lambda\mathbf{I} \end{array}\right] \\
&= \det(-\lambda\mathbf{I}) \det(\boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{E^cE^c} - \lambda\mathbf{I}) \\
&= (-1)^{|E|}\lambda^{|E|} \det(\boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{E^cE^c} - \lambda\mathbf{I}).
\end{aligned}
$$

Thus we see that the eigenvalues of $\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}$ are precisely 0 and the eigenvalues of $\boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{E^cE^c}$. This implies that

$$
\rho(\boldsymbol{\Omega}\boldsymbol{\alpha}_{E^c}) = \rho(\boldsymbol{\alpha}^{E^cE}(\mathbf{I} - \boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{E^cE^c}). \tag{C.53}
$$

By Theorem C.1, if $\rho(\boldsymbol{\alpha}^{E^cE}(\mathbf{I}-\boldsymbol{\alpha}^{EE})^{-1}\boldsymbol{\alpha}^{EE^c})<1$, the process with intensity (C) is subcritical.

As the three sub-intensities that consist $\boldsymbol{\xi}_E(t)$ all correspond to subcritical processes given the assumed conditions, $\boldsymbol{\xi}_E(t)$ is then subcritical. ■

Corollary C.2. *The regularity conditions for* PCMHP(d,e) *in Theorem 4.1 cover the Hawkes process and the MBP process as special cases.*

Proof. If $e=0$ (Hawkes), we have $\boldsymbol{\alpha}^{EE}=\boldsymbol{\alpha}^{EE^c}=\boldsymbol{\alpha}^{E^cE}=\mathbf{0}$ and $\boldsymbol{\alpha}^{E^cE^c}=\boldsymbol{\alpha}$. Thus, only the second condition is non-trivially satisfied, yielding $\rho(\boldsymbol{\alpha})<1$. If $e=d$ (MBP), we have $\boldsymbol{\alpha}^{E^cE^c}=\boldsymbol{\alpha}^{EE^c}=\boldsymbol{\alpha}^{E^cE}=\mathbf{0}$ and $\boldsymbol{\alpha}^{EE}=\boldsymbol{\alpha}$. Thus, only the first condition is non-trivially satisfied, yielding $\rho(\boldsymbol{\alpha})<1$. If $0<e<d$ (PCMHP), all branching submatrices are possibly nonzero and all three conditions are potentially non-trivially satisfied. ■

C.5 Additional Results and Proofs for PCMHP Inference

Proof (of Theorem 4.3). Since $E \supseteq Q$, $\{1,\ldots,e\}$ splits as $\{1,\ldots,q\} \cup \{q+1,\ldots,e\}$. Given this observation, the joint probability of the event volumes $\bigcup_{j\in Q}\{\mathsf{C}_k^j\}_{k=1}^{n^j}$ and event times $\mathscr{H}_{T^-}^{Q^c}$,

$$\mathbb{P}\left\{\mathsf{C}_1^1,\ldots,\mathsf{C}_{n^1}^1,\ldots,\mathsf{C}_1^q,\ldots,\mathsf{C}_{n^q}^q,\mathscr{H}_{T^-}^{q+1},\ldots,\mathscr{H}_{T^-}^{d}\right\},$$

can be written as

$$\mathbb{P}\left\{\mathsf{C}_1^1,\ldots,\mathsf{C}_{n^1}^1,\ldots,\mathsf{C}_1^q,\ldots,\mathsf{C}_{n^q}^q,\mathscr{H}_{T^-}^{q+1},\ldots,\mathscr{H}_{T^-}^{e},\mathscr{H}_{T^-}^{e+1},\ldots,\mathscr{H}_{T^-}^{d}\right\}. \tag{C.54}$$

Conditioning on the E^c event times $\left\{\mathscr{H}_{T^-}^{e+1},\ldots,\mathscr{H}_{T^-}^{d}\right\}$, Eq. (C.54) splits as

$$\mathbb{P}\left\{\mathsf{C}_1^1,\ldots,\mathsf{C}_{n^1}^1,\ldots,\mathsf{C}_1^q,\ldots,\mathsf{C}_{n^q}^q,\mathscr{H}_{T^-}^{q+1},\ldots,\mathscr{H}_{T^-}^{e}|\mathscr{H}_{T^-}^{e+1},\ldots,\mathscr{H}_{T^-}^{d}\right\}\cdot\mathbb{P}\left\{\mathscr{H}_{T^-}^{e+1},\ldots,\mathscr{H}_{T^-}^{d}\right\}. \tag{C.55}$$

In a PCMHP(d,e) model, events in dimension $j \in E$ are independent of events in dimension $k \neq j, k \in E$, which implies that Eq. (C.55) splits as a product over dimensions,

$$\prod_{j=1}^{q}\mathbb{P}\left\{\mathsf{C}_1^j,\ldots,\mathsf{C}_{n^j}^j\,\middle|\,\mathscr{H}_{T^-}^{E^c}\right\}\cdot\prod_{j=q+1}^{e}\mathbb{P}\left\{\mathscr{H}_{T^-}^{j}\,\middle|\,\mathscr{H}_{T^-}^{E^c}\right\}\cdot\mathbb{P}\left\{\mathscr{H}_{T^-}^{e+1},\ldots,\mathscr{H}_{T^-}^{d}\right\}. \tag{C.56}$$

Taking $-\log(\cdot)$ of both sides of Eq. (C.56) and converting the product to a sum over logarithms, we get

$$\begin{aligned}\mathscr{L}(\boldsymbol{\Theta};T) = &-\sum_{j=1}^{q}\log\mathbb{P}\left\{\mathsf{C}_1^j,\ldots,\mathsf{C}_{n^j}^j\,\middle|\,\mathscr{H}_{T^-}^{E^c};\boldsymbol{\Theta}\right\}-\sum_{j=q+1}^{e}\log\mathbb{P}\left\{\mathscr{H}_{T^-}^{j}\,\middle|\,\mathscr{H}_{T^-}^{E^c};\boldsymbol{\Theta}\right\}\\ &-\sum_{j=e+1}^{d}\log\mathbb{P}\left\{\mathscr{H}_{T^-}^{j};\boldsymbol{\Theta}\right\}\end{aligned} \tag{C.57}$$

Using Proposition C.3 on the first term and and Eq. (C.7) on the second and third terms, we arrive at the desired formula. ■

Remark C.3. ***Joint fitting.*** *To jointly fit multiple ($R > 1$)* PCMHP(d,e) *realizations, one needs to maximize the sum of log-likelihoods correponding to each realization. That is,*

$$\mathscr{L}(\boldsymbol{\Theta};T) = \sum_{r=1}^{R}\mathscr{L}^r(\boldsymbol{\Theta};T), \tag{C.58}$$

where $\mathscr{L}^r(\boldsymbol{\Theta};T)$ is the negative log-likelihood corresponding to the r^{th} realization.

Additional details on runtime complexity. Evaluating $\mathscr{L}(\boldsymbol{\Theta}; T)$ has a runtime complexity of

$$\mathscr{O}\left((C + \mathrm{n}^{E^c}) \cdot (\mathrm{n}^{Q} + \mathrm{n}^{Q^c})\right); \tag{C.59}$$

which comes from combining the time complexity of calculating the $\mathscr{L}_{\text{IC-LL}}$ loss $\mathscr{O}(\mathrm{n}^{Q}(C + \mathrm{n}^{E^c}))$ and the $\mathscr{L}_{\text{PP-LL}}$ loss $\mathscr{O}(\mathrm{n}^{Q^c}(C + \mathrm{n}^{E^c}))$. C denotes a constant independent of the dimension of the PCMHP and the data. Having closed-form expressions for convolutions on $\mathbf{h}_E(t)$ simplifies the convolutions in $\xi_E^j(\cdot)$ (see Eq. (4.5)) and $\Xi_E^j(\cdot)$ (see Eq. (C.37)) to function evaluations in $\mathscr{O}(1)$. We sketch the proof below.

To calculate $\mathscr{L}_{\text{IC-LL}}(\boldsymbol{\Theta}; T)$, we iterate over each dimension $j \in Q$ and every observation interval $[o_{k-1}^j, o_k^j)$, requiring n^Q loops. For each iteration, we calculate $\Xi_E^j(o_{k-1}^j, o_k^j)$, which has time complexity $\mathscr{O}(\mathrm{n}^{E^c})$, since we need to calculate the influence of each observed event in E^c. Thus, the total time complexity is $\mathscr{O}(\mathrm{n}^{Q} \cdot (C + \mathrm{n}^{E^c}))$, where C accounts for a constant number of calculations independent of the data (ex. calculating the baseline intensity and $\mathbf{h}_E(t)$).

Calculations for $\mathscr{L}_{\text{PP-LL}}(\boldsymbol{\Theta}; T)$ are similar. Instead of iterating over $j \in Q$ and C_k^j, we iterate over $j \in Q^c$ and $t_k^j \in \mathscr{H}_{T^-}^j$, requiring n^{Q^c} loops. Within each loop, we calculate $\xi_E^j(t_k^j)$, which has time complexity $\mathscr{O}(\mathrm{n}^{E^c})$. We note that this dominates the time complexity $\mathscr{O}(|Q| \cdot \mathrm{n}^{E^c})$ of calculating $\Xi_E^j(T; \boldsymbol{\Theta})$ in each loop. Thus, the total time complexity is $\mathscr{O}(\mathrm{n}^{Q^c} \cdot (C + \mathrm{n}^{E^c}))$.

In the case $E = Q = \emptyset$, the runtime complexity reduces to $\mathscr{O}((\mathrm{n}^{E^c})^2)$, which is consistent with the multivariate Hawkes process. If $E = Q = D$, runtime complexity reduces to $\mathscr{O}(\mathrm{n}^{E})$, which is consistent with the MBP (i.e. Poisson) process.

Remark C.4. *In general, the negative log-likelihood $\mathscr{L}(\boldsymbol{\Theta}; T)$ is nonconvex in $\boldsymbol{\Theta}$ due to the presence of nonlinear parameters, e.g. θ^{ij} for the exponential kernel (see Appendix C.6 for a special case of* PCMHP *(2,1) where we have a convex likelihood). Due to this, gradient-based algorithms have no guarantee of attaining the global optimum. Convexity analysis for the general form of $\mathscr{L}(\boldsymbol{\Theta}; T)$ is difficult due to the complexity of evaluating $\boldsymbol{h}_E(t)$ in a high-dimensional setting. In Appendix C.6, we prove convexity for a special case of* PCMHP$(2,1)$. *For nonconvex cases, global optimization (such as the Lipschitz Global Optimizer* LGO *[91]) can be leveraged to find the global optimum, for instance using a branch-and-bound or randomized search strategy over the parameter space or iterated local optimization.*

C.6 Convexity Analysis of the PCMHP$(2,1)$ Likelihood

Consider a PCMHP(2,1) process. Suppose the associated kernel can be expressed as $\varphi^{ij}(t) = \alpha^{ij} f^{ij}(t;\theta^{ij})$, where $f^{ij}(t;\theta^{ij})$ is a probability density over $[0,\infty)$ parametrized by θ^{ij}. Given a multiple impulse exogenous input function,

$$\boldsymbol{\mu}(t) = \sum_i \begin{pmatrix} a_i \cdot \delta(t-c_i) \\ b_i \cdot \delta(t-d_i) \end{pmatrix},$$

we show that for a fixed set of $\{\theta^{ij}\}$ parameters and given observed point data $\Psi^1 = \{s_k^1\}$ and $\Psi^2 = \{s_k^2\}$ over $[0,T)$ for dimensions 1 and 2, the negative point-process log-likelihood (PP-PP NLL) of the PCMHP(2,1) is convex in the $\{\alpha^{ij}\}$ parameters.

We first consider a single impulse as the exogenous input. Suppose

$$\boldsymbol{\mu}(t) = \begin{pmatrix} a \cdot \delta(t-c) \\ b \cdot \delta(t-d) \end{pmatrix},$$

where $a,b,c,d \geq 0$. Given this general impulse exogenous input, we aim to show that the PP-PP NLL is convex in the $\{\alpha^{ij}\}$ parameters. Given observed data points $\Psi^1 = \{s_k^1\}$ and $\Psi^2 = \{s_k^2\}$ over $[0,T)$ in the two dimensions, the PP-PP NLL is given by

$$\mathscr{L}(\boldsymbol{\alpha};T) = -\log \sum_{s_k^1} (\xi_1^1(s_k^1;\boldsymbol{\alpha})) - \log \sum_{s_k^2} (\xi_1^2(s_k^2;\boldsymbol{\alpha})) + \Xi_1^1(T;\boldsymbol{\alpha}) + \Xi_1^2(T;\boldsymbol{\alpha}), \tag{C.60}$$

where

$$\xi_1^1(t;\boldsymbol{\alpha}) = a\delta(t-c) + a h_1^{11}(t-c)[\![t>c]\!] + \sum_{t_k^2<t} \varphi_2^{12}(t-t_k^2) + h_1^{11}(t) * \sum_{t_k^2<t} \varphi_2^{12}(t-t_k^2) \tag{C.61}$$

$$\xi_1^2(t;\boldsymbol{\alpha}) = b\delta(t-d) + b h_1^{21}(t-d)[\![t>d]\!] + \sum_{t_k^2<t} \varphi_2^{22}(t-t_k^2) + h_1^{21}(t) * \sum_{t_k^2<t} \varphi_2^{12}(t-t_k^2) \tag{C.62}$$

$$\Xi_1^1(t;\boldsymbol{\alpha}) = a[\![t>c]\!] + a H_1^{11}(t-c) + \sum_{t_k^2<t} \Phi_2^{12}(t-t_k^2) + h_1^{11}(t) * \sum_{t_k^2<t} \Phi_2^{12}(t-t_k^2) \tag{C.63}$$

$$\Xi_1^2(t;\boldsymbol{\alpha}) = b[\![t>d]\!] + b H_1^{21}(t-d) + \sum_{t_k^2<t} \Phi_2^{22}(t-t_k^2) + h_1^{21}(t) * \sum_{t_k^2<t} \Phi_2^{12}(t-t_k^2), \tag{C.64}$$

and

$$\mathbf{h}_1(t) = \sum_{n=1}^{\infty} \begin{pmatrix} \alpha^{11} f^{11}(t;\theta^{11}) & 0 \\ \alpha^{21} f^{21}(t;\theta^{21}) & 0 \end{pmatrix}^{\otimes n} \tag{C.65}$$

$$H_1^{ij}(t-u) = \int_0^t h_1^{ij}(\tau-u)[\![\tau>u]\!]\mathrm{d}\tau = \int_u^t h_1^{ij}(\tau-u)\mathrm{d}\tau \tag{C.66}$$

$$\Phi_2^{ij}(t) = \alpha^{ij} \int_0^t f_2^{ij}(\tau;\{\theta^{ij}\})\mathrm{d}\tau = \alpha^{ij} F_2^{ij}(t;\{\theta^{ij}\}) \tag{C.67}$$

Our goal is to show that Eq. (C.60) is convex in the $\{\alpha^{ij}\}$ parameters for fixed $\{\theta^{ij}\}$. Our strategy is to show that each of the four tems in Eq. (C.60) is a convex function of $\{\alpha^{ij}\}$.

First, note that the map $l : x \mapsto -\sum \log x$ is convex and non-decreasing. As such, for $-\log\sum_{s_k^1}(\xi_1^1(s_k^1;\{\alpha^{ij}\}))\} = l \circ \xi_1^1(\cdot;\{\alpha^{ij}\})$ to be convex in $\{\alpha^{ij}\}$ (where $\xi_1^1(\cdot;\{\alpha^{ij}\})$ is interpreted as a map from α parameters to the conditional intensity, *i.e.*, $(0,1)^4 \mapsto \mathbb{R}^+$), it is sufficient to show that $\xi_1^1(\cdot;\{\alpha^{ij}\})$ is a convex function. Similarly, for $-\log\sum_{s_k^2}(\xi_1^2(s_k^2;\{\alpha^{ij}\})\}$ to be convex, $\xi_1^2(\cdot;\{\alpha^{ij}\})$ has to be convex. For the last two terms of Eq. (C.60) to be convex, $\Xi_1^1(T;\{\alpha^{ij}\})$ and $\Xi_1^2(T;\{\alpha^{ij}\})$ have to be convex.

I. Convexity of $\xi_1^1(\cdot;\{\alpha^{ij}\})$. Let us split Eq. (C.61) as follows:

$$\xi_1^1(t) = \overbrace{a\delta(t-c) + a h_1^{11}(t-c)}^{I.A} + \overbrace{\sum_{t_k^2<t} \varphi_2^{12}(t-t_k^2)}^{I.B} + \overbrace{h_1^{11}(t) * \sum_{t_k^2<t} \varphi_2^{12}(t-t_k^2)}^{I.C} \tag{C.68}$$

To show convexity of ξ_1^1, we need to show convexity of each of the indicated terms.

For the I.A term, we compute the Hessian matrix with respect to $\{\alpha^{ij}\}$ as

$$\mathbf{Hess}_{I.A} = \left[\frac{\partial^2 (I.A)}{(\partial\alpha^{ij})^2}\right] = \begin{pmatrix} \frac{\partial^2 h_1^{11}}{(\partial\alpha^{11})^2} & \frac{\partial^2 h_1^{11}}{(\partial\alpha^{12})^2} \\ \frac{\partial^2 h_1^{11}}{(\partial\alpha^{21})^2} & \frac{\partial^2 h_1^{11}}{(\partial\alpha^{22})^2} \end{pmatrix}. \tag{C.69}$$

Let us compute each of these terms. First, observe that from Eq. (C.65) we can write

$$\mathbf{h}_1(t) = \begin{pmatrix} \sum_{n=1}^{\infty} (\alpha^{11})^n (f^{11})^{\otimes n} & 0 \\ \alpha^{21} \sum_{n=1}^{\infty} (\alpha^{11})^{n-1} f^{21} * (f^{11})^{\otimes (n-1)} & 0 \end{pmatrix}. \tag{C.70}$$

Immediately, we see that

$$\begin{aligned} \frac{\partial^2 h_1^{11}}{(\partial\alpha^{11})^2} &= \sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^{n-2}(f^{11})^{\otimes n} = \frac{1}{(\alpha^{11})^2} \sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n (f^{11})^{\otimes n} \geq 0 \\ \frac{\partial^2 h_1^{11}}{(\partial\alpha^{12})^2} &= \frac{\partial^2 h_1^{11}}{(\partial\alpha^{21})^2} = \frac{\partial^2 h_1^{11}}{(\partial\alpha^{22})^2} = 0 \end{aligned} \tag{C.71}$$

Since $\|f^{ij}\| = 1$, $\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n (f^{11})^{\otimes n}$ is upper bounded by $\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n$, which can be shown to be convergent by the Integral Test.

It follows that the set of eigenvalues of $\mathbf{Hess}_{I.A}$ is $\{0, \frac{a}{(\alpha^{11})^2} \sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n (f^{11})^{\otimes n}(t-a)\}$. Since all of the eigenvalues of $\mathbf{Hess}_{I.A}$ are non-negative, $\mathbf{Hess}_{I.A}$ is positive-semidefinite and the term I.A is convex.

For the I.B term, since $\varphi^{ij} = \alpha^{ij} f^{ij}(\cdot;\theta^{ij})$ is linear in α^{ij}, it follows that the Hessian matrix of $\varphi_2^{12}(t-t_k^2)$ is identically zero, and thus I.B is trivially convex.

For the I.C term, since h^{11} and φ^{12} are differentiable functions of α^{ij}, it follows that

$$\frac{\partial^2}{(\partial\alpha^{ij})^2}\left(h_1^{11}(t) * \sum_{t_k^2<t}\varphi_2^{12}(t-t_k^2)\right) = h_1^{11}(t) * \sum_{t_k^2<t}\frac{\partial^2}{(\partial\alpha^{ij})^2}\varphi_2^{12}(t-t_k^2) + \frac{\partial^2}{(\partial\alpha^{ij})^2}h_1^{11}(t) * \sum_{t_k^2<t}\varphi_2^{12}(t-t_k^2)$$

Since the Hessian for $\varphi_2^{12}(t-t_k^2)$ is trivially 0, the first term of I.C is trivially zero. For the second term, since $\frac{\partial^2 h_1^{11}}{(\partial\alpha^{ij})^2} = 0$ unless $i=j=1$, we only need to consider the $i=j=1$ case. Using Eq. (C.71), we have

$$\begin{aligned}\frac{\partial^2}{(\partial\alpha^{11})^2}h_1^{11}(t) * \sum_{t_k^2<t}\varphi_2^{12}(t-t_k^2) &= \frac{1}{(\alpha^{11})^2}\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n (f^{11})^{\otimes n}(t) * \sum_{t_k^2<t}\varphi_2^{12}(t-t_k^2)\\ &= \frac{1}{(\alpha^{11})^2}\sum_{t_k^2<t}\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n (f^{11})^{\otimes n}(t) * \varphi_2^{12}(t-t_k^2)\\ &= \frac{\alpha^{12}}{(\alpha^{11})^2}\sum_{t_k^2<t}\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n (f^{11})^{\otimes n}(t) * f_2^{12}(t-t_k^2),\end{aligned}$$

which is upper-bounded by $\frac{\alpha^{12}}{(\alpha^{11})^2}|\mathscr{H}_t^2|\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n$, which is convergent and positive. Thus, I.C is convex.

Since I.A, I.B, I.C are all convex, ξ_1^1 is convex in $\{\alpha^{ij}\}$

II. Convexity of $\xi_1^2(\cdot;\{\alpha^{ij}\})$. Let us split Eq. (C.62) as follows:

$$\xi_1^2(t) = \overbrace{b\delta(t-d) + bh_1^{21}(t-d)}^{II.A} + \overbrace{\sum_{t_k^2<t}\varphi_2^{22}(t-t_k^2)}^{II.B} + \overbrace{h_1^{21}(t) * \sum_{t_k^2<t}\varphi_2^{12}(t-t_k^2)}^{II.C} \tag{C.72}$$

II.B and II.C are trivially convex, similar to I.B and I.C.

For II.A, observe that

$$\mathbf{Hess}_{II.A} = \left[\frac{\partial^2(II.A)}{(\partial\alpha^{ij})^2}\right] = b\begin{pmatrix}\frac{\partial^2 h_1^{21}}{(\partial\alpha^{11})^2} & \frac{\partial^2 h_1^{21}}{(\partial\alpha^{12})^2}\\ \frac{\partial^2 h_1^{21}}{(\partial\alpha^{21})^2} & \frac{\partial^2 h_1^{21}}{(\partial\alpha^{22})^2}\end{pmatrix}. \tag{C.73}$$

From Eq. (C.70), we see that

$$\begin{aligned}\frac{\partial^2 h_1^{21}}{(\partial\alpha^{11})^2} &= \alpha^{21}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^{n-3}f^{21}*(f^{11})^{\otimes(n-1)}\\ &= \frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^{n}f^{21}*(f^{11})^{\otimes(n-1)} \geq 0\\ \frac{\partial^2 h_1^{21}}{(\partial\alpha^{12})^2} &= \frac{\partial^2 h_1^{21}}{(\partial\alpha^{21})^2} = \frac{\partial^2 h_1^{21}}{(\partial\alpha^{22})^2} = 0\end{aligned}$$

Since $\|f^{ij}\| = 1$, it can be shown that $\frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^n f^{21}*(f^{11})^{\otimes(n-1)} < \infty$ by the Integral Test.

