Anytime Information Cascade Popularity Prediction via Self-Exciting Processes

Xi Zhang [1] **Akshay Aravamudan** [1] **Georgios C. Anagnostopoulos** [1]

Abstract

One important aspect of understanding behaviors of information cascades is to be able to accurately predict their popularity, that is, their message counts at any future time. Self-exciting Hawkes processes have been widely adopted for such tasks due to their success in describing cascading behaviors. In this paper, for general, marked Hawkes point processes, we present closed-form expressions for the mean and variance of future event counts, conditioned on observed events. Furthermore, these expressions allow us to develop a predictive approach, namely, Cascade Anytime Size Prediction via self-Exciting Regression model (CASPER), which is specifically tailored to popularity prediction, unlike existing generative approaches – based on point processes – for the same task. We showcase CASPER's merits via experiments entailing both synthetic and real-world data, and demonstrate that it considerably improves upon prior works in terms of accuracy, especially for early-stage prediction.

1. Introduction

Information cascades form, when people influence each others' perceptions and behaviors while partaking in a communication network. A prominent setting, where such influencing occurs, is social media, where online users propagate each others' contents and worldviews. Given an information cascade, whose initial progression has been observed, accurately predicting its total number of messages at any future time – a task commonly referred to as *anytime popularity prediction* – is particularly useful in applications such as marketing, spread of news and rumor control among other; for indicative examples, refer to (Yu et al., 2011; Wu et al., 2018; Singh & Chand, 2020; Gupta et al., 2020) respectively.

[1]Department of Computer Engineering & Sciences, Florida Institute of Technology, Melbourne, FL, USA. Correspondence to: Xi Zhang <zhang2012@my.fit.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

A large body of research has been devoted to popularity prediction; such works are surveyed in (Gao et al., 2019; Zhou et al., 2021). Generally speaking, relevant models and approaches can be categorized into three groups: feature-based models, deep learning models, and generative models. Feature-based models, as the ones proposed by (Szabo & Huberman, 2010; Bao et al., 2013; Yuan et al., 2016), extract various types of hand-crafted features, which are then utilized in conjunction with various techniques to make predictions. While such models can be quite explainable, their performance heavily depends on the quality of these feature sets, whose extraction is often laborious and requires domain expertise. On the other hand, deep learning models, like the ones put forward by (Li et al., 2017; Liao et al., 2019; Chen et al., 2019; Xu et al., 2021), aspire to learn effective features, but are often opaque to interpretation and/or explanation.

Moreover, both feature-based and deep learning approaches almost always require an abundance of data and computing effort to support model training and hyper-parameter tuning in order to achieve satisfactory prediction performances. For *anytime prediction*, *i.e.*, accurately predicting for any combination of *observation duration* $t_c \geq 0$ and *forecast horizon* $\Delta t \geq 0$, such approaches require training a distinct model for every desired $(t_c, \Delta t)$ pair, which exacerbates their computational burden even further.

Generative models, on the other hand, instead of directly tackling the prediction task, focus first on estimating cascade densities in order to characterize and describe cascade dynamics in the continuous temporal domain. Among them, models based on Hawkes (self-exciting) point processes, like the ones proposed in (Zhao et al., 2015; Kobayashi & Lambiotte, 2016; Tan & Chen, 2021), are particular advantageous since they exhibit emergent "rich-get-richer" phenomena, which explain well the heavy-tailed distributions of cascade sizes in empirical data.

In general, such models start by specifying a parameterized conditional intensity function, which attempts to capture the underlying characteristics of the observed information diffusion. For example, Chen & Tan (2018) choose a power-law function to model "aging" effect of influences and Kobayashi & Lambiotte (2016) use a sinusoidal function to model the circadian rhythm of Twitter users. Then, a

unique set of parameter values is learned for each cascade in order to describe its own self-exciting dynamics and predict its future evolution. In terms of parameters learning, in works like (Chen & Tan, 2018), all the cascade-specific parameters are inferred solely from the cascade's own past. In other works, like (Zhao et al., 2015; Kobayashi & Lambiotte, 2016), a few selected parameters are assumed to be shared across all cascade-generating processes, which are learned through fully-observed cascades in the train set, while the remaining parameters are learned from each cascade's observed history. In both cases, the parameters are learned via maximum likelihood estimation by employing the process' realization intensities. Finally, the conditional (on the process' observed *history* $\mathcal{H}_{t_c}$) mean count $\mathbb{E}\{N(t|\mathcal{H}_{t_c})\}$ – as derived from the fitted process – is typically employed to predict a cascade's popularity at time t.

Hawkes process based generative models are inherently interpretable and do not require intensive feature engineering. Moreover, their training is light-weight, as they normally feature much fewer parameters than the other two approaches. Moreover, for anytime popularity prediction, these models only need to be trained once for each observation duration t_c. This is particularly advantageous given that the other two approaches require training for each $(t_c, \Delta t)$ pair.