The set of eigenvalues of $\mathbf{Hess}_{II.A}$ is $\{0, b\frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^n f^{21}*(f^{11})^{\otimes(n-1)}(t-d)\}$. Since all of the eigenvalues of $\mathbf{Hess}_{II.A}$ are non-negative, $\mathbf{Hess}_{II.A}$ is positive-semidefinite and the term II.A is convex.

It follows that ξ_1^2 is convex.

III. Convexity of $\Xi_1^1(T;\{\alpha^{ij}\})$.

$$\Xi_1^1(t) = \overbrace{a[\![t>c]\!] + aH_1^{11}(t-c)}^{III.A} + \overbrace{\sum_{t_k^2<t}\Phi_2^{12}(t-t_k^2)}^{III.B} + \overbrace{h_1^{11}(t)*\sum_{t_k^2<t}\Phi_2^{12}(t-t_k^2)}^{III.C}$$

Similar to I and II, III.B and III.C have zero Hessians as Φ_2^{ij} is linear in α^{ij}.

For III.A, see that

$$\mathbf{Hess}_{III.A} = \left[\frac{\partial^2(III.A)}{(\partial\alpha^{ij})^2}\right] = a\begin{pmatrix}\frac{\partial^2 H_1^{11}}{(\partial\alpha^{11})^2} & \frac{\partial^2 H_1^{11}}{(\partial\alpha^{12})^2}\\ \frac{\partial^2 H_1^{11}}{(\partial\alpha^{21})^2} & \frac{\partial^2 H_1^{11}}{(\partial\alpha^{22})^2}\end{pmatrix}. \tag{C.74}$$

Integrating Eq. (C.70), we see that

$$\mathbf{H}_1(t-c) = \begin{pmatrix}\sum_{n=1}^{\infty}(\alpha^{11})^n\int_c^t(f^{11})^{\otimes n}(\tau-c)\mathrm{d}\tau & 0\\ \alpha^{21}\sum_{n=1}^{\infty}(\alpha^{11})^{n-1}\int_c^t f^{21}*(f^{11})^{\otimes(n-1)}(\tau-c)\mathrm{d}\tau & 0\end{pmatrix}. \tag{C.75}$$

We then have

$$\begin{aligned}\frac{\partial^2 H_1^{11}}{(\partial\alpha^{11})^2} &= \sum_{n=2}^{\infty}n(n-1)(\alpha^{11})^{n-2}\int_c^t(f^{11})^{\otimes n}(\tau-c)\mathrm{d}\tau\\ &= \frac{1}{(\alpha^{11})^2}\sum_{n=2}^{\infty}n(n-1)(\alpha^{11})^n\int_c^t(f^{11})^{\otimes n}(\tau-c)\mathrm{d}\tau \geq 0\\ \frac{\partial^2 H_1^{11}}{(\partial\alpha^{12})^2} &= \frac{\partial^2 H_1^{11}}{(\partial\alpha^{21})^2} = \frac{\partial^2 H_1^{11}}{(\partial\alpha^{22})^2} = 0\end{aligned}$$

Note that $\frac{1}{(\alpha^{11})^2}\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n \int_c^t (f^{11})^{\otimes n}(\tau-c)\mathrm{d}\tau$ is upper-bounded (term-by-term) by $\frac{1}{(\alpha^{11})^2}\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n$, which can be shown to be convergent by the Integral Test.

Since the eigenvalues of $\mathbf{Hess}_{III.A}$, $\{0, \frac{a}{(\alpha^{11})^2}\sum_{n=2}^{\infty} n(n-1)(\alpha^{11})^n \int_c^t (f^{11})^{\otimes n}(\tau-c)\mathrm{d}\tau\}$, are nonnegative, III.A is convex, and consequently Ξ_1^1 is convex.

IV. Convexity of $\Xi_1^2(T;\{\alpha^{ij}\})$.

$$\Xi_1^2(t) = \overbrace{b[\![t>d]\!] + bH_1^{21}(t-d)}^{IV.A} + \overbrace{\sum_{t_k^2<t} \Phi_2^{22}(t-t_k^2)}^{IV.B} + \overbrace{h_1^{21}(t) * \sum_{t_k^2<t} \Phi_2^{12}(t-t_k^2)}^{IV.C}$$

Similar to III, IV.B and IV.C have zero Hessians as α^{ij} is linear in Φ_2^{ij}.

For IV.A, see that

$$\mathbf{Hess}_{IV.A} = \left[\frac{\partial^2 (IV.A)}{(\partial\alpha^{ij})^2}\right] = a\begin{pmatrix} \frac{\partial^2 H_1^{21}}{(\partial\alpha^{11})^2} & \frac{\partial^2 H_1^{21}}{(\partial\alpha^{12})^2} \\ \frac{\partial^2 H_1^{21}}{(\partial\alpha^{21})^2} & \frac{\partial^2 H_1^{21}}{(\partial\alpha^{22})^2} \end{pmatrix}. \tag{C.76}$$

From Eq. (C.75), we get

$$\begin{aligned}
\frac{\partial^2 H_1^{21}}{(\partial\alpha^{11})^2} &= \alpha^{21}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^{n-3}\int_d^t f^{21} * (f^{11})^{\otimes(n-1)}(\tau-d)\mathrm{d}\tau \\
&= \frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^{n}\int_d^t f^{21} * (f^{11})^{\otimes(n-1)}(\tau-d)\mathrm{d}\tau \geq 0 \\
\frac{\partial^2 H_1^{21}}{(\partial\alpha^{12})^2} &= \frac{\partial^2 H_1^{21}}{(\partial\alpha^{21})^2} = \frac{\partial^2 H_1^{21}}{(\partial\alpha^{22})^2} = 0
\end{aligned}$$

Note that $\frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^n \int_d^t f^{21} * (f^{11})^{\otimes(n-1)}(\tau-d)\mathrm{d}\tau$ is upper-bounded by $\frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^n$, which is convergent by the Integral Test.

From this we see that the eigenvalues of $\mathbf{Hess}_{IV.A}$ are $\{0, b\frac{\alpha^{21}}{(\alpha^{11})^3}\sum_{n=3}^{\infty}(n-1)(n-2)(\alpha^{11})^n \int_d^t f^{21} * (f^{11})^{\otimes(n-1)}(\tau-d)\mathrm{d}\tau\}$. As these are nonnegative, we have that $\mathbf{Hess}_{IV.A}$ is positive-semidefinite, and so IV.A is convex. Consequently, Ξ_1^2 is convex.

As $\xi_1^1, \xi_1^2, \Xi_1^1, \Xi_1^2$ are all convex in $\{\alpha^{ij}\}$, it follows that the PP-PP NLL for the given exogenous function $\boldsymbol{\mu}(t)$ is convex in $\{\alpha^{ij}\}$.

Now, suppose that we have a multi-impulse exogenous input, given by

$$\boldsymbol{\mu}(t) = \sum_i \begin{pmatrix} a_i \cdot \delta(t-c_i) \\ b_i \cdot \delta(t-d_i) \end{pmatrix}.$$

Since the multi-impulse exogenous input is a sum of single exogenous input functions, the effect of each impulse is simply additive on the conditional intensity ξ_1^1, ξ_1^2 and the compensators Ξ_1^1, Ξ_1^2. Since a nonnegative-weighted sum of convex functions is still convex, it follows that the PP-PP NLL for the multi-impulse exogenous input is also convex in $\{\alpha^{ij}\}$ for fixed $\{\theta^{ij}\}$.

C.7 Approximating the Conditional Intensity $\boldsymbol{\xi}_E(t)$

As discussed in Section 4.4.1, in general, a closed-form solution for the conditional intensity $\boldsymbol{\xi}_E(t)$ in Eq. (4.5) is not guaranteed to exist since the infinite sum of convolutions $\mathbf{h}_E(t) = \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(t)$ cannot be expressed in closed form. In order to calculate $\boldsymbol{\xi}_E(t)$ given a sequence of observations and sample the PCMHP(d,e) process, we require the development of numerical techniques to approximate $\boldsymbol{\xi}_E(t)$. In this section, we describe two approximations that enable calculation of $\boldsymbol{\xi}_E(t)$.

Given the set of observed histories $\mathscr{H}_{t^-}^{E^c}$ on $[0,t)$, suppose that we are interested in approximating the intensity $\boldsymbol{\xi}_E(t)$. Eq. (4.5) can be used for this task, however, there are two issues that we first need to address. First, the formula involves taking the convolution of functions. If the functions involved are not too complex, the convolution can be calculated in closed form. In general, however, this is not the case and we need to approximate it numerically. Second, Eq. (4.5) contains the infinite sum of convolutions $\mathbf{h}_E(t)$, which in most cases cannot be written in closed form and has to be approximated as well.

To address these two issues, we introduce two approximations:

- Approximating continuous convolution with numerical convolution;
- And, approximating the infinite series $\mathbf{h}_E(t)$ with the sum of the first k terms $\mathbf{h}_E^k(t)$.

C.7.1 Numerical Convolution

Let $\mathbf{f}$ and $\mathbf{g}$ be matrix functions defined over $[0,t]$ such that the number of columns of $\mathbf{f}(s)$ and the number of rows $\mathbf{g}(s)$ are equal, *i.e.*, the matrix product $\mathbf{f}(s)\cdot\mathbf{g}(s)$ can be calculated. Assume that we are given a partition $\mathscr{P}[0,t]$ of $[0,t]$ with constant increment $\Delta^{\mathscr{P}}$, such that $\mathscr{P}[0,t] = \{t_0 = 0 < t_1 = \Delta^{\mathscr{P}} < t_2 = 2\Delta^{\mathscr{P}} < \ldots < (P-1)\Delta^{\mathscr{P}} = t_{P-1} < t = t_P\}$, where $P = \lceil t/\Delta^{\mathscr{P}} \rceil$. Let $\mathbf{f}[0:t]$ be the numerical array obtained by sampling the function $\mathbf{f}(t)$ on each point of $\mathscr{P}[0,t]$, *i.e.*, $[\mathbf{f}(0),\mathbf{f}(1),\ldots,\mathbf{f}(t)]$. The array $\mathbf{g}[0:t]$ is defined similarly.

We introduce the conv operator, a discrete approximation to continuous function convolution, where the convolution $\mathbf{f} * \mathbf{g}$ on $[0,t]$ is approximated as a sum of convolution terms over the partition $\mathscr{P}[0,t]$. Within each subinterval, we fix $\mathbf{g}$ at the left endpoint and perform the integration on $\mathbf{f}$. The univariate version of this convolution approximation scheme was considered by [102].

Proposition C.5. *Given a partition $\mathscr{P}[0,t]$ of $[0,t]$ and functions $\mathbf{f}$ and $\mathbf{g}$, $(\mathbf{f}*\mathbf{g})(t)$ is approximated by*

$$\text{conv}(\mathbf{f},\mathbf{g},\mathscr{P}[0,t]) = \sum_{t_i\in\mathscr{P}[0,t]} [\mathbf{F}(t_i) - \mathbf{F}(t-\min(t_{i+1},t))]\cdot\mathbf{g}(t_i), \tag{C.77}$$

where $\cdot$ is matrix multiplication and

$$\mathbf{F}(t) = \int_0^t \mathbf{f}(\tau)\mathrm{d}\tau.$$

Proof.

$$\begin{aligned}
\int_0^t \mathbf{f}(t-\tau)\cdot\mathbf{g}(\tau)\mathrm{d}\tau &= \sum_{t_i\in\mathscr{P}[0,t]} \int_{t_i}^{\min(t_{i+1},t)} \mathbf{f}(t-\tau)\cdot\mathbf{g}(\tau)\mathrm{d}\tau \\
&\approx \sum_{t_i\in\mathscr{P}[0,t]} \left[\int_{t_i}^{\min(t_{i+1},t)} \mathbf{f}(t-\tau)\mathrm{d}\tau\right]\cdot\mathbf{g}(t_i) \\
&= \sum_{t_i\in\mathscr{P}[0,t]} [\mathbf{F}(t_i) - \mathbf{F}(t-\min(t_{i+1},t))]\cdot\mathbf{g}(t_i).
\end{aligned}$$

■

To obtain an approximation for $\boldsymbol{\varphi}_E^{\otimes n}(t)$ over $\mathscr{P}[0,t]$ for $n\geq 2$, Proposition C.5 can be applied n times to $\boldsymbol{\varphi}_E(t)$. Summing the resulting expressions and applying the infinite series truncation in Appendix C.7.2 allow us to obtain the approximation $\mathbf{h}_E[0:t]$.

Given $\mathbf{h}_E[0:t]$, we see in Eq. (4.5) that calculating $\boldsymbol{\xi}_E[0:t]$ involves a second set of convolutions, where we pair $\mathbf{h}_E[0:t]$ with the background intensity $\boldsymbol{\mu}[0:t]$ and the influence contributions of the events in the E^c dimensions. Applying Proposition C.5 to these convolutions requires the integrated $\mathbf{h}_E(t)$, denoted as $\mathbf{H}_E(t)$. Proposition C.6 writes this as a function of the integrated Hawkes kernel $\boldsymbol{\Phi}_E(t)$ and $\mathbf{h}_E(t)$.

Proposition C.6.

$$\mathbf{H}_E(t) = \boldsymbol{\Phi}_E(t) + \mathbf{h}_E(t) * \boldsymbol{\Phi}_E(t) \tag{C.78}$$

Proof.

$$\begin{aligned}
\mathbf{H}_E(t) &= \int_0^t \mathbf{h}_E(s)\mathrm{d}s \\
&= \int_0^t \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(s)\mathrm{d}s \\
&= \sum_{n=1}^{\infty} \int_0^t \boldsymbol{\varphi}_E^{\otimes n}(s)\mathrm{d}s \\
&= \int_0^t \boldsymbol{\varphi}_E(s)\mathrm{d}s + \sum_{n=2}^{\infty} \int_0^t \boldsymbol{\varphi}_E^{\otimes n}(s)\mathrm{d}s \\
&= \boldsymbol{\Phi}_E(t) + \sum_{n=2}^{\infty} \int_0^t (\boldsymbol{\varphi}_E * \boldsymbol{\varphi}_E^{\otimes n-1})(s)\mathrm{d}s \\
&= \boldsymbol{\Phi}_E(t) + \int_0^t \sum_{n=2}^{\infty} (\boldsymbol{\varphi}_E^{\otimes n-1} * \boldsymbol{\varphi}_E)(s)\mathrm{d}s \\
&= \boldsymbol{\Phi}_E(t) + \int_0^t \sum_{n=1}^{\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \boldsymbol{\varphi}_E)(s)\mathrm{d}s \\
&= \boldsymbol{\Phi}_E(t) + \int_0^t (\mathbf{h}_E * \boldsymbol{\varphi}_E)(s)\mathrm{d}s \\
&= \boldsymbol{\Phi}_E(t) + \mathbf{H}_E(t) * \boldsymbol{\Phi}_E(t),
\end{aligned}$$

where the last line follows from the Fubini-Tonelli Theorem and that $\mathbf{h}_E$ and $\boldsymbol{\varphi}_E$ are $\mathscr{L}^1$-integrable. ■

C.7.2 Infinite Series Truncation

The infinite series $\mathbf{h}_E(t) = \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(t)$ is approximated by truncating the series up to the k^{*th} term, where k^* is selected based on a convergence threshold we discuss below, and replacing the continuous convolution $*$ with the numerical convolution operator conv. If we set

$$\mathrm{conv}^n(\boldsymbol{\varphi}_E, \mathscr{P}[0,t]) = \begin{cases} \boldsymbol{\varphi}_E[0:t] & n = 1, \\ \mathrm{conv}(\mathrm{conv}^{n-1}(\boldsymbol{\varphi}_E, \mathscr{P}[0,t]), \boldsymbol{\varphi}_E, \mathscr{P}[0,t]) & n > 1, \end{cases}$$

then our approximation to $\mathbf{h}_E(t)$ is given by

$$\mathbf{h}_E(t) \approx \sum_{n=1}^{k^*} \mathrm{conv}^n(\boldsymbol{\varphi}_E, \mathscr{P}[0,t]). \tag{C.79}$$

The accuracy of this approximation can be specified by a threshold on the max norm.

Definition C.3. Given a real matrix $\mathbf{M} = (m^{ij})$, we define its max norm $\|\mathbf{M}\|_{\max}$ as

$$\|\mathbf{M}\|_{\max} = \max_{i,j} |m^{ij}|. \tag{C.80}$$

Our approximation of the function $\mathbf{h}_E$ over $\mathscr{P}[0,t]$ is the set of points $\mathbf{h}_E^{k^\star}[0:t]$, where $k^\star$ is chosen to be the smallest $k \geq 1$ such that

$$\max_{s \in \mathscr{P}[0,t]} \left\| \text{conv}^k(\boldsymbol{\varphi}_E, \mathscr{P}[0,s]) \right\|_{\max} < \gamma^h, \tag{C.81}$$

where $\gamma^h > 0$ is a predetermined convergence threshold.

C.7.3 Algorithm to Approximate $\boldsymbol{\xi}_E(t)$

Algorithm 2 combines the approximation techniques in Appendices C.7.1 and C.7.2 to compute $\boldsymbol{\xi}_E[0:T]$ for a pretermined maximum time $T > 0$ and observed data $\mathscr{H}_{T^-}^{E^c}$. This algorithm involves three steps: (1) calculating the infinite sum approximation $\mathbf{h}_E^{k^\star}[0:T]$; (2) iterating over the events in E^c and getting the running total intensity contributed by these events; and (3) calculating $\boldsymbol{\xi}_E[0:T]$ using Eq. (4.5).

Algorithm 2: Approximation of $\boldsymbol{\xi}_E(t)$ by Discrete Convolution and Infinite Series Truncation

Input: kernel $\boldsymbol{\varphi}(t)$, kernel integral $\boldsymbol{\Phi}(t)$ exogenous input function $\boldsymbol{\mu}(t)$, partition $\mathscr{P}[0,T] = [0:T]$ with increment $\Delta^{\mathscr{P}} > 0$, observed E^c data points in $[0,T)$ $\{\mathscr{H}_T^j = \{t_k^j\} \mid j \in E^c, t_k^j < T\}$, threshold $\gamma^h > 0$
Output: $\boldsymbol{\xi}_E[0:T]$
initialize $P = \frac{T}{\Delta^{\mathscr{P}}}$; $\boldsymbol{\xi}_E[0:T] = \mathrm{a}[0:T] = \mathbf{0}^{d\times(P+1)}$; $\mathbf{h}_E[0:T] = \boldsymbol{\Delta}[0:T] = \boldsymbol{\varphi}_E[0:T]$;
do
  $\boldsymbol{\Delta}[0:T] = \text{conv}(\Delta, \boldsymbol{\varphi}_E, [0:T])$;
  $\mathbf{h}_E[0:T] = \mathbf{h}_E[0:T] + \boldsymbol{\Delta}[0:T]$;
while $\max_{p=0:P}\left(\left\|\boldsymbol{\Delta}[t_p]\right\|_{\max}\right) \geq \gamma^h$;
$\mathbf{H}_E[0:T] = \boldsymbol{\Phi}[0:T] + \text{conv}(\mathbf{h}_E, \boldsymbol{\Phi}, [0:T])$;
for $j \in E^c$ **do**
  for $t_k^j \in \mathscr{H}_t^j$ **do**
    for $t_p \in [0:T]$ **do**
      if $t_k^j < t_p$ **then**
        $\mathrm{a}[t_p] = \mathrm{a}[t_p] + \boldsymbol{\varphi}_{E^c}^j(t_p - t_k^j)$;
      end
    end
  end
end
$\boldsymbol{\xi}_E[0:T] = \boldsymbol{\mu}[0:T] + \mathrm{a}[0:T] + \text{conv}(\mathbf{h}_E, \boldsymbol{\mu} + \mathrm{a}, [0:T])$;
return $\boldsymbol{\xi}_E[0:T]$

The two hyperparameters $\Delta^{\mathscr{P}}$ and γ^h control the approximation error involved in calculating $\boldsymbol{\xi}_E(t)$. The higher $\Delta^{\mathscr{P}}$ and the lower γ^h, the better the approximation. This, of

course, comes with the tradeoff of a longer computation time. Note that $\Delta^{\mathscr{P}}$ depends on the timescale of the process considered. A simple heuristic for $\Delta^{\mathscr{P}}$ is setting it with consideration of the interevent distribution of the point process history $\mathscr{H}_T^{E^c}$, i.e. $\{t_k - t_{k-1}\}$. One can start by setting $\Delta^{\mathscr{P}}$ as M =median($\{t_k - t_{k-1}\}$), then setting it as $\frac{1}{2}M$, $\frac{1}{3}M$, and so on, and calculating the relative difference of $\boldsymbol{\xi}_E(t)$ for progressively smaller $\Delta^{\mathscr{P}}$. One then chooses $\Delta^{\mathscr{P}}$ with a relative error lower than a predefined error threshold.

An alternative way to calculate $\boldsymbol{\xi}_E(t)$ is to interpret the PCMHP(d,e) intensity in Definition 4.1 as an expectation of the multivariate Hawkes intensity, with respect to the events in the E dimensions conditioned on the events in the E^c dimensions. To compute this expectation, we simply sample Hawkes event histories over the E dimensions using the Hawkes thinning algorithm, calculating the Hawkes intensity given each sample, and then average the resulting intensities. The method is presented in Appendix C.8.

We have presented three approaches to compute $\boldsymbol{\xi}_E(t)$: (1) a closed-form solution for the PCMHP$(2,1)$ process with exponential kernel derived in Appendix C.3; (2) the approach developed in this section based on numerical convolution and infinite series truncation; and (3) a sample-based approach developed in Appendix C.8. A comparison of $\boldsymbol{\xi}_E(t)$ obtained from these three approaches for the PCMHP$(2,1)$ process is presented in Appendix C.9.

C.8 $\boldsymbol{\xi}_E(t)$ as a Conditional Expectation over MHP Samples

With our definition of the PCMHP(d,e) conditional intensity $\boldsymbol{\xi}_E(t)$ as the expectation of the conditional intensity of a d-dimensional Hawkes process given observed event sequences $\mathscr{H}_{T^-}^{E^c}$, a straightforward way to calculate $\boldsymbol{\xi}_E(t)$ is as follows:

1. Use Algorithm 3 to obtain $n_{samples}$ samples of $\mathscr{H}_{T^-}^{E}$.
2. Given the n^{th} sample $\{\mathscr{H}_{T,n}^{j} = \{t_{k,n}^{j}\} | j \in E\}$, calculate the Hawkes conditional intensity:
$$\boldsymbol{\lambda}_n(t) = \boldsymbol{\mu} + \sum_{j\in E^c} \sum_{t_{k,n}^j < t} \boldsymbol{\varphi}^{\boldsymbol{j}}(t - t_k^j) + \sum_{j\in E} \sum_{t_{k,n}^j < t} \boldsymbol{\varphi}^{\boldsymbol{j}}(t - t_k^j) \tag{C.82}$$
3. Calculate the average to get the MBP conditional intensity:
$$\boldsymbol{\xi}_E(t) = \frac{1}{n_{samples}} \sum_{n=1}^{n_{samples}} \boldsymbol{\lambda}_n(t) \tag{C.83}$$

The approximation error of this method depends on the number of samples we take $n_{samples}$. The higher $n_{samples}$ we take, the lower the standard error of our average Eq. (C.83).

Algorithm 3 is a modification of Ogata's thinning algorithm in Algorithm 1, where we consider the intensity contributed by the event sequences in $\mathscr{H}_{T^-}^{E^c}$ as part of the exogenous intensity, which in the usual case is just the constant $\boldsymbol{\mu}$.

Since the exogenous intensity is nonconstant due to this modification, Algorithm 1 has to be adjusted since it assumes a constant exogenous excitation. Specifically, we need to adjust the upper bound

$$\bar{\lambda}(t) = \sum_{i=1}^{d} \lambda^i(t^+) \tag{C.84}$$

so that its dominance holds until the next event after t.

The i^{th} component of the conditional intensity $\lambda(t)$ given event sequences $\mathscr{H}_{T^-}^{E^c}$ can be written as

$$\lambda^i(t) = \mu^i + \sum_{j\in E^c} \sum_{t_k^j < t} \varphi^{ij}(t - t_k^j) + \sum_{j\in E} \sum_{t_k^j < t} \varphi^{ij}(t - t_k^j). \tag{C.85}$$

Given that every component in $\boldsymbol{\varphi}$ is non-decreasing, a natural upper bound on $\varphi^{ij}(t - t_k^j)$ is $\varphi^{ij}(0)$. Thus we can write

$$\begin{aligned}\lambda^i(t) &\le \mu^i + \sum_{j\in E^c} \varphi^{ij}(0) \sum_{t_k^j < t} 1 + \sum_{j\in E} \sum_{t_k^j < t} \varphi^{ij}(t - t_k^j) \\ &= \mu^i + \sum_{j\in E^c} \varphi^{ij}(0) |\mathscr{H}_t^j| + \sum_{j\in E} \sum_{t_k^j < t} \varphi^{ij}(t - t_k^j)\end{aligned}$$

The upper bound $\bar{\lambda}(t)$ has to hold until the stochastic time of next event. The only non-stochastic upper bound for $|\mathcal{H}_t^j|$ until the next event is $|\mathcal{H}_T^j|$. Setting this upper bound we get

$$\bar{\lambda}(t) = \sum_{i=1}^{d} \left[\mu^i + \sum_{j \in E^c} |\mathcal{H}_T^j| \varphi^{ij}(0) + \sum_{j \in E} \sum_{t_k^j < t} \varphi^{ij}(t - t_k^j) \right], \tag{C.86}$$

which is now a correct upper bound for this setup.

Algorithm 3: Simulating an e-dimensional Hawkes Process on $[0, T)$ Given Observed Event Sequences $\bigcup_{j \in E^c} \mathcal{H}_T^j$ by Thinning

Input: kernel matrix $\boldsymbol{\varphi}(t)$, exogenous excitation $\boldsymbol{\mu}$; observed event sequences $\mathcal{H}_T^j = \{t_k^j\}$ for $j \in E^c$; time horizon $T > 0$

Output: $\mathcal{H}_T^j = \{t_k^j\}$ for $j \in E$

initialize $t = 0; \mathcal{H}_T^j = \emptyset$ for $j \in E$;

while $t < T$ **do**

- $\bar{\lambda} = \sum_{i \in E} \left[\mu^i + \sum_{j \in E^c} |\mathcal{H}_T^j| \varphi^{ij}(0) + \sum_{j \in E} \sum_{t_k^j \le t} \varphi^{ij}(t - t_k^j) \right]$;
- $u \sim$ `uniform(0,1)`;
- $w = -\log \frac{u}{\bar{\lambda}}$;
- $t = t + w$;
- $U \sim$ `uniform(0,1)`;
- **if** $U\bar{\lambda} \le \sum_{i \in E} \lambda^m(t)$ **then**
  - $j = 1$;
  - **while** $U\bar{\lambda} \le \sum_{i=1}^{j} \lambda^m(t)$ **do**
    - $j = j + 1$;
  - **end**
  - $t_k^j = t$;
  - $\mathcal{H}_T^j = \mathcal{H}_T^j \cup \{t_k^j\}$;
- **end**

end

if $t_k^j < T$ **then**

- **return** $\mathcal{H}_T^j$ for $j \in E$;

else

- **return** $\mathcal{H}_T^1, \ldots, \mathcal{H}_T^j \setminus \{t_k^j\}, \ldots, \mathcal{H}_T^e$;

end

C.9 Comparison of $\boldsymbol{\xi}_E(t)$ Evaluation Methods

As noted in Section 4.4.2, a closed form solution for $\boldsymbol{\xi}_E(t)$ of a generic PCMHP(d,e) process cannot be written down except in special cases. One of these special cases as we have shown in Appendix C.3 is the PCMHP$(2,1)$ process with the exponential kernel. Here we present a comparison of the two approximation schemes, developed in Appendix C.7 and Appendix C.8, with the closed form solution for the exponential PCMHP$(2,1)$ process.

Fig. C.2 shows a comparison of $\xi_1^1(t)$ and $\xi_1^2(t)$ computed through the numerical convolution approximation, as an expectation over Hawkes processes, and the closed form solution in the Appendix. We see that there is agreement in all cases, showing the viability of our two approximation methods.

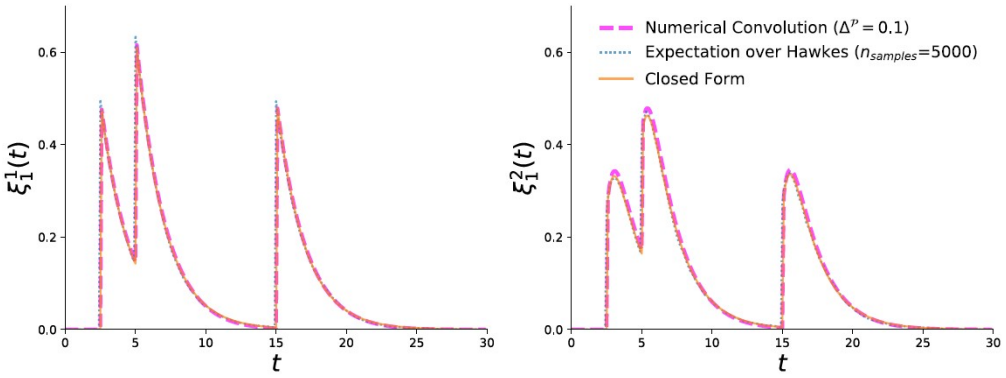


Figure C.2: Comparison of the exponential PCMHP$(2,1)$ conditional intensity obtained three ways: (1) the method based on numerical convolution in Appendix C.7, (2) the expectation-over-Hawkes method presented in this section, and (3) the closed-form solution in Appendix C.3. Parameter set: $\boldsymbol{\theta} = [1,1,0.2,0.5]$, $\boldsymbol{\alpha} = [0.5,0.5,0.5,0.5]$. Event sequence in dimension 2: $\mathscr{H}_{30}^2 = \{2.5,5,15\}$.

C.10 Numerical Scheme to Calculate PCMHP Likelihood

For this section, we assume that $Q = E$ (see Section 4.4.2). Algorithm 2 can be used to calculate the negative log-likelihood of a PCMHP(d, e) process given interval-censored counts in the E dimensions and event sequences in the E^c dimensions. However, it only returns the conditional intensity $\boldsymbol{\xi}_E(t)$ on the points on the partition $\mathscr{P}[0, T]$ and does not return the compensator $\boldsymbol{\Xi}_E(t)$. These two observations are problematic if we want to compute the negative log-likelihood $\mathscr{L}(\boldsymbol{\Theta}; T)$ because (1) we need the compensator $\boldsymbol{\Xi}_E(t)$ to compute $\mathscr{L}(\boldsymbol{\Theta}; T)$, and (2) we would have to evaluate $\boldsymbol{\xi}_E(t)$ for every timestamp in $\bigcup_{j \in E^c} \mathscr{H}_t^j$ and the compensator on the left and right endpoints of the observation intervals. Points where we need to evaluate these functions may not coincide with the points on $\mathscr{P}[0, T]$. Though we can interpolate our points of interest (*i.e.*, the event timestamps and the censor points) on the partition, it is prone to error. A more accurate approach would be to calculate the $\boldsymbol{\xi}_E(t)$ and $\boldsymbol{\Xi}_E(t)$ directly at the points of interest.

In this section, we introduce Algorithm 4, which allows us to compute $\mathscr{L}(\boldsymbol{\Theta}; T)$ given a set of event sequences and observed counts.

Let $\mathscr{T} = \bigcup_{j \in E^c} \{T_p^j\}$ be the collection of all observed event timestamps in the E^c dimensions. Let $\mathscr{O} = \bigcup_{j \in E} \{o_p^j\} = \bigcup_{j \in E} \mathscr{O}^j$ be the collection of all censor points in the E dimensions. Let $\mathscr{P} = \{0 = t_0 < \cdots < t_P = T\}$ be a partition of the time interval $[0, T]$ with step size $\frac{P}{T}$. Define the points-of-interest set $\tilde{\mathscr{T}}$ as the union of these three sets supplied with labels $\mathscr{C}_i$. Let

$$\tilde{\mathscr{T}} = \{(t_i, \mathscr{C}_i)\} = \mathscr{T} \cup \mathscr{O} \cup \mathscr{P}, \tag{C.87}$$

where t_i is the i^{th} point-of-interest, $\mathscr{C}_i = \{(r_j, d_j)\}$ is a set of ordered pairs containing point t_i's roles $\{r_j\}$ and corresponding dimensions of interest $\{d_j\}$. Here, $r_j \in \{ts, o, p\}$, where ts, o and p represent event timestamp, observation censor point and partition point, respectively. In the case that $r_j = p$, d_j is unspecified. Otherwise, $d_j \in D$.

We iterate over $\tilde{\mathscr{T}}$ in Algorithm 4, storing the contribution of each point-of-interest to the interval-censored log-likelihood and point-process log-likelihood defined in Eqs. (4.11) and (4.12), respectively. The overall log-likelihood $\mathscr{L}(\boldsymbol{\Theta}; T)$ is then the sum of these two.

Runtime Complexity. To get an estimate of the runtime complexity, observe that Algorithm 4 can be decomposed into three major steps.

1. pre-compute $\mathbf{h}_E[t_0 : t_P]$ and $\mathbf{H}_E[t_0 : t_P]$,
2. iterate over $k \in \tilde{\mathscr{T}}$, computing $\mathrm{a}[t_k]$ and $\mathrm{A}[t_k]$, $\boldsymbol{\xi}_E$ if t_k is an E^c-timestamp (to calculate the PP-LL contribution), and $\boldsymbol{\Xi}_E$ if t_k is an E-censor point, and

Algorithm 4: Computing the Negative Log Likelihood of a Partial MBP Process

Input: kernel matrix $\boldsymbol{\varphi}(t)$, kernel integral matrix $\boldsymbol{\Phi}(t)$, exogenous input function $\boldsymbol{\mu}(t)$, exogenous input integral $\mathbf{S}(t)$, points-of-interest set $\tilde{\mathcal{T}} = \{(t_i, (r_j, d_j))\} = \mathcal{T} \cup \left(\bigcup_{j \in E} \mathcal{O}^j\right) \cup \mathscr{P}$, censored counts $\{\mathsf{C}_k^j\}$ for $j \in E$, threshold $\gamma^h > 0$
Output: NLL, the negative log likelihood of the partial MBP process
initialize $\mathbf{h}_E[t_0 : t_P] = \mathbf{0}^{d \times d \times (P+1)}$; $\boldsymbol{\Delta}[t_0 : t_P] = \boldsymbol{\varphi}_E[t_0 : t_P]$; $\mathrm{a}[t_0 : t_{|\tilde{\mathcal{T}}|}] = \mathrm{A}[t_0 : t_{|\tilde{\mathcal{T}}|}] = \mathbf{0}^{d \times (|\tilde{\mathcal{T}}|+1)}$; $\tilde{\mathbf{H}}_E[t_0 : t_{|\tilde{\mathcal{T}}|+1}] = \mathbf{0}^{d \times d \times (|\tilde{\mathcal{T}}|+1)}$; $l^j = 0, \boldsymbol{\omega}^j = \mathbf{0}^{|\mathcal{O}^j|} \forall\, j \in E$; PPLL = ICLL = 0;
do
  $\boldsymbol{\Delta}[t_0 : t_P] = \mathrm{conv}(\Delta, \boldsymbol{\varphi}_E, [t_0 : t_P])$;
  $\mathbf{h}_E[t_0 : t_P] = \mathbf{h}_E[t_0 : t_P] + \boldsymbol{\Delta}[t_0 : t_P]$;
while $\max_{p=0:P} \left(\left\|\boldsymbol{\Delta}[t_p]\right\|_{\max}\right) \geq \gamma^h$;
$\mathbf{H}_E[t_0 : t_P] = \boldsymbol{\Phi}[t_0 : t_P] + \mathrm{conv}(\mathbf{h}_E, \boldsymbol{\Phi}, [t_0 : t_P])$;
for $k = 0 : |\tilde{\mathcal{T}}|$ **do**
  for $j \in E^c$ **do**
    for $t_l \in \mathscr{H}_{t_k}^j$ **do**
      $\mathrm{a}[t_k] = \mathrm{a}[t_k] + \boldsymbol{\varphi}_{E^c}^j(t_k - t_l)$;
      $\mathrm{A}[t_k] = \mathrm{A}[t_k] + \boldsymbol{\Phi}_{E^c}^j(t_k - t_l)$;
    end
  end
  for $r_j, d_j \in \mathscr{C}_k$ **do**
    if $r_j = ts$ **then**
      $\boldsymbol{\xi}_E = \boldsymbol{\mu}(t_k) + \mathrm{a}[t_k]$;
      if $|E| \neq 0$ *and* $k \neq 0$ **then**
        $\boldsymbol{\xi}_E = \boldsymbol{\xi}_E + \mathrm{conv}(\mathbf{h}_E, \boldsymbol{\mu} + \mathbf{a}, [t_0 : t_k])$;
      end
      $\mathrm{PPLL} = \mathrm{PPLL} + \log \xi_E^{d_j}$
    end
    if $r_j = o$ **then**
      $\boldsymbol{\Xi}_E = \mathbf{S}(t_k) + \mathrm{A}[t_k]$;
      if $|E| \neq 0$ *and* $k \neq 0$ **then**
        $\boldsymbol{\Xi}_E = \boldsymbol{\Xi}_E + \mathrm{conv}(\mathbf{h}_E, \mathbf{S} + \mathbf{A}, [t_0 : t_k])$;
      end
      $\omega^j[l] = \Xi_E^{d_j}$;
      $l^j = l^j + 1$;
    end
  end
  if $k = |\tilde{\mathcal{T}}|$ **then**
    $\boldsymbol{\Xi}_E = \boldsymbol{\Xi}_E + \mathrm{conv}(\mathbf{h}_E, \mathbf{S} + \mathbf{A}, [t_0 : t_k])$;
    for $j \in E^c$ **do**
      $\mathrm{PPLL} = \mathrm{PPLL} - \Xi_E^j$;
    end
  end
end
for $j \in E$ **do**
  $\mathrm{diff}[1 : |\mathcal{O}^j|] = \omega^j[1 : |\mathcal{O}^j|] - \omega^j[0 : |\mathcal{O}^j| - 1]$;
  $\mathrm{ICLL} = \mathrm{ICLL} + \sum_{k=1}^{|\mathcal{O}^j|} (\mathsf{C}_k^j \cdot \log \mathrm{diff}[k] - \mathrm{diff}[k])$;
end
$\mathrm{NLL} = -(\mathrm{PPLL} + \mathrm{ICLL})$;
return NLL

3. iterate over E to calculate the total IC-LL.

To get the runtime complexity of the first step, we first observe that $\mathbf{h}_E[t_0 : t_P]$ is composed of an infinite sum of self-convolutions of $\boldsymbol{\varphi}_E$ over the partition $\mathscr{P}$ (see Eq. (C.79)). Given the partition length Δ^P, there are $\lceil \frac{T}{\Delta^P} \rceil$ partition points. For each partition point, we

perform a matrix multiplication of $\mathcal{O}(d \cdot e)$ entries, since $\boldsymbol{\varphi}_E$ is a $d \times d$ matrix with e nonzero columns. Thus a single convolution scales as $\mathcal{O}(\lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e)$. Consequently, $\mathbf{h}_E[t_0 : t_P]$ is obtained by recursively applying the convolution operation and summing up the k^* terms (see Eq. (C.81)), yielding $\mathcal{O}(k^\star \cdot \lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e)$ operations.

For the second step, we assume that $\tilde{\mathcal{T}} = \{\mathcal{T}, \mathcal{O}, \mathcal{P}\}$ are pairwise disjoint, so that we can write $|\tilde{\mathcal{T}}| = |\mathcal{T}| + |\mathcal{O}| + |\mathcal{P}| = \mathrm{n}^E + \mathrm{n}^{E^c} + \lceil \frac{T}{\Delta^P} \rceil$. For each point t_k in $\tilde{\mathcal{T}}$, we pre-compute $\mathrm{a}[t_k]$ and $\mathrm{A}[t_k]$, which takes $\mathcal{O}(\mathrm{n}^{E^c} \cdot d)$ operations (since we loop over all Hawkes datapoints in E^c and we have to compute a and A values in each dimension. Next, we compute $\boldsymbol{\xi}_E$ only for points $t_k \in \mathcal{T}$ (since we need them for PP-LL calculations). Each point $t_k \in \mathcal{T}$ (of which there are $\frac{\mathrm{n}^{E^c}}{\mathrm{n}^E + \mathrm{n}^{E^c} + \lceil \frac{T}{\Delta^P} \rceil}$) requires $\mathcal{O}(d + \lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e)$ operations (evaluation of exogenous intensity and the convolution operation). Similarly, each point $t_k \in \mathcal{O}$ (of which there are $\frac{\mathrm{n}^E}{\mathrm{n}^E + \mathrm{n}^{E^c} + \lceil \frac{T}{\Delta^P} \rceil}$) requires $\mathcal{O}(d + \lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e)$ operations. Thus, the second step has runtime complexity $\mathcal{O}\Big(|\tilde{\mathcal{T}}| \cdot \big[\mathrm{n}^{E^c} \cdot d + \frac{\mathrm{n}^E + \mathrm{n}^{E^c}}{|\tilde{\mathcal{T}}|} \cdot \lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e\big]\Big)$. The first term, being quadratic in the number of events, dominates over the second, being linear, hence we can simplify the complexity as $\mathcal{O}\Big(|\tilde{\mathcal{T}}| \cdot \mathrm{n}^{E^c} \cdot d\Big)$.

The third step has complexity $\mathcal{O}(\mathrm{n}^E)$.

Consider two cases.

(a) If $E^c = \emptyset$, then $\mathrm{n}^{E^c} = 0$, simplifying the second step's complexity to $\mathcal{O}\Big(\mathrm{n}^E \cdot \lceil \frac{T}{\Delta^P} \rceil \cdot d \cdot e\big]\Big)$. The total runtime complexity is then linear in n^E, similar to that of the MBP (Poisson) process.

(b) If $E^c \neq \emptyset$, we observe that the first term of the second step's complexity is quadratic in n^{E^c}, similar to the MHP. In practice (particularly for high-frequency data), the number of observed Hawkes events n^{E^c} dominates in magnitude over the other parameters, which leads to the second step dominating the time complexity. Thus, the runtime complexity of Algorithm 4 scales as $\mathcal{O}\Big(|\mathrm{n}^E + \mathrm{n}^{E^c} + \lceil \frac{T}{\Delta^P} \rceil| \cdot \mathrm{n}^{E^c} \cdot d\Big)$.

Thus the runtime complexity of evaluating the PCMHP likelihood intuitively lies in between the complexity of the MBP (Poisson) process and MHP.

C.11 Gradient $\mathscr{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}; T)$ Calculations

In this section we calculate the gradient of the negative log-likelihood $\mathscr{L}(\boldsymbol{\Theta}; T)$. Our starting point is Eq. (4.6). Taking the gradient with respect to the parameter vector $\boldsymbol{\Theta}$ of both sides and using linearity of the gradient operator, we have

$$\mathscr{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}; T) = \sum_{j \in E} \partial_{\boldsymbol{\Theta}} \mathscr{L}^j_{\text{IC-LL}}(\boldsymbol{\Theta}; T) + \sum_{j \in E^c} \partial_{\boldsymbol{\Theta}} \mathscr{L}^j_{\text{PP-LL}}(\boldsymbol{\Theta}; T). \tag{C.88}$$

Taking the gradient of Eq. (4.10) and Eq. (4.11), we obtain:

$$\partial_{\boldsymbol{\Theta}} \mathscr{L}^j_{\text{IC-LL}}(\boldsymbol{\Theta}; T) = \sum_{k=1}^{n^j} [\partial_{\boldsymbol{\Theta}} \Xi^j(o^j_k) - \partial_{\boldsymbol{\Theta}} \Xi^j(o^j_{k-1})] \left[1 - \frac{C^j_k}{\Xi^j(o^j_k) - \Xi^j(o^j_{k-1})} \right] \tag{C.89}$$

$$\partial_{\boldsymbol{\Theta}} \mathscr{L}^j_{\text{PP-LL}}(\boldsymbol{\Theta}; T) = \sum_{t^j_k \in \mathscr{H}^j_T} \frac{-\partial_{\boldsymbol{\Theta}} \xi^j(t^j_k)}{\xi^j(t^j_k)} + \partial_{\boldsymbol{\Theta}} \Xi^j(T). \tag{C.90}$$

Now we would need to calculate the gradients of the conditional intensity $\boldsymbol{\xi}_E$ and the compensator $\boldsymbol{\Xi}_E$. Recall that our parameter vector $\boldsymbol{\Theta}$ consists of the Hawkes kernel parameters contained in $\boldsymbol{\varphi}$ and the exogenous parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$, in view of Eq. (C.93).