Nevertheless, models based on Hawkes processes suffer from a couple of important drawbacks. First, they are often criticized for their lackluster prediction performances, as discussed in (Mishra et al., 2016; Cao et al., 2017; Chen et al., 2019). This drawback can be traced, in essence, to the fact that these models are employed for a predictive task, while being generative: instead of optimizing prediction accuracy directly, they rely on likelihood maximization for parameter estimation and, subsequently, use the estimated parameters to produce point forecasts. The latter estimation approach tends to yield poorer predictions, especially when training data is limited.

Secondly, while the conditional mean count has been the de facto point estimate of future cascade popularity, there is no computationally-useful expression of it for general marked Hawkes point process. That being the case, to make predictions, existing works employing such processes are forced to either rely on expensive simulations, as in (Xiao et al., 2016; Ling et al., 2020), or rely on model-specific estimation algorithms, such as the ones proposed in (Wang et al., 2017; Kobayashi & Lambiotte, 2016; Chen & Tan, 2018). This lack of a computationally-useful expression for the conditional mean count limits the potential of Hawkes process-based approaches in the context of popularity prediction.

The contributions of this work are two-fold. First, in our main result, Theorem 4.7, we derive closed-form expressions for the conditional (on the observed history $\mathcal{H}_{t_c}$) mean and variance at $t \geq t_c$ of the counting process $N(t)$, which is associated to an Marked Hawkes Point Processs (MHPPs) with arbitrary, Lebesgue-integrable conditional intensity function and unpredictable marks. Compared to prior works, which also concern themselves with such moments, our main result does not assume special forms of the intensity functions involved and it is in closed-form as opposed to being computed by numerically solving one or more differential equations, which may be intractable.

Our novel results are eventually reached by first viewing an MHPP's counting process $N(t)$ as an equivalent branching process and, thus, determining – in Theorem 4.2 – the probability generating function (PGF) of $N_k(t)$ for each generation $k \geq 0$. Such generation-wise results aid in furthering our understanding of MHPPs' nature and evolution over time and allowed us to determine closed form expressions for $N(t)$'s mean and variance in Theorem 4.5. Our final – and main – result is shown by interpreting $N(t|\mathcal{H}_{t_c})$ in Lemma 4.6 as a sum of already-observed event counts plus an unconditional count process of suitable intensity. Finally, let us mention that we confirm our theoretical findings through extensive simulations.

The second contribution of this work is a new approach to anytime popularity prediction, which we dubbed Cascade Anytime Size Prediction via self-Exciting Regression (CASPER). Its proposal is motivated by the aforementioned challenges faced by existing, state-of-the-art, Hawkes process-based approaches to popularity prediction. Unlike these generative approaches, the proposed predictive framework fits an MHPP-based model by directly minimizing the empirical mean squared prediction error between future counts and predicted counts. For producing the latter, CASPER leverages our new theoretical results and predicts future counts of individual cascade given its observed past via our new conditional mean count expression. Furthermore, costly simulations or approximations are avoided by using this expression for prediction.

The merits of CASPER's predictive nature are first demonstrated via experiments using synthetic data, which compare the results obtained via generative versus predictive training (according to CASPER) in terms of prediction performance. Additional experimental results on real-world Twitter data show that CASPER considerably improves upon established MHPP based models in forecasting accuracy, especially for early-stage prediction, when the amount of already-observed events is meager.

The rest of the paper is organized as follows: Section 2 surveys related works, while Section 3 introduces the marked Hawkes process and the assumptions made throughout this work. Section 4 presents our main theoretical results, followed by Section 5, where CASPER is described. Section 6

presents experimental results for both synthetic and real-world (Twitter) data. Finally, we briefly discuss our work and future works in Section 7.

2. Related Works

For cascade popularity prediction tasks, we have already outlined different approaches (feature-based, deep learning, and generative models) and prediction procedures with existing (marked) Hawkes point process based models. In this section, we briefly describe several notable works in finding the counting distribution/moments of (marked) Hawkes point processes.

A temporal Hawkes point process $N(t)$ is called *stationary*, if the distribution of $N(t, t+\tau]$ only depends on the interval's length $\tau \geq 0$, which implies a conditional intensity with constant base function, *i.e.*, $b(t) \triangleq b \geq 0, \forall t \geq 0$. Assuming stationarity, Hawkes (1971a;b) have provided the second order counting properties of the process and its multivariate equivalent by finding their mean and covariance density through point spectra analysis. Bacry & Muzy (2016) extend these early works to the marked case, and provide a non-parametric kernel estimation method that relies on the previously found second order statistical properties. In the same vein, Hawkes & Oakes (1974) demonstrate the existence of an equivalent branching (cluster) process representation of the process and go on to provide integral equations for $N(t)$'s PGF. However, the calculation of the PGF is intractable, and, hence, only asymptotic results were entertained. Following this work, Oakes (1975) finds explicit equations for the mean and the variance of the counting distribution as a function of $t > 0$ for the special case of exponential excitation functions.