Let $\theta \in \{\gamma^i, \nu^i\}$. Taking derivatives of Eq. (4.5) and Eq. (C.37), we see that

$$\begin{aligned}
\partial_{\gamma^k} \xi^i(t) &= \delta_{ik} \cdot \delta^k(t) + h^{ik}_E(t) \\
\partial_{\nu^k} \xi^i(t) &= \delta_{ik} + H^{ik}_E(t) \\
\partial_{\gamma^k} \Xi^i(t) &= \delta_{ik} + H^{ik}_E(t) \\
\partial_{\nu^k} \Xi^i(t) &= t \cdot \delta_{ik} + h^{ik}_E(t) * t.
\end{aligned}$$

Let θ be a parameter of the Hawkes kernel $\boldsymbol{\varphi}$. For example, if we use the exponential kernel, then $\theta \in \{\theta^{ij}, \alpha^{ij}\}$. Again, taking derivatives of Eq. (4.5) and Eq. (C.37), we get

$$\begin{aligned}
\partial_\theta \boldsymbol{\xi}(t) = &\sum_{i \in E^c} \sum_{t^i_k < t} \partial_\theta \boldsymbol{\varphi}^i_{E^c}(t - t^i_k) + \mathbf{h}_E(t) * \sum_{i \in E^c} \sum_{t^i_k < t} \partial_\theta \boldsymbol{\varphi}^i_{E^c}(t - t^i_k) \\
&+ \partial_\theta \mathbf{h}_E(t) \cdot \boldsymbol{\gamma} + \partial_\theta \mathbf{h}_E(t) * \left(\boldsymbol{\nu} + \sum_{i \in E^c} \sum_{t^i_k < t} \partial_\theta \boldsymbol{\varphi}^i_{E^c}(t - t^i_k) \right),
\end{aligned} \tag{C.91}$$

$$\begin{aligned}
\partial_\theta \boldsymbol{\Xi}(t) = &\sum_{i \in E^c} \sum_{t^i_k < t} \partial_\theta \boldsymbol{\Phi}^i_{E^c}(t - t^i_k) + \mathbf{h}_E(t) * \sum_{i \in E^c} \sum_{t^i_k < t} \partial_\theta \boldsymbol{\Phi}^i_{E^c}(t - t^i_k) \\
&+ \partial_\theta \mathbf{h}_E(t) * \left(\boldsymbol{\gamma} + \boldsymbol{\nu} \cdot t + \sum_{i \in E^c} \sum_{t^i_k < t} \partial_\theta \boldsymbol{\Phi}^i_{E^c}(t - t^i_k) \right).
\end{aligned} \tag{C.92}$$

The derivatives involving $\boldsymbol{\varphi}_{E^c}$ and $\boldsymbol{\Phi}_{E^c}$ are straightforward to calculate given the form of the Hawkes kernel. The derivative involving $\partial_\theta \mathbf{h}_E(t)$ can be calculated from $\boldsymbol{\varphi}_E(t)$ and its self-convolutions using the following observation:

$$
\begin{aligned}
\partial_\theta \mathbf{h}_E &= \partial_\theta \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n} \\
&= \sum_{n=1}^{\infty} \partial_\theta \boldsymbol{\varphi}_E^{\otimes n} \\
&= \partial_\theta \boldsymbol{\varphi}_E + \sum_{n=2}^{\infty} \partial_\theta \boldsymbol{\varphi}_E^{\otimes n} \\
&= \partial_\theta \boldsymbol{\varphi}_E + \sum_{n=2}^{\infty} \partial_\theta (\boldsymbol{\varphi}_E^{\otimes n-1} * \boldsymbol{\varphi}_E) \\
&= \partial_\theta \boldsymbol{\varphi}_E + \sum_{n=2}^{\infty} (\boldsymbol{\varphi}_E^{\otimes n-1} * \partial_\theta \boldsymbol{\varphi}_E + \partial_\theta \boldsymbol{\varphi}_E^{\otimes n-1} * \boldsymbol{\varphi}_E) \\
&= \partial_\theta \boldsymbol{\varphi}_E + \sum_{n=2}^{\infty} \left(\boldsymbol{\varphi}_E^{\otimes n-1} * \partial_\theta \boldsymbol{\varphi}_E + \partial_\theta \boldsymbol{\varphi}_E * \boldsymbol{\varphi}_E^{\otimes n-1} + \sum_{k=1}^{n-2} \boldsymbol{\varphi}_E^{\otimes k} * \partial_\theta \boldsymbol{\varphi}_E * \boldsymbol{\varphi}_E^{\otimes n-k-1} \right).
\end{aligned}
$$

We can leverage this recursive calculation to compute $\partial_\theta \mathbf{h}_E(t)$ efficiently. The method is presented in Algorithm 5.

Algorithm 5: Computing $\partial_\theta \mathbf{h}_E$ recursively over a partition of $[0, T]$

Input: partition, convergence threshold γ

Output: $\partial_\theta \mathbf{h}_E$

$B = \partial_\theta \boldsymbol{\varphi}_E$;

$S, A = B$

do

  $B = \boldsymbol{\varphi}_E * B$;

  $A = B + A * \boldsymbol{\varphi}_E$;

  $S = S + A$;

while $||A|| \geq \gamma$;

return S

Lastly, observe that Eq. (C.91) and Eq. (C.92) both contain a term involving convolution with $\partial_\theta \mathbf{h}_E(t)$. From Proposition C.5, computing this convolution requires us to have an expression for $\partial_\theta \mathbf{H}_E(t)$, the integral of $\partial_\theta \mathbf{h}_E(t)$. We can obtain this expression from $\partial_\theta \mathbf{h}_E(t)$

as follows:

$$\begin{aligned}\mathbf{h}_E(t) &= \sum_{n=1}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(t) \\ &= \boldsymbol{\varphi}_E(t) + \sum_{n=2}^{\infty} \boldsymbol{\varphi}_E^{\otimes n}(t) \\ &= \boldsymbol{\varphi}_E(t) + \sum_{n=1}^{\infty} (\boldsymbol{\varphi}_E^{\otimes n} * \boldsymbol{\varphi}_E)(t) \\ &= \boldsymbol{\varphi}_E(t) + (\mathbf{h}_E * \boldsymbol{\varphi}_E)(t).\end{aligned}$$

Integrating both sides, we get

$$\mathbf{H}_E(t) = \boldsymbol{\Phi}_E(t) + (\mathbf{h}_E * \boldsymbol{\Phi}_E)(t)$$

Taking the derivative of both sides with respect to θ and then applying the convolution product rule, we get

$$\begin{aligned}\partial_\theta \mathbf{H}_E(t) &= \partial_\theta \boldsymbol{\Phi}_E(t) + \partial_\theta (\mathbf{h}_E(t) * \boldsymbol{\Phi}_E(t)) \\ &= \partial_\theta \boldsymbol{\Phi}_E(t) + \partial_\theta \mathbf{h}_E(t) * \boldsymbol{\Phi}_E(t) + \mathbf{h}_E(t) * \partial_\theta \boldsymbol{\Phi}_E(t).\end{aligned}$$

In every iteration of Algorithm 4, the gradients $\partial_\theta \boldsymbol{\xi}(t)$ and $\partial_\theta \Xi(t)$ can be computed alongside $\boldsymbol{\xi}(t)$ and $\Xi(t)$ for every $\theta \in \boldsymbol{\Theta}$. These values are then passed one layer above to Eq. (C.89) and Eq. (C.90) to get the ICLL and PPLL gradients, respectively. Finally, the return values are passed to Eq. (C.88) to compute the overall gradient $\mathscr{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}; T)$. Computing the gradients does not affect the runtime complexity of Algorithm 4.

C.12 Sampling from PCMHP

Given $\bigcup_{j\in E^c} \mathcal{H}_{t^-}^j$, Algorithm 2 allows us to calculate the PCMHP(d,e) conditional intensity $\boldsymbol{\xi}_E(t)$. However, to 'continue' the process, we need to be able to sample events that occur beyond t. We can obtain samples from a PCMHP(d,e) process using the thinning algorithm presented in Algorithm 6. Here, we consider a specific form of the exogenous rate $\boldsymbol{\mu}(t)$, given by

$$\boldsymbol{\mu}(t) = \boldsymbol{\gamma} \cdot \boldsymbol{\delta}(t) + \boldsymbol{\nu}, \tag{C.93}$$

where $\boldsymbol{\gamma}, \boldsymbol{\nu} \in (\mathbb{R}^+)^d$ are learnable parameters.

Intuitively, this corresponds to a spike of magnitude $\boldsymbol{\gamma}$ at $t = 0$ and a constant rate $\boldsymbol{\nu}$ over time. This form of $\boldsymbol{\mu}(t)$ follows [102, 103].

C.12.1 Thinning Algorithm

Algorithm 6 is a modified version of Ogata's thinning algorithm [83] for the multivariate Hawkes process, with two modifications. First, this version of the thinning algorithm uses the appropriate upper bound ξ_E^{ub} for PCMHP(d,e), considering that the PCMHP(d,e) has a different conditional intensity from Hawkes, and the latter's upper bound will not be valid for PCMHP(d,e) except in the special case $E = \emptyset$. We derive this upper bound in Appendix C.12.2. Second, given that the conditional intensity $\boldsymbol{\xi}_E(t)$ in general cannot be written in closed form, we would need to apply the approximations detailed in Appendix C.7. In addition, sampling forward in time requires us to be able to compute $\boldsymbol{\xi}_E(t)$ at every candidate event time, and due to the convolution term, we have to keep track of the Hawkes intensity contributed by every previous event accepted by the thinning algorithm. We introduce a stepsize parameter $\Delta^t > 0$ that controls the discretization in time when we approximate the Hawkes intensity.

Algorithm 6: Simulating a Partial MBP Process on $[0, T)$ with Thinning

Input: kernel matrix $\boldsymbol{\varphi}(t)$, kernel integral matrix $\boldsymbol{\Phi}(t)$, exogenous parameters $\boldsymbol{\gamma}, \boldsymbol{\nu}$; dimension labels E, E^c; time horizon $T > 0$; partition $\mathscr{P}[0, T] = [0 : T]$ with increment $\Delta^{\mathscr{P}}$; threshold $\gamma^h > 0$; stepsize $\Delta^t > 0$

Output: $\mathscr{H}_T^j = \{t_k^j\}$ for $j = 1 : d$

initialize $t = 0; \mathscr{H}_T^j = \emptyset$ for $j = 1 : d; \mathscr{T}_\cup = \mathbf{a}_\cup = []$;

do

  $\boldsymbol{\Delta}[0 : T] = \text{conv}(\Delta, \boldsymbol{\varphi}_E, [0 : T])$;

  $\mathbf{h}_E[0 : T] = \mathbf{h}_E[0 : T] + \boldsymbol{\Delta}[0 : T]$;

while $\max_{p=0:P} \left(\left\| \boldsymbol{\Delta}[t_p] \right\|_{\max} \right) \geq \gamma^h$;

$\mathbf{H}_E[0 : T] = \boldsymbol{\Phi}[0 : T] + \text{conv}(\mathbf{h}_E, \boldsymbol{\Phi}, [0 : T])$;

$\mathbf{h}_E^{ub} = \max \mathbf{h}_E[0 : T]$, entrywise;

while $t < T$ **do**

  $\boldsymbol{\xi}_E^{ub} = \boldsymbol{\nu} + \mathbf{h}_E^{ub} \cdot (\boldsymbol{\gamma} + T\boldsymbol{\nu} + \sum_{j \in E^c} |\mathscr{H}_T^j| \boldsymbol{\Phi}_{E^c}^j(T)) + \sum_{j \in E^c} \sum_{t_k^j \leq t} \boldsymbol{\varphi}_{E^c}^j(t - t_k^j)$;

  $\xi_E^{ub} = \sum_{i=1}^d (\xi_E^{ub})^i$;

  $u \sim$ `uniform(0,1)`;

  $w = -\log \frac{u}{\xi_E^{ub}}$;

  $\mathscr{T} = \text{discretize}([t, t + w), \Delta^t)$;

  $\mathbf{a}[0 : |\mathscr{T}|] = \mathbf{0}$;

  for $i = 0 : |\mathscr{T}|$ **do**

    $\mathbf{a}[i] = \sum_{j \in E^c} \sum_{s \in \mathscr{H}_T^j} \boldsymbol{\varphi}_{E^c}^j(\mathscr{T}[i] - s)$;

  end

  $t = t + w$;

  $\mathscr{T}_\cup = \mathscr{T}_\cup \cup \mathscr{T}$;

  $\mathbf{a}_\cup[0 : |\mathscr{T}_\cup|] = [\mathbf{a}_\cup, \mathbf{a}]$;

  $\boldsymbol{\xi}_E = \boldsymbol{\nu} + \mathbf{h}_E(t) \cdot \boldsymbol{\gamma} + \text{conv}\left(\mathbf{h}_E, \boldsymbol{\nu} + \mathbf{a}_\cup, \mathscr{T}_\cup \cup \{t\}\right) + \sum_{j \in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}_{E^c}^j(t - t_k^j)$;

  $U \sim$ `uniform(0,1)`;

  if $U\xi_E^{ub} \leq \sum_{i=1}^d \xi_E^m(t)$ **then**

    $j = 1$;

    while $U\xi_E^{ub} \leq \sum_{i=1}^j \xi_E^m(t)$ **do**

      $j = j + 1$;

    end

    $t_k^j = t$;

    $\mathscr{H}_T^j = \mathscr{H}_T^j \cup \{t_k^j\}$;

  end

end

if $t_{k^j}^j < T$ **then**

  return $\mathscr{H}_T^j$ for $j = 1 : d$;

else

  return $\mathscr{H}_T^1, \ldots, \mathscr{H}_T^j \setminus \{t_k^j\}, \ldots, \mathscr{H}_T^d$;

end

Remark C.5. *Algorithm 6 has two discretization parameters $\Delta^{\mathscr{P}}$ and Δ^t. The first increment $\Delta^{\mathscr{P}}$ controls the discretization for $\mathbf{h}_E$, while the second increment Δ^t controls the discretization for the Hawkes intensity term $\sum_{j \in E^c} \sum_{s \in \mathscr{H}_T^j} \boldsymbol{\varphi}_{E^c}^j(t - s)$. In general, we would like to keep both increments as small as possible, but this comes at a tradeoff of computation time. However, there is a higher priority to keep $\Delta^{\mathscr{P}}$ small as $\mathbf{h}_E$, being an infinite sum, could hit convergence issues if $\Delta^{\mathscr{P}}$ is not small enough.*

C.12.2 Derivation of Thinning Upper Bounds for PCMHP(d,e)

In this section we derive the appropriate upper bounds for the PCMHP(d,e) process used in the thinning algorithm.

We were able to derive two different upper bounds. The first one can be calculated in a simple manner, and we use this in our implementation of Algorithm 3 to generate samples from PCMHP(d,e). However, this upper bound has the potential to be large if for some $(i,j) \in D \times D$, $\max_{t \le T} \mathbf{h}_E^{ij}(t)$ is high, causing the thinning algorithm to propose and subsequently reject many trial points. We therefore introduce a second upper bound that bounds the intensity more closely, leading to faster sampling. We use this in our popularity prediction use case in Section 4.7.

Upper Bound 1. We need to find an upper bound at arbitrary time $t > 0$ (that holds up until the next stochastic event) for each of the terms in the conditional intensity below.

$$\boldsymbol{\xi}_E(t) = \boldsymbol{\mu}(t) + (\mathbf{h}_E * \boldsymbol{\mu})(t) + \sum_{j \in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}_{E^c}^j(t - t_k^j) + \left(\mathbf{h}_E(t) * \left[\sum_{j \in E^c} \sum_{t_k^j < t} \boldsymbol{\varphi}_{E^c}^j(t - t_k^j) \right] \right), \tag{C.94}$$

where

$$\boldsymbol{\mu}(t) = \boldsymbol{\gamma} \cdot \boldsymbol{\delta}(t) + \boldsymbol{\nu}(t).$$

Let $\bar{\mathbf{h}}_E$ be the matrix whose (i,j) entry is $\max_{t \le T} \mathbf{h}_E^{ij}(t)$.

Let $\bar{\boldsymbol{\nu}} = \max \boldsymbol{\nu}_{s \le T}(t)$. Then the first term is bounded above by $\bar{\boldsymbol{\nu}}$.

For the second term, we have the following upper bound that holds for all t:

$$\begin{aligned}
(\mathbf{h}_E * \boldsymbol{\mu})(t) &= \mathbf{h}_E(t) * \left(\boldsymbol{\gamma} \cdot \boldsymbol{\delta}(t) + \boldsymbol{\nu}(t) \right) \\
&= \mathbf{h}_E(t) \cdot \boldsymbol{\gamma} + \int_0^t \mathbf{h}_E(t-s) \cdot \boldsymbol{\nu}(s) \mathrm{d}s \\
&\le \bar{\mathbf{h}}_E \cdot \boldsymbol{\gamma} + \int_0^T \bar{\mathbf{h}}_E \cdot \bar{\boldsymbol{\nu}} \mathrm{d}s \\
&= \bar{\mathbf{h}}_E \cdot (\boldsymbol{\gamma} + T\bar{\boldsymbol{\nu}}).
\end{aligned}$$

For the third term, as in the usual Thinning algorithm for Hawkes, an upper bound that holds until the next event is $\sum_{j \in E^c} \sum_{t_k^j \le t} \boldsymbol{\varphi}_{E^c}^j(t - t_k^j)$.

Lastly, for the fourth term, an upper bound that holds until the next event is given by:

$$\begin{aligned}
\mathbf{h}_E(t) * \sum_{j\in E^c}\sum_{t_k^j<t} \boldsymbol{\varphi}^j_{E^c}(t-t_k^j) &\le \mathbf{h}_E(t) * \sum_{j\in E^c}\sum_{t_k^j\le t} \boldsymbol{\varphi}^j_{E^c}(t-t_k^j) \\
&= \int_0^t \mathbf{h}_E(t-s)\cdot \sum_{j\in E^c}\sum_{t_k^j\le t} \boldsymbol{\varphi}^j_{E^c}(s-t_k^j)\mathrm{d}s \\
&= \bar{\mathbf{h}}_E \cdot \int_0^t \sum_{j\in E^c}\sum_{t_k^j\le t} \boldsymbol{\varphi}^j_{E^c}(s-t_k^j)\mathrm{d}s \\
&= \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c}\sum_{t_k^j\le t} \int_0^t \boldsymbol{\varphi}^j_{E^c}(s-t_k^j)\mathrm{d}s \\
&\overset{(a)}{=} \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c}\sum_{t_k^j\le t} \int_{-t_k^j}^{t-t_k^j} \boldsymbol{\varphi}^j_{E^c}(u)\mathrm{d}u \\
&\overset{(b)}{=} \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c}\sum_{t_k^j\le t} \int_0^{t-t_k^j} \boldsymbol{\varphi}^j_{E^c}(u)\mathrm{d}u \\
&= \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c}\sum_{t_k^j\le t} \boldsymbol{\Phi}^j_{E^c}(t-t_k^j) \\
&\overset{(c)}{\le} \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c}\sum_{t_k^j\le t} \boldsymbol{\Phi}^j_{E^c}(t) \\
&= \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c} |\mathscr{H}_t^j| \boldsymbol{\Phi}^j_{E^c}(t) \\
&\overset{(d)}{\le} \bar{\mathbf{h}}_E \cdot \sum_{j\in E^c} |\mathscr{H}_t^j| \boldsymbol{\Phi}^j_{E^c}(T).
\end{aligned}$$

where in (a) we made a change of variable $u = s - t_k^j$, in (b) we used the fact that $\varphi^{ij}(t) = 0$ for $t < 0$, in (c) we used the fact that since $\varphi^{ij}(t)$ is non-increasing, $\Phi^{ij}(t)$ is non-decreasing and thus $\Phi^{ij}(t) \ge \Phi^{ij}(t-s)$ for $s \ge 0$, and finally in (d) we bound $\boldsymbol{\Phi}^j_{E^c}$ by its maximum value on $[0, T]$.

Thus an upper bound at t until the next event is given by

$$\text{u.b.}(t) = \bar{\boldsymbol{\nu}} + \bar{\mathbf{h}}_E \cdot \left(\boldsymbol{\gamma} + T\bar{\boldsymbol{\nu}} + \sum_{j\in E^c} |\mathscr{H}_t^j| \boldsymbol{\Phi}^j_{E^c}(T) \right) + \sum_{j\in E^c}\sum_{t_k^j\le t} \boldsymbol{\varphi}^j_{E^c}(t-t_k^j). \tag{C.95}$$

Suppose we are given an exponential kernel $\varphi^{ij}(t) = \kappa^{ij}\theta^{ij}\exp^{-\theta^{ij}t}$, with corresponding $\Phi^{ij}(t) = \kappa^{ij}\left(1 - \exp^{-\theta^{ij}t}\right)$. We can simplify the upper bound by using $\boldsymbol{\Phi}^j_{E^c}(\infty) = \boldsymbol{\kappa}^j$ instead

of $\mathbf{\Phi}^j_{E^c}(T)$, since $\mathbf{\Phi}^j_{E^c}(\infty) \geq \mathbf{\Phi}^j_{E^c}(t)$ for any time t. Then we would have

$$\text{u.b.}(t) = \bar{\boldsymbol{\nu}} + \bar{\mathbf{h}}_E \cdot \left(\boldsymbol{\gamma} + T\bar{\boldsymbol{\nu}} + \sum_{j \in E^c} |\mathscr{H}^j_t| \boldsymbol{\kappa}^j \right) + \sum_{j \in E^c} \sum_{t^j_k \leq t} \boldsymbol{\varphi}^j_{E^c}(t - t^j_k).$$

Upper Bound 2. Suppose the exogenous intensity is given by

$$\boldsymbol{\mu}(t) = \boldsymbol{\gamma} \cdot \boldsymbol{\delta}(t) + \boldsymbol{\nu}.$$

An upper bound for the conditional intensity at t until the next stochastic event is

$$\begin{aligned} &\boldsymbol{\nu} + \sum_{j \in E^c} \sum_{t^j_k \leq t} \boldsymbol{\varphi}^j(t - t^j_k) + \max_{s \geq t} \mathbf{h}_E(s) \cdot \boldsymbol{\gamma} \\ &\quad + \mathbf{H}_E(\infty) \cdot \boldsymbol{\nu} + \left[\text{upper bound of } \mathbf{h}_E(t) * \sum_{j \in E^c} \sum_{t^j_k < t} \boldsymbol{\varphi}^j(t - t^j_k) \right], \end{aligned} \tag{C.96}$$

which we get by upper bounding each term in Eq. (C.94). The fourth term $\mathbf{H}_E(\infty) \cdot \boldsymbol{\nu}$ is obtained by noting that $\mathbf{H}_E(t)$ is a non-decreasing function and hence upper-bounded by $\mathbf{H}_E(\infty)$.

The tricky part here is obtaining an expression for the rightmost term. Let $t' \geq t$. Our goal is to find a function $\mathbf{f}(t)$ that satisifies

$$\mathbf{f}(t') \geq \mathbf{h}_E(t') * \sum_{j \in E^c} \sum_{t^j_k < t'} \boldsymbol{\varphi}^j(t - t^j_k) = \int_0^{t'} \mathbf{h}_E(t' - s) \cdot \sum_{j \in E^c} \sum_{t^j_k < s} \boldsymbol{\varphi}^j(s - t^j_k) \mathrm{d}s \tag{C.97}$$

We split the rightmost integral as

$$\int_0^t \mathbf{h}_E(t' - s) \cdot \sum_{j \in E^c} \sum_{t^j_k < s} \boldsymbol{\varphi}^j(s - t^j_k) \mathrm{d}s + \int_t^{t'} \mathbf{h}_E(t' - s) \cdot \sum_{j \in E^c} \sum_{t^j_k < s} \boldsymbol{\varphi}^j(s - t^j_k) \mathrm{d}s \tag{C.98}$$

and aim to bound these two terms separately.

To proceed, introduce

$$\hat{h}^{ij}_E(t) = \begin{cases} \max_s h^{ij}_E(s) & t < \operatorname{argmax}_s h^{ij}_E(s) \\ h^{ij}_E(t) & \text{otherwise} \end{cases}$$

Let $\hat{\mathbf{h}}_E(t) = [\hat{h}^{ij}_E(t)]$. Observe that

$$\hat{\mathbf{h}}_E(t) \geq \mathbf{h}_E(t) \tag{C.99}$$

for all $t \geq 0$.

Let $t' \geq s \geq t \geq 0$. We have $\hat{\mathbf{h}}_E(t'-s) \geq \mathbf{h}_E(t'-s)$. Multiplying both sides by $\sum_{j\in E^c}\sum_{t_k^j<s}\boldsymbol{\varphi}^j(s-t_k^j)$ and integrating over $s \in [0,t)$, we get

$$\int_0^t \hat{\mathbf{h}}_E(t'-s)\cdot\sum_{j\in E^c}\sum_{t_k^j<s}\boldsymbol{\varphi}^j(s-t_k^j)\mathrm{d}s \geq \int_0^t \mathbf{h}_E(t'-s)\cdot\sum_{j\in E^c}\sum_{t_k^j<s}\boldsymbol{\varphi}^j(s-t_k^j)\mathrm{d}s \tag{C.100}$$

Next, note that $\hat{\mathbf{h}}_E(t)$ is non-increasing. That is, given $t' \geq t \geq s$,

$$\hat{\mathbf{h}}_E(t-s) \geq \hat{\mathbf{h}}_E(t'-s) \tag{C.101}$$

So we have

$$\int_0^t \hat{\mathbf{h}}_E(t-s)\cdot\sum_{j\in E^c}\sum_{t_k^j<s}\boldsymbol{\varphi}^j(s-t_k^j)\mathrm{d}s \geq \int_0^t \mathbf{h}_E(t'-s)\cdot\sum_{j\in E^c}\sum_{t_k^j<s}\boldsymbol{\varphi}^j(s-t_k^j)\mathrm{d}s \tag{C.102}$$

Thus we have an upper bound for the first term.

For the second term, observe that since we are working under the assumption of no events between t and t',

$$\begin{aligned}
\int_t^{t'} \hat{\mathbf{h}}_E(t'-s)\cdot\sum_{j\in E^c}\sum_{t_k^j<s}\boldsymbol{\varphi}^j(s-t_k^j)\mathrm{d}s &= \int_t^{t'} \hat{\mathbf{h}}_E(t'-s)\cdot\sum_{j\in E^c}\sum_{t_k^j\leq t}\boldsymbol{\varphi}^j(s-t_k^j)\mathrm{d}s \\
&\leq \int_t^{t'} \hat{\mathbf{h}}_E(t'-s)\cdot\sum_{j\in E^c}\sum_{t_k^j\leq t}\boldsymbol{\varphi}^j(t-t_k^j)\mathrm{d}s \\
&= \int_t^{t'} \hat{\mathbf{h}}_E(t'-s)\mathrm{d}s\cdot\sum_{j\in E^c}\sum_{t_k^j\leq t}\boldsymbol{\varphi}^j(t-t_k^j) \\
&= [\hat{\mathbf{H}}_E(t'-t)-\hat{\mathbf{H}}_E(0)]\cdot\sum_{j\in E^c}\sum_{t_k^j\leq t}\boldsymbol{\varphi}^j(t-t_k^j) \\
&= \hat{\mathbf{H}}_E(t'-t)\cdot\sum_{j\in E^c}\sum_{t_k^j\leq t}\boldsymbol{\varphi}^j(t-t_k^j) \\
&\leq \hat{\mathbf{H}}_E(T-t)\cdot\sum_{j\in E^c}\sum_{t_k^j\leq t}\boldsymbol{\varphi}^j(t-t_k^j)
\end{aligned}$$

This is the upper bound for the second term. Here, $\hat{\mathbf{H}}_E(t)$ is given by

$$\hat{H}_E^{ij}(t) = \begin{cases} \max_s h_E^{ij}(s)\cdot t & t < u = \operatorname{argmax}_s h_E^{ij}(s) \\ H_E^{ij}(t) + [\max_s h_E^{ij}(s)\cdot u - H_E^{ij}(u)] & \text{otherwise} \end{cases}$$

C.13 Prediction of Expected Counts with PCMHP(d, e)

Once a PCMHP(d, e) model is fitted to a set of event sequences $\bigcup_{j \in E^c} \mathcal{H}^j_{T^-}$ and interval-censored counts $\bigcup_{j \in E} \{\mathsf{C}^j_k\}_{k=1}^{n^j}$, this can be used to predict the expected count of events in any interval of time past $t = T$.

Suppose that we observe events and counts on $[0, T^{train})$ and we wish to predict the expected count of events in every dimension on every subinterval of the partition $\mathscr{P}[T^{train}, T^{test})$, where $T^{test} > T^{train}$. The most straightforward way to do this is to continue the PCMHP(d, e) process on $[T^{train}, T^{test})$ by using Algorithm 6 to draw sample histories. For each sample, we count the number of events in each dimension on each subinterval of $\mathscr{P}[T^{train}, T^{test})$. The expected count on a subinterval would then be the average of the counts on the selected subinterval over the set of sample histories.

The problem with the previous approach is that it is not computationally efficient to perform the sampling, as we have to sample all dimensions simultaneously in Algorithm 6, especially problematic if some dimensions have a high background intensity.

Let $\mathscr{P}[T^{train}, T^{test}) = \bigcup_{i=1}^{P-1} [o_i, o_{i+1})$, where $o_1 = T^{train}$ and $o_P = T^{test}$, be a partition of $[T^{train}, T^{test})$. A computationally efficient scheme to predict expected counts can be done with the following three-step approach.

1. Sample only the E^c dimensions on $[T^{train}, T^{test})$.

2. For each sample, compute expected counts on $\mathscr{P}[T^{train}, T^{test})$ as $\{\boldsymbol{\Xi}_E(o_{i+1})$ - $\boldsymbol{\Xi}_E(o_i) | i \in 1 \cdots P-1\}$.

3. Compute the average of $\{\boldsymbol{\Xi}_E(o_{i+1})$ - $\boldsymbol{\Xi}_E(o_i) | i \in 1 \cdots P-1\}$ across samples.

This approach relies on two properties of the PCMHP(d, e) process. First, $\boldsymbol{\xi}_E(t)$ and $\boldsymbol{\Xi}_E(t)$ only depends on the E^c dimensions, and so we only need to actually sample these dimensions to calculate the intensity and compensator as these are independent of events that occurs in the E dimensions. Second, similar to what is stated in Proposition C.1 for the Hawkes process, the compensator $\boldsymbol{\Xi}_E(t)$ of a PCMHP(d, e) process can be interpreted as the expected count of events on $[0, t)$ given event sequences $\bigcup_{j \in E^c} \mathcal{H}^j_{t^-}$. Given a sample event sequence $\bigcup_{j \in E^c} \mathcal{H}^j_{o^-_{i+1}}$, the difference $\boldsymbol{\Xi}_E(o_{i+1})$ - $\boldsymbol{\Xi}_E(o_i)$ then represents the expected count of events in $[o_i, o_{i+1})$. By averaging over samples, we are averaging over histories $\bigcup_{j \in E^c} \mathcal{H}^j_{o^-_{i+1}}$), and so we get expected counts on $\mathscr{P}[T^{train}, T^{test})$.

Note that there is a subtle difference between the two approaches. The first approach returns event sequences in each dimension, which we then count and average over to get

expected values. On the other hand, the second approach directly estimates the expected counts as it uses the compensator of the process.

Fig. C.3 shows a comparison of the two methods over 1000 samples. We assume here that we observe data up until $T^{train} = 10$ and we wish to get expected counts on $\mathscr{P}[10,20) = [10,11),\cdots,[19,20)$. The solid lines show the estimate of the expected count. It is evident that there is very good agreement between the two approaches. The uncertainty clouds around the lines mean different things. Since the second approach directly estimates the expected counts, the cloud around the blue line represents the variance of the expected counts, whereas the red cloud represents the variance of the counts across different histories.

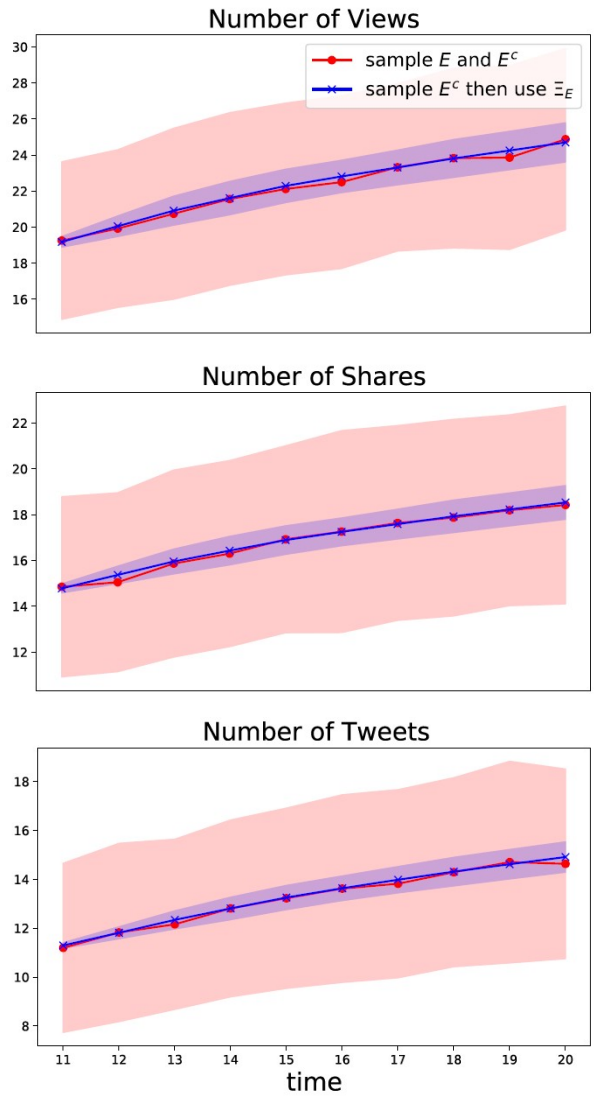


Figure C.3: Comparing the two ways of predicting expected counts with PCMHP(d, e) The first method samples all dimensions, while the second method samples only the Hawkes dimensions and uses the compensator of the process to estimate expected counts. In the figure, we observe data until $T^{train} = 10$ and compute event counts over $[10,11), \cdots, [19,20)$.

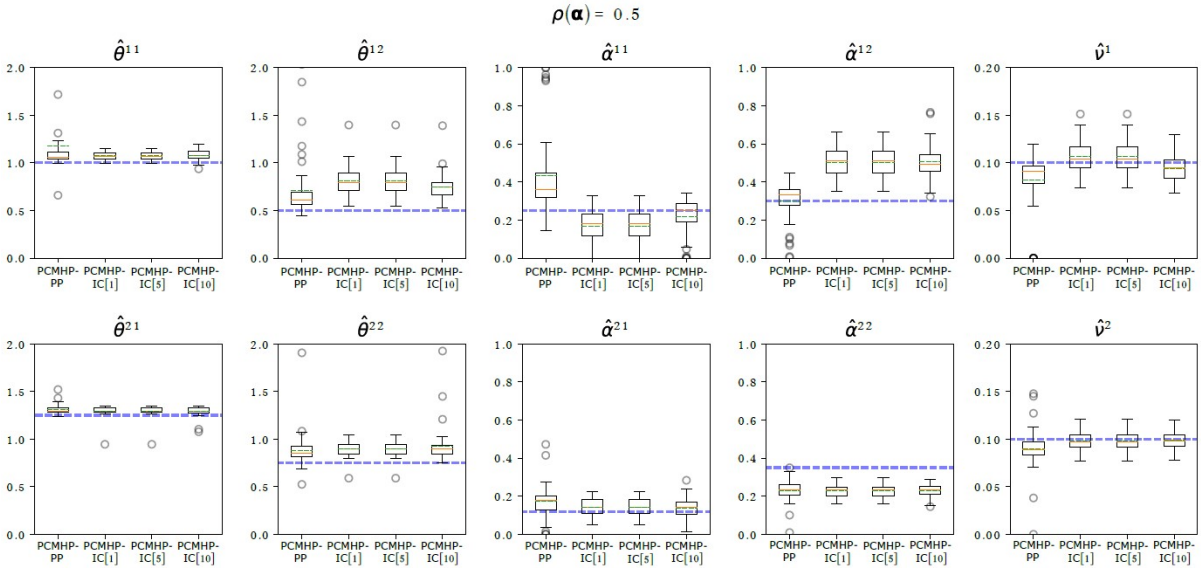


Figure C.4: *The MHP model parameters can be reliably estimated with the* PCMHP *model. Parameter recovery results for* $\rho(\boldsymbol{\alpha}) = 0.5$. In each subplot we show the parameter estimates obtained from the PCMHP$(2,1)$ model fitted on samples from a 2-dimensional MHP model using PCMHP-PP and PCMHP-IC. We consider three variants of interval censoring (observation window lengths 1, 5, and 10). The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively. The dashed blue lines show the original parameters of the MHP model.

C.14 Additional Results for Synthetic Parameter Recovery

C.14.1 Individual Parameter Estimates

Here we present individual parameter fits of the PCMHP$(2,1)$ model on 2-dimensional MHP data. Table C.1 lists the parameters for three considered values of the spectral radius: $\rho(\boldsymbol{\alpha}) \in \{0.5, 0.75, 0.9\}$. The $\rho(\boldsymbol{\alpha}) = 0.5$ indicates a clearly subcritical MHP; the $\rho(\boldsymbol{\alpha}) = 0.9$ corresponds to a MHP approaching the critical regime; and $\rho(\boldsymbol{\alpha}) = 0.75$ corresponds to an intermediate case between these two.

Fig. C.4, Fig. C.5 and Fig. C.6 show the PCMHP$(2,1)$-estimated $\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\nu}\}$ for the parameter sets corresponding to $\rho(\boldsymbol{\alpha}) = 0.5$, $\rho(\boldsymbol{\alpha}) = 0.75$ and $\rho(\boldsymbol{\alpha}) = 0.9$ in Table C.1.

Parameter recovery. Below we discuss results for the case $\rho(\boldsymbol{\alpha}) = 0.75$ (Fig. C.5). Results for $\rho(\boldsymbol{\alpha}) = 0.5$ and $\rho(\boldsymbol{\alpha}) = 0.9$ are similar. The horizontal dashed blue lines show the values used for generating the data. For each parameter, we plot four boxplots. The leftmost boxplot is the PCMHP-PP fit (*i.e.*, the fit on the timestamp dataset). The next three are the PCMHP-IC[1], PCMHP-IC[5], and PCMHP-IC[10] fits.

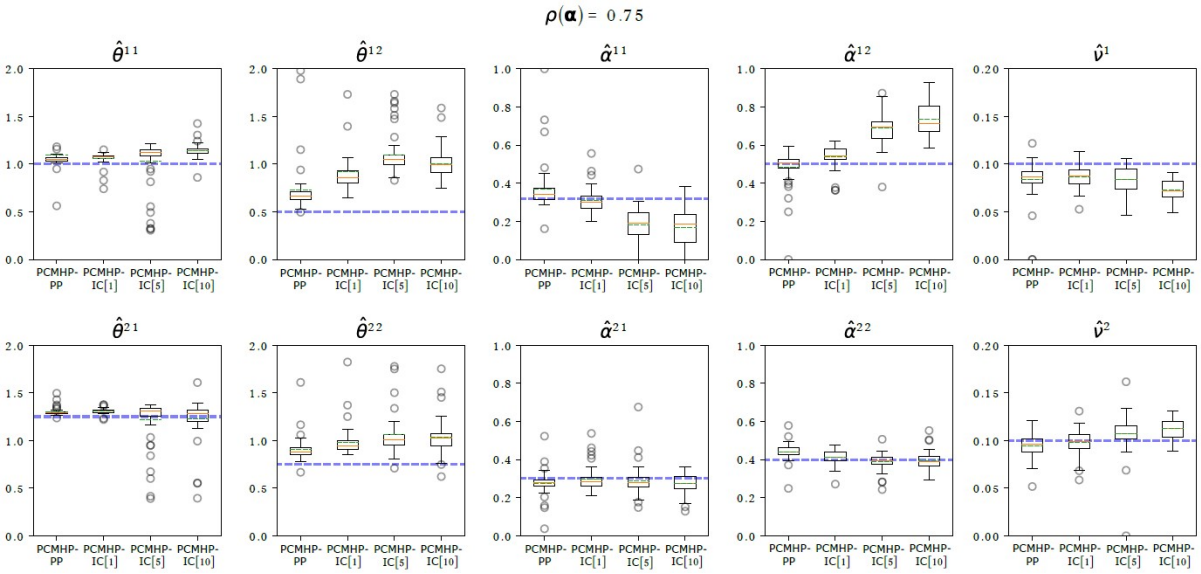


Figure C.5: *The MHP model parameters can be reliably estimated with the* PCMHP *model. Parameter recovery results for* $\rho(\boldsymbol{\alpha}) = 0.75$. In each subplot we show the parameter estimates obtained from the PCMHP(2, 1) model fitted on samples from a 2-dimensional MHP model using PCMHP-PP and PCMHP-IC. We consider three variants of interval censoring (observation window lengths 1, 5, and 10). The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively. The dashed blue lines show the original parameters of the MHP model.

Table C.1: Hawkes spectral radii and model parameters used in the parameter recovery synthetic experiment. We fix the initial impulse parameters $\gamma^0 = \gamma^1 = 0$ in our simulations.

$\rho(\boldsymbol{\alpha})$	θ^{11}	θ^{12}	θ^{21}	θ^{22}	α^{11}	α^{12}	α^{21}	α^{22}	ν^1	ν^2
0.5	1.0	0.5	1.25	0.75	0.25	0.3	0.12	0.35	0.1	0.1
0.75	1.0	0.5	1.25	0.75	0.32	0.5	0.3	0.4	0.1	0.1
0.9	1.0	0.5	1.25	0.75	0.4	0.5	0.3	0.6	0.1	0.1

We see that the PCMHP-PP estimates are tight around the generating MHP parameters for all parameters. This indicates that the model mismatch information loss (*i.e.*, of type (1)) appears to have a small impact on fitting quality. Arguably, we observe a slight overestimation for θ^{12} and θ^{22}, and a clear overestimation of α^{11} and underestimation of α^{22}. The ν^i parameters appear tightly recovered by PCMHP on the timestamp dataset.

On the partially interval-censored dataset, we continue to observe good fits. This indicates that PCMHP can successfully recover the generating MHP parameters even after interval-censoring. However, we see that the θ^{ij} parameters become increasingly overestimated as the observation window widens, particularly for θ^{12} and θ^{22}. Similarly, α^{12} is overestimated, and α^{11} is underestimated. The approximation quality degrades as the

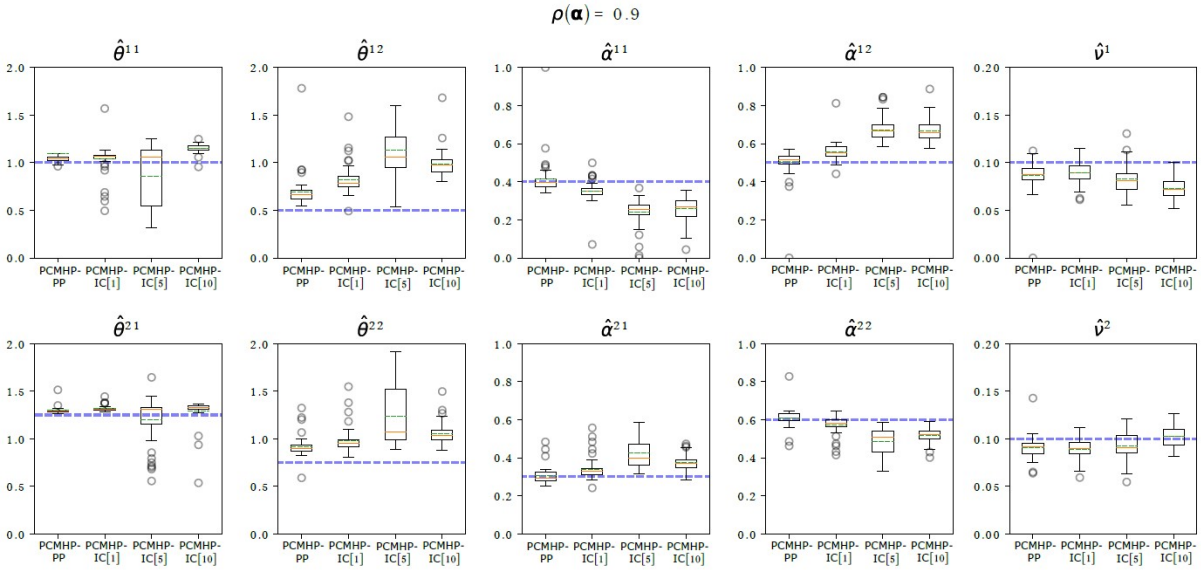


Figure C.6: *The MHP model parameters can be reliably estimated with the* PCMHP *model. Parameter recovery results for* $\rho(\boldsymbol{\alpha}) = 0.9$. In each subplot we show the parameter estimates obtained from the PCMHP$(2,1)$ model fitted on samples from a 2-dimensional MHP model using PCMHP-PP and PCMHP-IC. We consider three variants of interval censoring (observation window lengths 1, 5, and 10). The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively. The dashed blue lines show the original parameters of the MHP model.

observation window widens, indicating an increasing information loss of type (2).

Recovery of the spectral radius. We see in Fig. C.7 that the estimated spectral radius $\rho(\boldsymbol{\alpha})$ is close to the actual value regardless of the model mismatch and interval-censoring. The recovered spectral radius is estimated close to the original MHP spectral radius for all considered cases.

C.14.2 Convergence Analysis

We study the error convergence of the PCMHP$(2,1)$-estimated parameters $\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\nu}}\}$ and spectral radius $\rho(\hat{\boldsymbol{\alpha}})$ under different settings of (1) the time window T over which we fit the model and (2) the number of sequences used for joint fitting (see Remark C.3). Next, we fix T and investigate spectral radius recovery of the underlying MHP by fitting PCMHP$(2,1)$ over datasets of increasing spectral radii.

Varying T and $N_{sequences}$. The first column of Fig. C.8 shows RMSE$(\boldsymbol{\alpha})$, RMSE$(\boldsymbol{\theta})$, RMSE$(\boldsymbol{\nu})$ and $\Delta\rho$ as a function of the fitting window length T, respectively. Meanwhile, the second

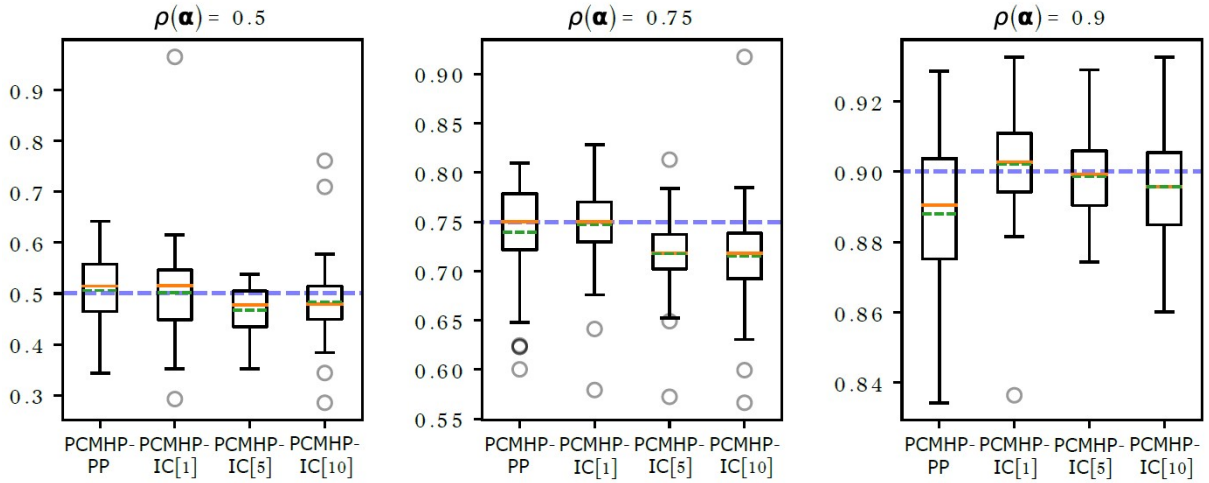


Figure C.7: *The spectral radius estimated by the* PCMHP *model approximates well the spectral radius of the generating MHP.* In each column we show the spectral radius estimated from the PCMHP$(2,1)$ model fitted on samples from a 2-dimensional MHP model (see parameters in Table C.1). Dashed lines show the spectral radii of the MHP model.

column of Fig. C.8 shows RMSE($\boldsymbol{\alpha}$), RMSE($\boldsymbol{\theta}$), RMSE($\boldsymbol{\nu}$) and $\Delta\rho$ as a function of the number of sequences in the joint fit, respectively. Both sets of plots correspond to the case $\rho(\boldsymbol{\alpha}) = 0.75$ parameter set in Table C.1. We see in Figs. C.8(a) and C.8(g) that RMSE($\boldsymbol{\alpha}$) and $\Delta\rho$ both converge to stable values as we increase the length of the fitting time window and the number of sequences in the joint fit.

Varying the spectral radius. In Fig. C.9(a) we evaluate how well our different models recover the MHP spectral radius. We fix $T = 100$, $N_{sequences} = 50$ and we plot the spectral radius deviation $\Delta\rho$ as a function of the generating MHP spectral radius $\rho(\boldsymbol{\alpha})$. We generate samples from 2D MHPs with increasing $\rho(\boldsymbol{\alpha}) \in \{0.1, 0.15, \ldots, 0.9, 0.95\}$ and we fit MHP, PCMHP-PP, and PCMHP-IC[k] for $k \in \{1,2,5,10,20\}$. We see that when the spectral radius is not too small, it is recovered well. This is intuitive since for small spectral radius, the MHP generates only a few events to fit on. We also notice that fitting with partially interval-censored data (*i.e.* the PCMHP-IC fits) tends to underestimate the spectral radius. In Fig. C.9(b) we plot the standard deviation of the estimated $\Delta\rho$ when fitted to MHP samples of different spectral radii for different model configurations.

Varying the number of MBP dimensions e. Here, we consider the PCMHP$(5,e)$ and vary $e \in \{0,1,2,3,4,5\}$. We fix $T = 100$ and $\rho(\boldsymbol{\alpha}) = 0.92$. We plot the parameter recovery error for $\boldsymbol{\alpha}$ (Fig. C.10(a)), $\boldsymbol{\theta}$ (Fig. C.10(b)) and $\boldsymbol{\nu}$ (Fig. C.10(c)). We see in Figs. C.10(a) and C.10(c) that RMSE($\boldsymbol{\alpha}$) and RMSE($\boldsymbol{\nu}$) increase with e, while in Fig. C.10(b) that RMSE($\boldsymbol{\theta}$) plateaus for

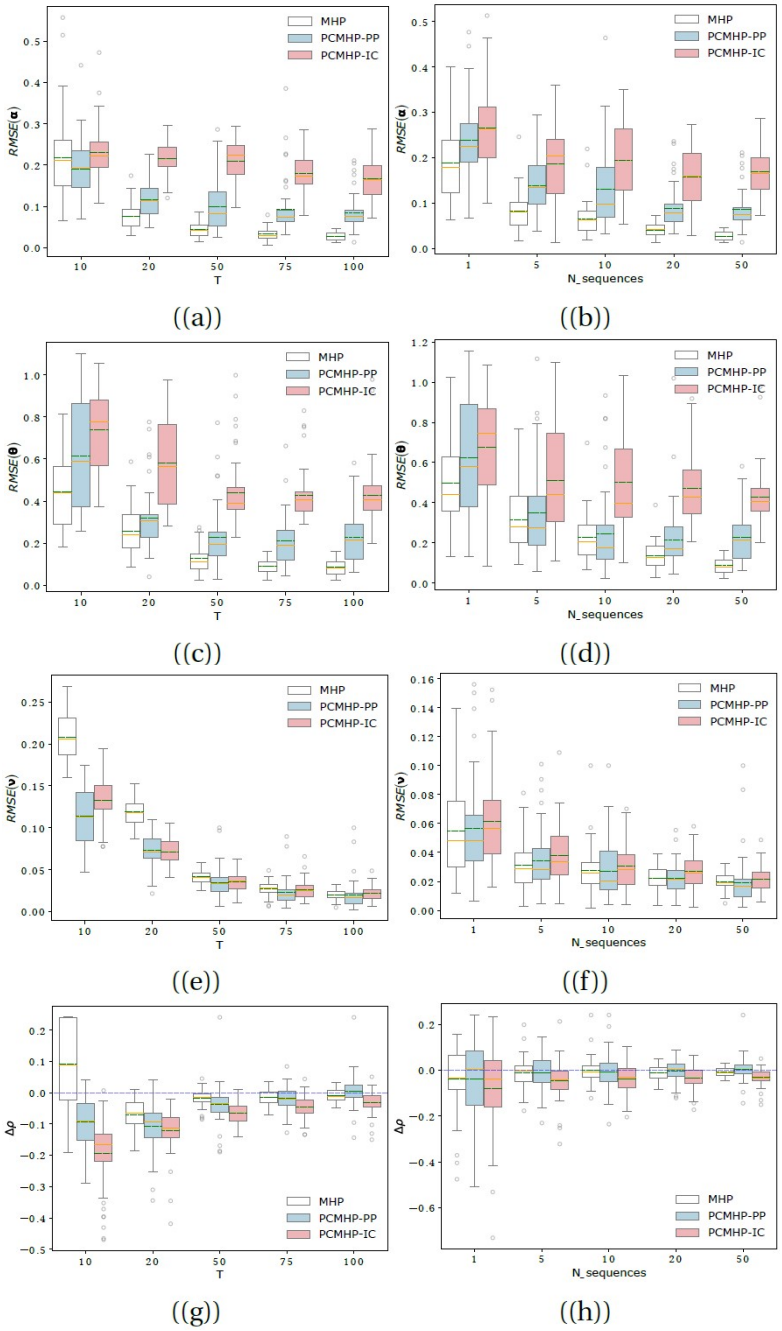


Figure C.8: The error of $\hat{\boldsymbol{\alpha}}$ (first row), $\hat{\boldsymbol{\theta}}$ (second row), $\hat{\boldsymbol{\nu}}$ (third row) and recovered spectral radius (fourth row) are plotted vs. varying T (left column) and $N_{sequences}$ (right column). Samples are drawn from a 2D MHP with $\rho(\boldsymbol{\alpha}) = 0.75$ and parameters in Table C.1. Default hyperparameters are $T = 100$, $N_{sequences} = 50$ and interval size = 5. In each plot we compare performance for three model fits: MHP, PCMHP-PP and PCMHP-IC as three boxplots. The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively.

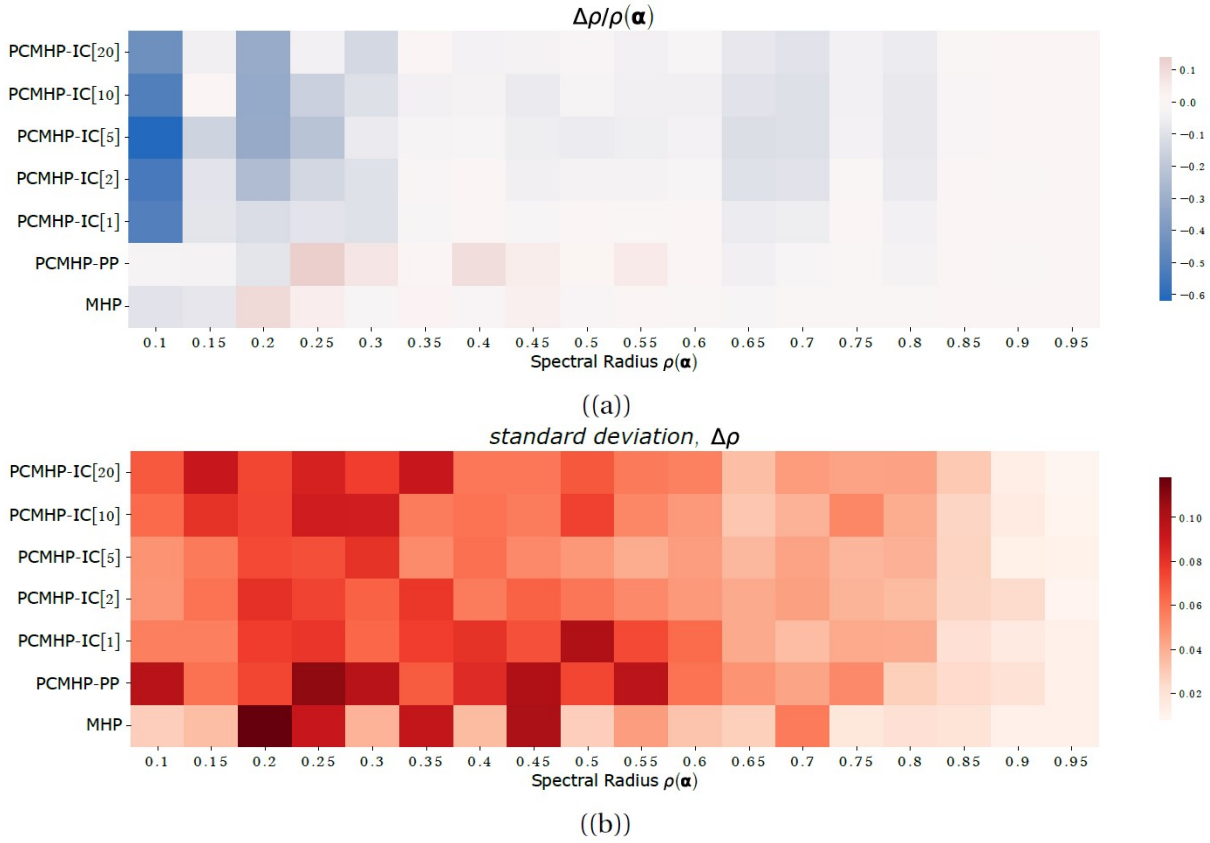
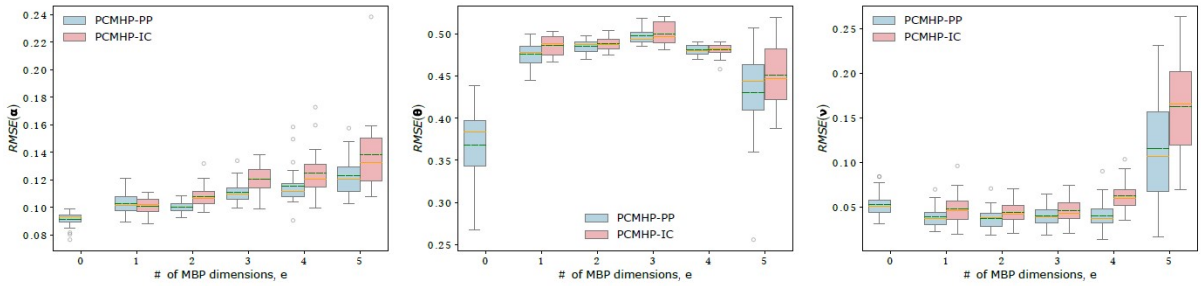


Figure C.9: (a) The deviation of recovering the spectral radius by our various approaches, as a function of the spectral radius itself. The x-axis shows a wide array of spectral radius values and the y-axis presents the different models used for fitting. The color shows the mean deviation $\Delta\rho/\rho(\boldsymbol{\alpha})$ over multiple fittings. (b) Standard deviation of $\Delta\rho$ vs. the spectral radius of the generating MHP. Rows correspond to the spectral radius $\rho(\boldsymbol{\alpha})$ of the MHP samples used for data generation, while the columns represent the different models used for fitting.

intermediate e. The behavior for $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$ is not surprising since a higher e implies more MHP dimensions are replaced with MBP, increasing the model mismatch information loss.

On the consistency of the PCMHP estimator. The convergence of RMSE($\boldsymbol{\alpha}$) and $\Delta\rho$ of the PCMHP-PP (Figs. C.8(a) and C.8(g)) for high T and number of sequences provide evidence of the consistency of the PCMHP MLE estimator for the MHP. Similar to [102], we perform numerical experiments in lieu of an analytical treatment to study limiting behavior since – as far as we are aware – there is no previous literature on the asymptotic theory (consistency and asymptotic normality) of the MLE for multivariate nonstationary MHPs (which the PCMHP falls under). Previous work has focused on the consistency of the univariate stationary [81], multivariate stationary [21, 44] and univariate nonstationary [18] cases. The asymptotic



((a)) ((b)) ((c))

Figure C.10: *Recovery error increases with the number of MHP dimensions we replace with MBP.* Error of $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\nu}}$ are plotted as functions of the number of MBP dimensions e. Samples are drawn from a 5-dimensional MHP with spectral radius $\rho(\boldsymbol{\alpha}) = 0.92$. Hyperparameters are $T = 100$, $N_{sequences} = 20$ and interval size = 1. We compare the PCMHP-PP, and PCMHP-IC model fits in each column. The mean and median estimates are indicated by the dashed green lines and solid orange lines, respectively.

theory of MLE for multivariate nonstationary MHPs and the partially interval-censored case are both open topics and fruitful directions for future work.

C.15 Additional Details for Popularity Prediction Experiment

In this section, we provide additional details on the fitting procedure for the PCMHP(3,3) and PCMHP(3,2) models on the online video popularity task and the filtering procedure we performed on `ACTIVE` to identify dynamic videos for performance evaluation.

C.15.1 Technical Details for Fitting

We present here details on fitting PCMHP(3,3) and PCMHP(3,2) models on the `ACTIVE` dataset of views/shares/tweets. Specifically, we discuss four points:

1. assigning weights to each dimension's contribution to $\mathscr{L}(\boldsymbol{\Theta}; T)$,
2. regularizing the exogenous parameter $\boldsymbol{\nu}$,
3. hyperparameter tuning, and
4. the optimization algorithm.

Assigning weights to each dimension's contribution to $\mathscr{L}(\boldsymbol{\Theta}; T)$. For the first point, recall that the negative log-likelihood for a PCMHP process given a mix of event sequences and interval-censored data is defined as the sum of the likelihood contribution in each dimension. Recall for the PCMHP that

$$\mathscr{L}(\boldsymbol{\Theta}; T) = \sum_{j \in E} \mathscr{L}^j_{\text{IC-LL}}(\boldsymbol{\Theta}; T) + \sum_{j \in E^c} \mathscr{L}^j_{\text{PP-LL}}(\boldsymbol{\Theta}; T).$$

A problem that we encountered with this formulation is that it treats the dimensions equally, but the scale of each dimension's values could differ largely from one another. For instance, in the online video popularity case study, view counts are orders of magnitudes higher than share counts and tweet counts, which would then cause the likelihood contribution of views to dominate the total likelihood, prioritizing fit on the views over the tweets and shares. This is problematic, in particular, for the PCMHP(3,2) model as we want to fit the tweets dimension well, as it drives the self- and cross-excitating behavior of the process. To solve this, we assign a dimension weight hyperparameter w^j to each dimension $j \in D$ which we multiply to the log-likelihood contribution of dimension j. Instead of $\mathscr{L}(\boldsymbol{\Theta}; T)$, we use a dimension-weighted version $\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w})$, given by

$$\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w}) = \sum_{j \in E} w^j \mathscr{L}^j_{\text{IC-LL}}(\boldsymbol{\Theta}; T) + \sum_{j \in E^c} w^j \mathscr{L}^j_{\text{PP-LL}}(\boldsymbol{\Theta}; T). \tag{C.103}$$

The corresponding gradients for this dimension-weighted version can be obtained by a trivial adjustment of Appendix C.11.

Regularizing the exogenous parameter $\boldsymbol{\nu}$. For the PCMHP models we fit in the online popularity case study in Section 4.7 and COVID-19 case study in Section 4.8, we optimize the parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\nu}\}$. To reduce the size of the parameter space, the initial impulse parameters $\boldsymbol{\gamma}$ are fixed as hyperparameters. Similar to [7], we found that adding a term that regularizes the exogenous parameter $\boldsymbol{\nu}$ on the dimension-weighted likelihood $\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w})$ in Eq. (C.103) further improves performance. We introduce another hyperparameter w^ν that controls the level of regularization on the $\mathscr{L}^1$ norm $\|\boldsymbol{\nu}\|_1$. Intuitively, using a higher w^ν is equivalent to biasing the model optimization to avoid taking on high values for the baseline intensities $\boldsymbol{\nu}$, thereby resulting to larger values in $\boldsymbol{\alpha}$. The regularized version of Eq. (C.103), where $\mathbf{w} = \{w^1, \cdots, w^d, w^\nu\}$, is then given by

$$\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w}) = \sum_{j \in E} w^j \mathscr{L}^j_{\text{IC-LL}}(\boldsymbol{\Theta}; T) + \sum_{j \in E^c} w^j \mathscr{L}^j_{\text{PP-LL}}(\boldsymbol{\Theta}; T) + w^\nu \|\boldsymbol{\nu}\|_1. \tag{C.104}$$

This is the version of the log-likelihood that we use to fit the models in Section 4.7 and Section 4.8. The parameter set that we optimize for is $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\nu}\}$.

Hyperparameter tuning. The hyperparameters that we tune in our model are shown in Table C.2. There are three different types of hyperparameters we consider. The first two types are the dimension weights $\{w^j\}$ and the $\boldsymbol{\nu}$ regularization weight w^ν we discussed above. The third hyperparameter is how we fix the $\boldsymbol{\gamma}$ parameter. We consider two different modes: (1) max, where we fix $\boldsymbol{\gamma}$ for a particular video to be fitted to the maximum value of the daily view, share, and tweet count on days 1-10, and (2) start, where we fix $\boldsymbol{\gamma}$ to the video's initial daily view, share, and tweet count. The candidate hyperparameters we sweep over are also shown in Table C.2. Note that we have different candidate hyperparameters for PCMHP(3,2) and PCMHP(3,3), which we selected based on heuristics.

We use days 1-90 to perform hyperparameter tuning and fitting. Specifically, we use days 1-75 as the training set for hyperparameter selection, and use days 76-90 as the validation set. Once we determine the best-performing hyperparameter set for a video, we refit the model on days 1-90. Lastly, the performance of the tuned model is evaluated on the test set, days 91-120.

Optimization algorithm. To optimize $\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w})$ over $\boldsymbol{\Theta}$, we use `IPOPT`, a nonlinear optimization solver for large-scale problems [123]. The solver requires the gradient of the objective and the Hessian of the Lagrangian. We wrap the procedure in Appendix C.11 as a function to compute the gradient iterately. For the Hessian of the Lagrangian, we use the limited-memory quasi-Newton approximated provided by `IPOPT`.

Table C.2: Hyperparameters of the PCMHP(d,e) models considered in the YouTube popularity prediction experiment

Hyperparameter	Description	(d,e): Values
$w^1 \cdots w^d$	dimension weights in $\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w})$	$(3,3)$: $\{[1000,1,1]\}$ $(3,2)$: $\{[1,1,1],[1,1,1000]\}$
w^ν	weight of the $\|\boldsymbol{\nu}\|_1$ term in $\mathscr{L}(\boldsymbol{\Theta}; T, \mathbf{w})$	$(3,3)$: $\{10,1000\}$ $(3,2)$: $\{1000\}$
γ^{init}	value of exogenous impulse at $t=0$	{max, start}

C.15.2 Filtering for Dynamic Videos

In our performance evaluation and baseline comparison in Section 4.7, we filtered the ACTIVE dataset for YouTube videos that have rich dynamics on days $21-90$. We showed that PCMHP$(3,2)$ performs best on these dynamic videos.

We present here the filtering procedure we implemented to arrive at the evaluation dataset in Fig. 4.4.

1. Filter ACTIVE for videos that have a mean daily tweet count on days 20-90 of at least 1. This ensures that the videos we consider have minimal activity on days 20-90.

2. Filter for videos that have less than 1000 tweets on days 1-90, as we only fit PCMHP$(3,2)$ on this set to avoid computational explosion, as discussed in Section 4.7.

3. Compute the standard deviation of the daily view count, daily share count and daily tweet count on days 20-90. Filter for videos that have higher than median standard deviation for each measure.

C.15.3 Performance Comparison of PCMHP and HIP

Fig. C.11 shows a comparison of the performance of PCMHP$(3,3)$, PCMHP$(3,2)$, and HIP on ACTIVE 20% (Appendix C.15.3) and DYNAMIC VIDEOS (Appendix C.15.3).

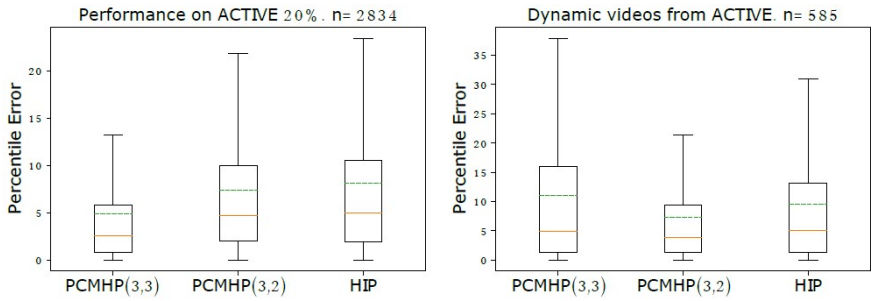


Figure C.11: Performance comparison of PCMHP(3,3), PCMHP(3,2) and HIP on (a) a random sample that comprises 20% of the videos in ACTIVE, and (b) the set of dynamic videos from ACTIVE. The dashed line and solid line indicate the mean and median estimates, respectively.

Table C.3: Goodness-of-fit measures on the PCMHP(2,1) models fitted on the COVID-19 daily case count and news article timestamp dataset.

Country	News KS p-value	Cases SK p-value	Cases Fit Score
UK	**0.47**	**0.09**	0.60
USA	0.04	**0.33**	0.48
Brazil	**0.18**	**0.22**	0.79
China	**0.17**	0	0.60
France	**0.33**	0	0.51
Germany	**0.99**	0	0.36
India	0	**0.09**	0.97
Italy	**0.09**	**0.24**	0.61
Spain	**0.15**	0.04	0.28
Sweden	**0.24**	0	0.72
Philippines	**0.96**	0	0.73

C.16 Additional Results for COVID-19 Experiment

In this section, we (1) elaborate on the goodness-of-fit tests we perform on the model fit of PCMHP(2,1) on the COVID-19 daily case count and news article dataset and (2) present an interpretation of PCMHP-fitted parameters for each country in our global sample.

C.16.1 Goodness-of-Fit Tests

To check model fit of PCMHP(2,1) on the cases-news data in Section 4.8, we perform separate goodness-of-fit tests for the news dimension and the case dimension. We found that the PCMHP(2,1) fits are statistically significant for 9 out of 11 countries on the news dimension and 5 out of 11 countries on the cases' dimension. For UK, Italy, and Brazil, the model fits are significant on both dimensions.

Observations in the news dimension are in the form of event timestamps $t_1^2, \cdots, t_{n^2}^2$, where n^2 is the number of news articles. The time-rescaling theorem [13] says that

$$\Xi^2(t_{j+1}^2) - \Xi^2(t_j^2) \sim \text{Exp}(1), \tag{C.105}$$

where $j \in 1 \cdots n^2 - 1$. Applying the previous formula on the observed data gives us n^2 samples from Exp(1). Fig. C.12 shows the $Q-Q$ plots for $\{\Xi^2(t_{j+1}^2) - \Xi^2(t_j^2)\}$, comparing the samples to the theoretical distribution for Exp(1). We can also use the Kolmogorov-Smirnov (KS) test to check the significance of Eq. (C.105), which we show in Table C.3.

Observations in the case dimension are in the form of daily case counts $\mathsf{C}^1_1, \cdots, \mathsf{C}^1_{120}$, corresponding to counts in $[0,1), \cdots, [119,120)$. Under the PCMHP$(2,1)$ process, the dimension 1 subprocess is a Poisson process, and so

$$\mathsf{C}^1_j \sim \text{Poi}(\Xi^1(j) - \Xi^1(j-1)). \tag{C.106}$$

To convert the C^1_j observations to samples from a single distribution, we use the Anscombe transform [2]. Given $x \sim \text{Poi}(m)$, the transformation

$$A : x \to 2\sqrt{x + \frac{3}{8}} \tag{C.107}$$

approximately yields standard normal samples: $A(x) \sim \mathcal{N}\left(2\sqrt{m+\frac{3}{8}}, 1\right)$ given large m. Combining Eq. (C.106) and Eq. (C.107) then subtracting the Gaussian mean, we see that

$$2\left(\sqrt{\mathsf{C}^1_j + \frac{3}{8}} - \sqrt{\Xi^1(j) - \Xi^1(j-1) + \frac{3}{8}}\right) \sim \mathcal{N}(0,1). \tag{C.108}$$

Fig. C.13 shows the $Q-Q$ plots for $\{2\left(\sqrt{\mathsf{C}^1_j + \frac{3}{8}} - \sqrt{\Xi^1(j) - \Xi^1(j-1) + \frac{3}{8}}\right)\}$, comparing the samples to the theoretical distribution for $\mathcal{N}(0,1)$. We can then apply the Skew-Kurtosis (SK) test for normality to check significance of Eq. (C.108), which we show in Table C.3.

We also introduce a sampling-based score to measure quality of fit of PCMHP$(2,1)$ to the daily case count. For each day j, we sample N_i observations from $\text{Poi}(\Xi^1(j) - \Xi^1(j-1))$. We then check whether the actual case count C^1_j is within the $[2.5\%, 97.5\%]$ band of the distribution of the N_i samples. We do this for each day $j \in 1 \ldots 120$ and average the result. In summary, we calculate

$$\frac{1}{120}\sum_{j=1}^{120} [\![\mathsf{C}^1_j \in [2.5\%, 97.5\%] \text{ interval of } \text{Poi}(\Xi^1(j) - \Xi^1(j-1))]\!], \tag{C.109}$$

The metric is simply the percentage of days where the actual count falls within the interval predicted by the model. The fit scores are tabulated in Table C.3.

C.16.2 Interpreting Individual Country Fits

Table C.4 contains the PCMHP$(2,1)$ parameter estimates $\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ for each country. Fig. C.14 shows the PCMHP$(2,1)$ fits for UK, USA, Brazil, China, France, Germany, Spain, Sweden, Philippines.

The PCMHP$(2,1)$ model parameters shown in Table C.4 are interpretable. Here, we discuss them in the order θ^{ij}, α^{ij}, and finally ν^i. We treat the parameters as approximations of the corresponding MHP parameters, similar to Section 4.6.

Table C.4: Parameters of the PCMHP(2, 1) models fitted on daily COVID-19 case counts and COVID-19 news article timestamps for the 11 countries we consider.

Country	θ^{11}	θ^{12}	θ^{21}	θ^{22}	α^{11}	α^{12}	α^{21}	α^{22}	ν^1	ν^2
UK	0.89	0.11	1.84	1.98	0.89	145.27	0.0004	0.29	1.99	0.36
USA	0.76	0.03	0.88	0.44	0.93	419.8	0	0.71	0.0054	0.85
Brazil	0.13	0.01	1.89	2.46	1.08	2198.7	0	0.38	0.24	1.47
China	0.93	4	1.39	4	0.59	0.1	0.012	0.39	0.0001	0.02
France	0.31	4	2.01	3.1	0.77	86.89	0.0015	0.38	0.025	0.07
Germany	0.98	0.18	2.05	1.9	0.75	247.16	0.0005	0.36	1.75	0.3
India	0.06	2.02	1.98	1.82	1.6	8.75	0	0.76	28.52	1.14
Italy	0.79	0.08	1.95	2.21	0.88	38.55	0.0002	0.91	0.11	0.62
Spain	0.4	0.08	1.57	1.62	0.73	316.27	0.0001	0.61	2.69	0.19
Sweden	0.48	0.53	2	1.89	0.98	2.13	0.0005	0.32	11.11	0.28
Philippines	0.18	2.54	2.31	1.94	1.11	14.6	0	0.4	0.04	0.19

The parameter θ^{ij} encodes the *speed of influence decay from dimension j to i*. Small values of θ^{ij} indicate that j influences i over a longer period, whereas large values imply a short half-life of the influence of j on i. From Table C.4 we see that India has the smallest θ^{11} (self-excitement of case numbers), which indicates the slow, consistent progression of COVID in India during the early phase. Brazil has the smallest θ^{12} (influence of news on cases); this indicates that Brazil cases did not immediately spike once Brazil made its way into the English-speaking news but instead steadily increased over an extended period. Similar interpretations can be made for θ^{21} and θ^{22}.

The parameter α^{ij} measures the *strength of influence from j to i* – it captures the expected number of i events triggered by a single j event. The multivariate Hawkes process, and consequently the PCMHP process, assumes $\alpha^{ij} > 0$ for all i and j; this implies that events can only self- or cross-excite other events but not inhibit. This is a modeling assumption that can be relaxed in future work. We can see from Table C.4 that Brazil, India, and the Philippines all have $\alpha^{11} > 1$, implying that COVID was highly contagious in these countries during the early stage (every infection generated more than one infection in average). We interpret α^{12} as how strong news preempts an increase in cases. Given that Brazil has the largest α^{12} in our sample, the news was particularly preemptive of cases there during the early stage. Conversely, α^{21} measures the expected number of news articles published after a single case. In countries with high α^{21}, news serves a reactionary role to an increase in cases. Among the countries considered, China has the highest α^{21}, probably because it was the first country in which COVID-19 has spread at a nation level.

The exogenous parameters ν^1 and ν^2 represent the *exogenous rates for cases and news*,

respectively. These parameters capture external factors that cannot be accounted by the self- and cross-excitation of cases and news. ν^1 is a measure of imported cases from other countries, while ν^2 captures the base level of reporting. India has the highest importation of cases among the countries considered as it has the highest ν^1, whereas Brazil, given its high ν^2, has the highest base reporting.

We remind our readers that these are English-speaking news (based mostly in English-speaking countries), indicating that Brazil had privileged coverage by news media. We hypothesize that it is probably due to its government's skepticism of the very existence of COVID-19 in the early days.

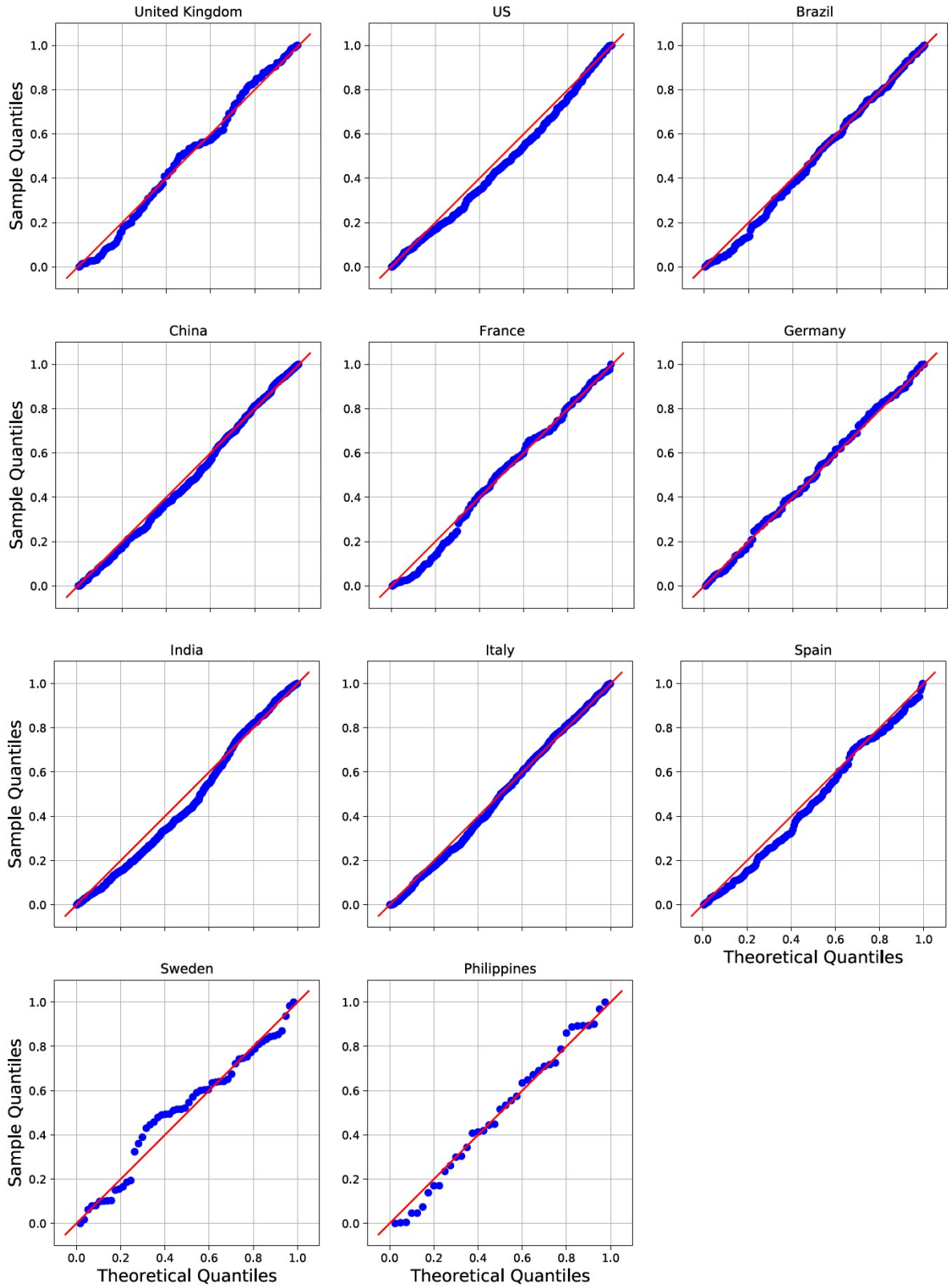


Figure C.12: Q-Q plots for observations in the news dimension in the COVID-19 country PCMHP(2,1) fits

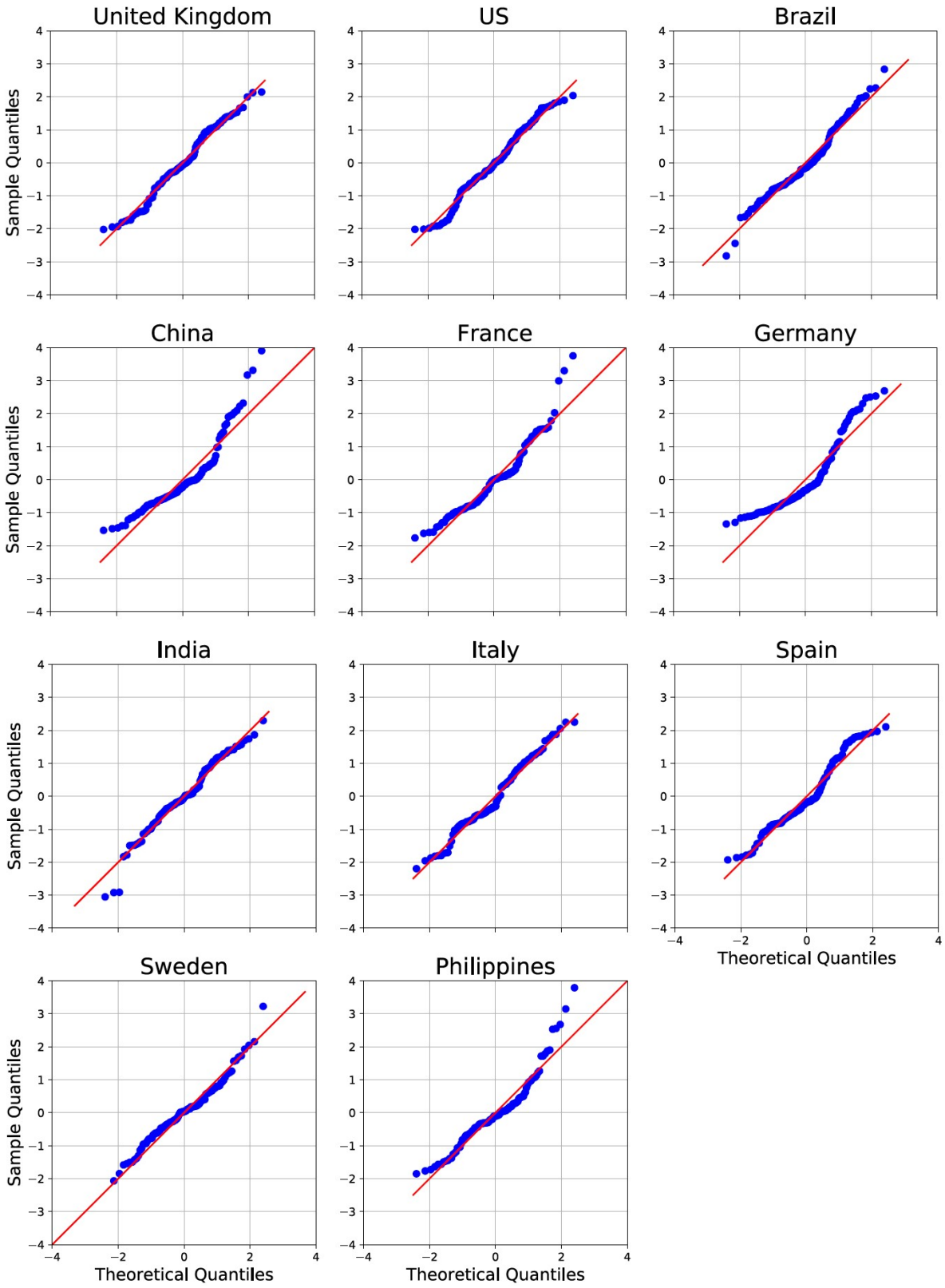


Figure C.13: Q-Q plots for observations in the cases dimension in the COVID-19 country PCMHP(2, 1) fits.

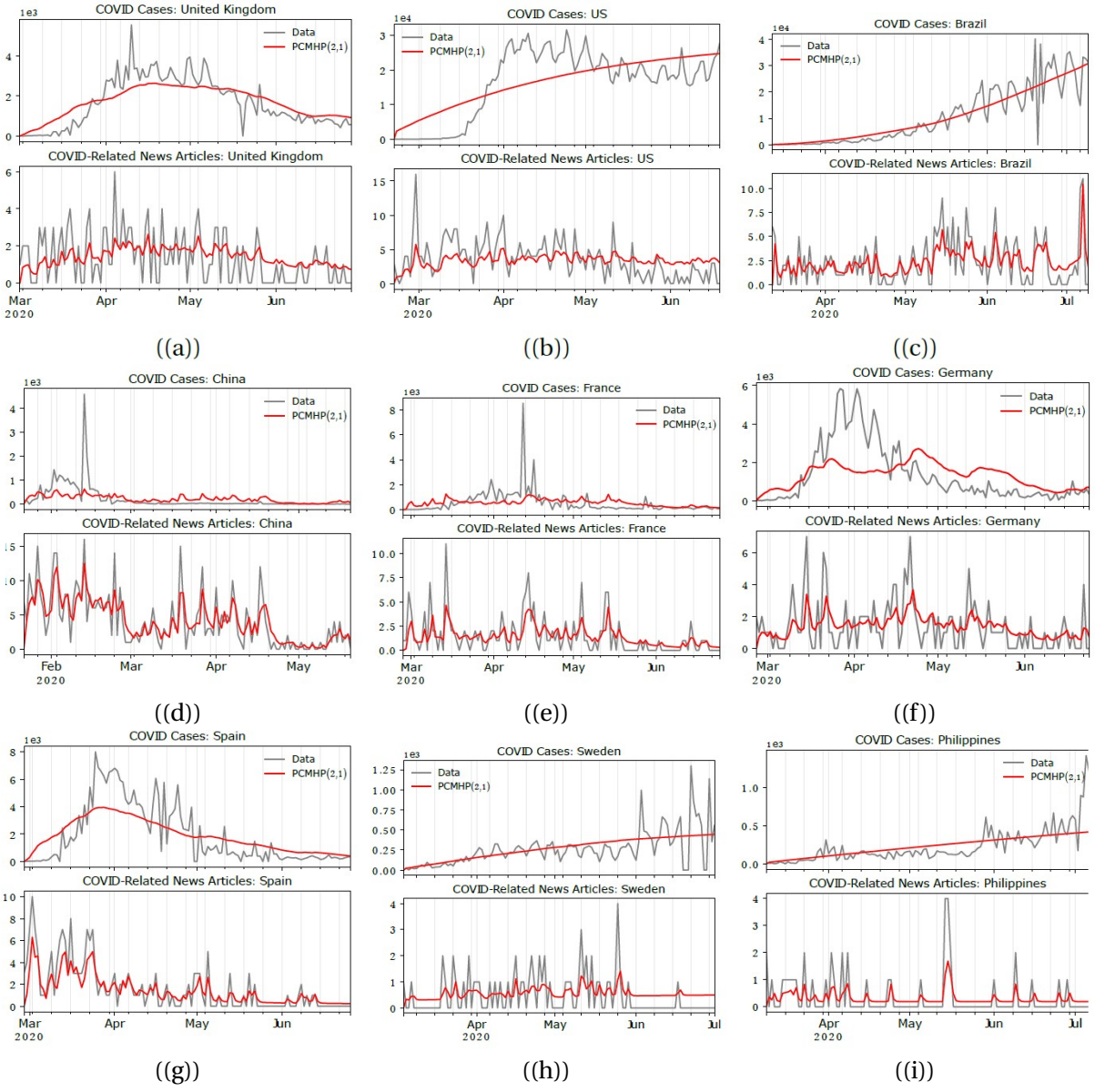


Figure C.14: Fit of the PCMHP(2, 1) model on the daily COVID-19 case count and COVID-19-related news articles, for the other countries in the global sample.

BIBLIOGRAPHY

[1] M. AGOVINO, M. R. CARILLO, AND N. SPAGNOLO, *Effect of Media News on Radicalization of Attitudes to Immigration*, Journal of Economics, Race, and Policy, (2021).

[2] F. J. ANSCOMBE, *THE TRANSFORMATION OF POISSON, BINOMIAL AND NEGATIVE-BINOMIAL DATA*, Biometrika, 35 (1948), pp. 246–254.

[3] AYLIEN, *AYLIEN Coronavirus Dataset*, 2020.

[4] E. BAKSHY, J. M. HOFMAN, W. A. MASON, AND D. J. WATTS, *Everyone's an influencer: quantifying influence on twitter*, in WSDM 2011.

[5] P. BAO, *Modeling and predicting popularity dynamics via an influence-based self-excited hawkes process*, in CIKM 2016.

[6] A. L. BERTOZZI, E. FRANCO, G. MOHLER, M. B. SHORT, AND D. SLEDGE, *The challenges of modeling and forecasting the spread of COVID-19*, PNAS, 117 (2020), pp. 16732–16738.

[7] Y. BESSY-ROLAND, A. BOUMEZOUED, AND C. HILLAIRET, *Multivariate Hawkes process for cyber insurance*, Annals of Actuarial Science, 15 (2021), pp. 14–39.

[8] M. BETZ, *Constraints and opportunities: what role for media development in countering violent extremism?*, (2016).

[9] K. BOGAERTS, A. KOMAREK, AND E. LESAFFRE, *Survival analysis with interval-censored data: a practical approach with examples in R, SAS, and BUGS*, Chapman and Hall/CRC, 2017.

[10] E. BOOTH, J. LEE, M.-A. RIZOIU, AND H. FARID, *Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention*, Journal of Sociology, (2024).

[11] E. BOREL, *Sur l'emploi du theoreme de Bernoulli pour faciliter le calcul d'une infinite de coefficients.*, CR Acad. Sci. Paris, 1942.

[12] P. BRÉMAUD AND L. MASSOULIÉ, *Stability of nonlinear Hawkes processes*, The Annals of Probability, 24 (1996), pp. 1563 – 1588.

[13] E. N. BROWN, R. BARBIERI, V. VENTURA, R. E. KASS, AND L. M. FRANK, *The Time-Rescaling Theorem and Its Application to Neural Spike Train Data Analysis*, Neural Computation, 14 (2002), pp. 325–346.

[14] R. BROWNING, D. SULEM, 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, (2021), pp. 1–28.

[15] B. CARPENTER, A. GELMAN, M. D. HOFFMAN, D. LEE, B. GOODRICH, M. BETANCOURT, M. BRUBAKER, J. GUO, P. LI, AND A. RIDDELL, *Stan: A probabilistic programming language*, Journal of statistical software, 76 (2017).

[16] P. R. CENTER, *Pew research center*, 2023.

[17] D.-G. D. CHEN, J. SUN, AND K. E. PEACE, *Interval-censored time-to-event data: methods and applications*, CRC Press, 2012.

[18] F. CHEN AND P. HALL, *Inference for a nonstationary self-exciting point process with an application in ultra-high frequency financial data modeling*, Journal of Applied Probability, 50 (2013), pp. 1006–1024.

[19] F. CHEYSSON AND G. LANG, *Spectral estimation of Hawkes processes from count data*, The Annals of Statistics, 50 (2022), pp. 1722 – 1746.

[20] W.-H. CHIANG, X. LIU, AND G. MOHLER, *Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates*, International Journal of Forecasting, 38 (2022), pp. 505–520.

[21] E. S. CHORNOBOY, L. P. SCHRAMM, AND A. F. KARR, *Maximum likelihood identification of neural point process systems*, Biological Cybernetics, 59 (1988), pp. 265–275.

[22] R. CHUNARA, J. R. ANDREWS, AND J. S. BROWNSTEIN, *Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak*, The American Journal of Tropical Medicine and Hygiene, 86 (2012), pp. 39–45.

[23] M. CINELLI, G. DE FRANCISCI MORALES, A. GALEAZZI, W. QUATTROCIOCCHI, AND M. STARNINI, *The echo chamber effect on social media*, PNAS, 118 (2021).

[24] K. CLAYTON ET AL., *Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media*, Political Behavior, 42 (2020), pp. 1073–1095.

[25] L. G. COOPER, *Chapter 6 market-share models*, in Marketing, vol. 5, Elsevier, 1993, pp. 259–314.

[26] R. CRANE AND D. SORNETTE, *Robust dynamic classes revealed by measuring the response function of a social system*, PNAS, 105 (2008), pp. 15649–15653.

[27] D. DALEY AND D. VERE-JONES, *An introduction to the theory of point processes. Vol. I*, Probability and its Applications (New York), Springer-Verlag, New York, second ed., 2003.

[28] A. DE, I. VALERA, N. GANGULY, S. BHATTACHARYA, AND M. GOMEZ-RODRIGUEZ, *Learning and forecasting opinion dynamics in social networks*, NIPS'16, 2016, pp. 397–405.

[29] E. DONG, H. DU, AND L. GARDNER, *An interactive web-based dashboard to track COVID-19 in real time*, The Lancet Infectious Diseases, 20 (2020), pp. 533–534.

[30] M. DU AND J. SUN, *Statistical analysis of interval-censored failure time data*, Chinese Journal of Applied Probability and Statistics, 37 (2021), pp. 627–654.

[31] R. DUNBAR, *Neocortex size as a constraint on group size in primates*, Journal of Human Evolution, 22 (1992), pp. 469–493.

[32] E. ERDOGAN, S. MA, A. BEYGELZIMER, AND I. RISH, *Statistical Models for Unequally Spaced Time Series*, pp. 626–630.

[33] K. FERGUSON, *Countering violent extremism through media and communication strategies*, Reflections, 27 (2016), p. 28.

[34] K. FOKIANOS, A. RAHBEK, AND D. TJØSTHEIM, *Poisson autoregression*, Journal of the American Statistical Association, 104 (2009), pp. 1430–1439.

[35] G. B. FOLLAND, *Real analysis: Modern techniques and their applications*, Wiley, New York, 1999.

[36] K. FUJITA, A. MEDVEDEV, S. KOYAMA, R. LAMBIOTTE, AND S. SHINOMOTO, *Identifying exogenous and endogenous activity in social media*, Phys. Rev. E, 98 (2018), p. 052304.

[37] M. GARETTO, E. LEONARDI, AND G. L. TORRISI, *A time-modulated Hawkes process to model the spread of COVID-19 and the impact of countermeasures*, Annual Reviews in Control, 51 (2021), pp. 551–563.

[38] S. GELPER, R. VAN DER LANS, AND G. VAN BRUGGEN, *Competition for attention in online social networks: Implications for seeding strategies*, Management Science, (2021).

[39] M. GHASSEMI, N. DALMASSO, S. LAMBA, V. POTLURU, T. BALCH, S. SHAH, AND M. VELOSO, *Online learning for mixture of multivariate hawkes processes*, in ICAIF 2022.

[40] GIFCT, *Content-sharing algorithms, processes, and positive interventions working group*, (2021).

[41] M. GOMEZ-RODRIGUEZ, D. BALDUZZI, AND B. SCHÖLKOPF, *Uncovering the temporal dynamics of diffusion networks*, in ICML 2011, 2011.

[42] M. GRUPPI, B. D. HORNE, AND S. ADAL, *Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles*, 2020.

[43] A. M. GUESS, P. BARBERA, S. MUNZERT, AND J. YANG, *The consequences of online partisan media*, PNAS, 118 (2021).

[44] X. GUO, A. HU, R. XU, AND J. ZHANG, *Consistency and Computation of Regularized MLEs for Multivariate Hawkes Processes*, Oct. 2018.
arXiv:1810.02955 [math, stat].

[45] S. GUPTA, G. JAIN, AND A. A. TIWARI, *Polarised social media discourse during covid-19 pandemic: evidence from youtube*, Behaviour & Information Technology, (2022), pp. 1–22.

[46] G. HACIYAKUPOGLU, J. Y. HUI, V. SUGUNA, D. LEONG, AND M. F. B. A. RAHMAN, *Countering fake news: A survey of recent global initiatives*, (2018).

[47] A. G. HAWKES, *Spectra of some self-exciting and mutually exciting point processes*, Biometrika, 58 (1971), pp. 83–90.

[48] A. HENSCHKE AND A. REED, *Toward an Ethical Framework for Countering Extremist Propaganda Online*, Studies in Conflict & Terrorism, (2021), pp. 1–18.

[49] M. D. HOFFMAN, A. GELMAN, ET AL., *The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.*, J. Mach. Learn. Res., 15 (2014), pp. 1593–1623.

[50] M. HOROWITZ, S. CUSHION, M. DRAGOMIR, S. GUTIÉRREZ MANJÓN, AND M. PANTTI, *A framework for assessing the role of public service media organizations in countering disinformation*, Digital Journalism, 10 (2022).

[51] J. H. HUBBARD AND B. B. HUBBARD, *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, Prentice Hall, second ed., 2002.

[52] S. JACKSON, *The double-edged sword of banning extremists from social media*, (2019).

[53] R. JOHAL, *Factiva: Gateway to business information*, Journal of Business & Finance Librarianship, 15 (2009), pp. 60–64.

[54] A. JOHNS, F. BAILO, E. BOOTH, AND M.-A. RIZOIU, *Labelling, shadow bans and community resistance: did meta's strategy to suppress rather than remove covid misinfo and conspiracy theory on facebook slow the spread?*, Media International Australia.

[55] Z. KAMONT, *Hyperbolic functional differential inequalities and applications*, vol. 486, Springer Science & Business Media, 2012.

[56] Y. KIM, J. HUANG, AND S. EMERY, *Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection*, Journal of medical Internet research, 18 (2016), p. e41.

[57] G. KING, B. SCHNEER, AND A. WHITE, *How the news media activate public expression and influence national agendas*, Science, 358 (2017), pp. 776–780.

[58] M. KIRCHNER, *Hawkes and INAR(∞) processes*, Stochastic Processes and their Applications, 126 (2016), pp. 2494–2525.

[59] ——, *An estimation procedure for the Hawkes process*, Quantitative Finance, 17 (2017), pp. 571–595.

[60] R. KOBAYASHI AND R. LAMBIOTTE, *Tideh: Time-dependent hawkes process for predicting retweet dynamics*, in ICWSM 2016, pp. 191–200.

[61] Q. KONG, E. BOOTH, F. BAILO, A. JOHNS, AND M.-A. RIZOIU, *Slipping to the Extreme: A Mixed Method to Explain How Extreme Opinions Infiltrate Online Discussions*, in AAAI ICWSM, vol. 16, 2022, pp. 524–535.

[62] Q. KONG, P. CALDERON, R. RAM, O. BOICHAK, AND M.-A. RIZOIU, *Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems*, in ACM Web Conference, 2023.

[63] Q. KONG, M. A. RIZOIU, AND L. XIE, *Describing and Predicting Online Items with Reshare Cascades via Dual Mixture Self-exciting Processes*, International Conference on Information and Knowledge Management, Proceedings, (2020), pp. 645–654.

[64] R. KROHN AND T. WENINGER, *Modelling online comment threads from their start*, in 2019 IEEE International Conference on Big Data, 2019, pp. 820–829.

[65] B. KULKARNI, S. AGARWAL, A. DE, S. BHATTACHARYA, AND N. GANGULY, *Slant+: A nonlinear model for opinion dynamics in social networks*, in ICDM, IEEE, 2017, pp. 931–936.

[66] J. LEE, E. BOOTH, H. FARID, AND M.-A. RIZOIU, *Misinformation is not about Bad Facts: An Analysis of the Production and Consumption of Fringe Content*.
mar 2024.

[67] L. LI AND H. ZHA, *Learning parametric models for social infectivity in multi-dimensional hawkes processes*, in AAAI 2014.

[68] T. LI AND Y. KE, *Tweedie-hawkes processes: Interpreting the phenomena of outbreaks*, in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[69] Y. LIAO, S. WANG, E. HAN, J. LEE, AND D. LEE, *Characterization and early detection of evergreen news articles*, in ML and Knowledge Discovery in Databases, 2020.

[70] J. LISCHKA AND M. GARZ, *Clickbait news & algorithmic curation: Game theory framework of the relation bet. journalism, users and platforms*, New Media & Society.

[71] J. MA, W. GAO, P. MITRA, S. KWON, B. J. JANSEN, K.-F. WONG, AND M. CHA, *Detecting rumors from microblogs with recurrent neural networks*, in IJCAI 2016, 2016.

[72] L. MA, Y. FENG, D.-G. D. CHEN, AND J. SUN, *Interval-censored time-to-event data and their applications in clinical trials*, Clinical Trial Biostatistics and Biopharmaceutical Applications, 307 (2014).

[73] K. MALEKNEJAD, R. MOLLAPOURASL, AND P. MIRZAEI, *Numerical solution of volterra functional integral equation by using cubic b-spline scaling functions*, Numerical Methods for Partial Differential Equations, 30 (2014), pp. 699–722.

[74] MEDIA BIAS FACT CHECK, *News.com.au – Bias and Credibility*, 2024.

[75] G. J. MILINOVICH, G. M. WILLIAMS, A. C. A. CLEMENTS, AND W. HU, *Internet-based surveillance systems for monitoring emerging infectious diseases*, The Lancet Infectious Diseases, 14 (2014), pp. 160–168.

[76] S. MISHRA, M. A. RIZOIU, AND L. XIE, *Feature driven and point process approaches for popularity prediction*, CIKM 2016, pp. 1069–1078.

[77] F. MORSTATTER AND H. LIU, *Discovering, assessing, and mitigating data bias in social media*, Online Social Networks and Media, 1 (2017), pp. 1–13.

[78] P. MUKHERJEE, S. DUTTA, AND A. DE BRUYN, *Did clickbait crack the code on virality?*, Journal of the Academy of Marketing Science, 50 (2022), pp. 482–502.

[79] D. J. NAVARRO, T. L. GRIFFITHS, M. STEYVERS, AND M. D. LEE, *Modeling individual differences using dirichlet processes*, Journal of Mathematical Psychology, 50 (2006), pp. 101–122.
Special Issue on Model Selection: Theoretical Developments and Applications.

[80] E. NEKMAT, *Nudge effect of fact-check alerts: source influence and media skepticism on sharing of news misinformation in social media*, Social Media+ Society, 6 (2020).

[81] Y. OGATA, *The asymptotic behaviour of maximum likelihood estimators for stationary point processes*, Annals of the Institute of Statistical Mathematics, 30 (1978), pp. 243–261.

[82] Y. OGATA, *On Lewis' simulation method for point processes*, IEEE Transactions on Information Theory, 27 (1981), pp. 23–31.

[83] Y. OGATA, *Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes*, Journal of the American Statistical Association, 83 (1988), pp. 9–27.

[84] OPENAI, *Chatgpt*, 2023.
Software tool.

[85] G. PAOLACCI, J. CHANDLER, AND P. G. IPEIROTIS, *Running experiments on amazon mechanical turk*, Judgment and Decision making, 5 (2010), pp. 411–419.

[86] O. PAPASPILIOPOULOS, G. O. ROBERTS, AND M. SKÖLD, *A general framework for the parametrization of hierarchical models*, Statistical Science, (2007), pp. 59–73.

[87] S. B. PARIKH, V. PATIL, R. MAKAWANA, AND P. K. ATREY, *Towards impact scoring of fake news*, in MIPR 2019, IEEE.

[88] F. PEDREGOSA AND ET.AL., *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[89] M. PEUCKER, T. J. FISHER, AND J. DAVEY, *Mainstream media use in far-right online ecosystems*, tech. rep., Aug. 2022.

[90] C. L. PHILLIPS, J. M. PARR, E. A. RISKIN, AND T. PRABHAKAR, *Signals, systems, and transforms*, Prentice Hall Upper Saddle River, 2003.

[91] J. D. PINTER, *Global Optimization in Action*, vol. 6 of Nonconvex Optimization and Its Applications, Springer US, Boston, MA, 1996.

[92] H. PINTO, J. M. ALMEIDA, AND M. A. GONÇALVES, *Using early view patterns to predict the popularity of youtube videos*, in WSDM, 2013, pp. 365–374.

[93] E. PORTER AND T. J. WOOD, *Fact checks actually work, even on facebook. but not enough people see them.*, The Washington Post, (2021).

[94] E. RADER, A. VELASQUEZ, K. D. HALES, AND H. KWOK, *The gap between producer intentions and consumer behavior in social media*, in Proceedings of the 2012 ACM International Conference on Supporting Group Work, 2012, pp. 249–252.

[95] C. RADSCH, *Media development and countering violent extremism: An uneasy relationship, a need for dialogue*, Center for International Media Assistance. (2016), (2016).

[96] R. RAM, E. THOMAS, D. KERNOT, AND M.-A. RIZOIU, *Detecting Extreme Ideologies in Shifting Landscapes: an Automatic & Context-Agnostic Approach*, aug 2022.

[97] J. G. RASMUSSEN, *Lecture Notes: Temporal Point Processes and the Conditional Intensity Function*, arXiv:1806.00221 [stat], (2018).
arXiv: 1806.00221.

[98] K. REHFELD, N. MARWAN, J. HEITZIG, AND J. KURTHS, *Comparison of correlation analysis techniques for irregularly sampled time series*, Nonlinear Processes in Geophysics, 18 (2011), pp. 389–404.

[99] N. REIMERS AND I. GUREVYCH, *Sentence-bert: Sentence embeddings using siamese bert-networks*, in EMNLP 2019.

[100] M. A. RIZOIU, T. GRAHAM, R. ZHANG, Y. ZHANG, R. ACKLAND, AND L. XIE, *DEBATENIGHT: The role and influence of socialbots on twitter during the first 2016 U.S. presidential debate*, 12th International AAAI Conference on Web and Social Media, ICWSM 2018, (2018), pp. 300–309.

[101] M.-A. RIZOIU, Y. LEE, S. MISHRA, AND L. XIE, *A Tutorial on Hawkes Processes for Events in Social Media*, Frontiers of Multimedia Research, 12 2017, pp. 191–218.

[102] M.-A. RIZOIU, A. SOEN, S. LI, P. CALDERON, L. DONG, A. K. MENON, AND L. XIE, *Interval-censored Hawkes processes*, Journal of Machine Learning Research, 23 (2022), pp. 1–84.

[103] M. A. RIZOIU, L. XIE, S. SANNER, M. CEBRIAN, H. YU, AND P. VAN HENTERYCK, *Expecting to be HIP: Hawkes intensity processes for social media popularity*, WWW 2017, (2017), pp. 735–744.

[104] M.-A. RIZOIU AND L. X. XIE, *Online popularity under promotion: Viral potential, forecasting, and the economics of time*, ICWSM, 11 (2017), pp. 182–191.

[105] M. M. U. RONY, N. HASSAN, AND M. YOUSUF, *Diving deep into clickbaits: Who use them to what extents in which topics with what effects?*, 2017.

[106] P. J. SCHNEIDER AND T. A. WEBER, *Estimation of self-exciting point processes from time-censored data*, Physical Review E, 108 (2023), p. 015303.

[107] F. SCHOENBERG, *Estimating Covid-19 transmission time using Hawkes point processes*, The Annals of Applied Statistics, 17 (2023).

[108] A. SHARMA, A. GHOSH, AND M. FITERAU, *Generative sequential stochastic model for marked point processes*, in ICML Time Series Workshop, 2019.

[109] H. SHEN, D. WANG, C. SONG, AND A.-L. BARABÁSI, *Modeling and predicting popularity dynamics via reinforced poisson processes*, in AAAI, vol. 28, 2014.

[110] K. SHESHADRI AND M. P. SINGH, *The public and legislative impact of hyperconcentrated topic news*, Science Advances, 5 (2019).

[111] L. SHLOMOVICH, E. A. COHEN, N. ADAMS, AND L. PATEL, *Parameter estimation of binned hawkes processes*, Journal of Computational and Graphical Statistics, 31 (2022), pp. 990–1000.

[112] L. SHLOMOVICH, E. A. K. COHEN, AND N. ADAMS, *A parameter estimation method for multivariate binned Hawkes processes*, Statistics and Computing, 32 (2022), p. 98.

[113] K. SHU, S. WANG, AND H. LIU, *Beyond news contents: The role of social context for fake news detection*, WSDM '19, 2019, p. 312,Äì320.

[114] J. SUN, *The statistical analysis of interval-censored failure time data*, vol. 3, Springer, 2006.

[115] W. H. TAN AND F. CHEN, *Predicting the popularity of tweets using internal and external knowledge: an empirical bayes type approach*, AStA, 105 (2021), pp. 335–352.

[116] S. TEAM, *Cmdstanpy (0.9.76)*, 2023.

[117] C. TOMASI, *Estimating gaussian mixture densities with em–a tutorial*, Duke University, (2004), pp. 1–8.

[118] M. TSAGKIAS, W. WEERKAMP, AND M. DE RIJKE, *Predicting the volume of comments on online news stories*, in CIKM 2009, 2009.

[119] H. J. T. UNWIN, I. ROUTLEDGE, S. FLAXMAN, M.-A. RIZOIU, S. LAI, J. COHEN, D. J. WEISS, S. MISHRA, AND S. BHATT, *Using Hawkes Processes to model imported and local malaria cases in near-elimination settings*, PLOS Computational Biology, 17 (2021).

[120] I. VALERA AND M. GOMEZ-RODRIGUEZ, *Modeling Adoption and Usage of Competing Products*, in ICDM, Nov. 2015.

[121] L. VAN DER MAATEN AND G. HINTON, *Visualizing data using t-sne*, Journal of Machine Learning Research, 9 (2008), pp. 2579–2605.

[122] T. VENTURINI AND R. ROGERS, *Api-based research or how can digital sociology and journalism studies learn from the facebook and cambridge analytica data breach*, Digital Journalism, 7 (2019), pp. 532–540.

[123] A. WACHTER AND L. T. BIEGLER, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical Programming, 106 (2006), pp. 25–57.

[124] L. WENG, A. FLAMMINI, A. VESPIGNANI, AND F. MENCZER, *Competition among memes in a world with limited attention*, Scientific reports, 2 (2012), p. 335.

[125] S. C. WOOLLEY AND P. N. HOWARD, *Computational propaganda: Political parties, politicians, and political manipulation on social media*, Oxford University Press, 2018.

[126] S. WU, M.-A. RIZOIU, AND L. XIE, *Estimating Attention Flow in Online Video Networks*, ACM HCI, 3 (2019), pp. 1–25.

[127] ——, *Variation across scales: Measurement fidelity under twitter data sampling*, in Proceedings of the international AAAI conference on web and social media, vol. 14, 2020, pp. 715–725.

[128] H. XIE, R. ZHANG, AND H. BRUNNER, *Collocation methods for general volterra functional integral equations with vanishing delays*, SIAM Journal on Scientific Computing, 33 (2011), pp. 3303–3332.

[129] H. XU AND H. ZHA, *A dirichlet mixture model of hawkes processes for event sequence clustering*, Advances in neural information processing systems, 30 (2017).

[130] Q. YAN, S. TANG, S. GABRIELE, AND J. WU, *Media coverage and hospital notifications: Correlation analysis and optimal media impact duration to manage a pandemic*, Journal of Theoretical Biology, 390 (2016), pp. 1–13.

[131] Q. YAN, Y. TANG, D. YAN, J. WANG, L. YANG, X. YANG, AND S. TANG, *Impact of media reports on the early spread of COVID-19 epidemic*, Journal of Theoretical Biology, 502 (2020), p. 110385.

[132] G. K. YOUNG, *How much is too much: the difficulties of social media content moderation*, Information & Communications Technology Law, 31 (2022), pp. 1–16.

[133] A. ZADEH AND R. SHARDA, *How can our tweets go viral? point-process modelling of brand content*, Information & Management, 59 (2022), p. 103594.

[134] A. ZAREZADE, A. KHODADADI, M. FARAJTABAR, H. R. RABIEE, AND H. ZHA, *Correlated cascades: Compete or cooperate*, AAAI 2017, (2017), pp. 238–244.

[135] R. ZHANG, C. WALDER, AND M.-A. RIZOIU, *Variational Inference for Sparse Gaussian Process Modulated Hawkes Process*, Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020).

[136] 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 SIGKDD 2015, 2015, pp. 1513–1522.

[137] S. ZUO, H. JIANG, Z. LI, T. ZHAO, AND H. ZHA, *Transformer hawkes process*, in International conference on machine learning, PMLR, 2020, pp. 11692–11702.