For another special case of the Hawkes process, where both the base and excitation function can be expressed as a mixture of exponential terms, Errais et al. (2010) and Dassios & Zhao (2011) find the second-order moments of the intensity via infinitesimal generators that are then used to solve differential equations. Haimovich et al. (2020) find the conditional (on observed events) mean and variance of such a process and adopt the conditional mean as a point predictor for future event counts over arbitrary times.

Recently, Cui et al. (2020) derived the moments of the counting distribution for the most general form of Hawkes processes, which employ arbitrary base and excitation intensities. They accomplished this via an elementary approach by considering infinitesimal generators and by leveraging Dynkin's formula. A benefit of this approach is that it can also be applied to the marked process case. However, unlike our work, they do not consider the conditional case; furthermore, they lack closed-form expressions for the resulting moments. Related to our work, O'Brien et al. (2020) derive the counts' PGF for such self-exciting processes and find their conditional mean count to use as a point predictor for future counts. However, unlike us, they have to resort to solving the differential equations governing the equivalent branching process representation in lieu of closed-form expressions. Additionally, they do not consider the marked case, which is widely used in cascade modeling (Zhao et al., 2015; Mishra et al., 2016; Chen & Tan, 2018).

3. Marked Hawkes Point Process

A *marked temporal point process* extends the ordinary temporal point process by associating each event time with a stochastic *mark* taking values in the *mark space* $\mathcal{M}$. Such a process can also be regarded as a point process on the product space $(\mathbb{R}^+ \times \mathcal{M})$ with conditional intensity $\lambda^*(t, m)$, where $\mathbb{R}^+$ denotes the set of non-negative real numbers[1]. Its marginal temporal process is called *ground process* with *ground intensity* $\lambda^*(t)$. In this paper, we consider a Marked Hawkes Point Process (MHPP) $N(t)$ with ground intensity

$$\lambda^*(t) \triangleq b(t) + \sum_{i:t_i<t} \phi_{m_i}(t - t_i) \tag{1}$$

where $b(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+$ is its *base intensity* and $\phi_m(\cdot) : \mathbb{R}^+ \times \mathcal{M} \to \mathbb{R}^+$ is the mark-indexed *excitation function*. The former models exogenous (to the process) influences, while the latter models self-excitation effects. Here, the pairs (t_i, m_i) reflect the process' event times and associated marks observed up to time t. Throughout this paper, we will make the following two assumptions:

A.1 Both the base intensity function $b(\cdot)$ and the family of excitation functions $\{\phi_m(\cdot)\}_{m\in\mathcal{M}}$ are Lebesgue-integrable and non-zero almost everywhere on $\mathbb{R}^+$. If we define $\eta \triangleq \int_{s=0}^{\infty} b(s)ds$ and $\gamma_m \triangleq \int_{s=0}^{\infty} \phi_m(s)ds$ for all $m \in \mathcal{M}$, then this implies that $\eta, \gamma_m \in (0, \infty)$ for all $m \in \mathcal{M}$.

A.2 Any MHPP we will consider is assumed to feature unpredictable marks following mark distribution $g(m)$, *i.e.*, the mark distribution is independent of past event times and marks.

By virtue of Assumption A.2, the conditional intensity of such MHPPs takes the form:

$$\begin{aligned}\lambda^*(t, m) =& \lambda^*(t)g(m) = \\ &= \left(b(t) + \sum_{i:t_i<t} \phi_{m_i}(t - t_i)\right) g(m)\end{aligned} \tag{2}$$

[1] Without loss of generality, we will assume that such processes start at time $t = 0$. Results for an arbitrary starting time t_0 can be easily obtained by applying a mere time shift.

First noted by Hawkes & Oakes (1974), any Hawkes process can be equivalently viewed as a *branching process*. Under this prism, the mechanism for generating events from an MHPP with conditional intensity specified by Eq. (2) proceeds as follows: **(a)** 0^{th}-generation (*immigrant*) events have event times drawn from a Poisson process of intensity $b(\cdot)$; and, **(b)** for $k \geq 0$, each k^{th}-generation event (t^k, m^k) will independently give birth to a number of offspring events that belong to the $(k+1)^{\text{th}}$ generation and whose event times are sampled from a Poisson process with intensity $\phi_{m^k}(t - t^k)$. All aforementioned events bare unpredictable marks.

4. MHPP Count Moments

By viewing an MHPP $N(t)$ as an equivalent branching process, we have that $N(t) = \sum_{k\geq 0} N_k(t)$, where $N_k(t)$ is the number of k^{th}-generation events up to time t. In this section, we first derive the probability generating function (PGF) of $N_k(t)$ in Section 4.1. Subsequently, the first two moments of $N(t)$ are identified in Section 4.2. Due to space limitations, the proofs of our results are provided in Appendix A.

In what follows, $\delta(\cdot)$ will stand for the Dirac delta distribution located at $t = 0$ and $u(\cdot)$ for the Heavyside step function, which equals 1 for non-negative argument values and equals 0 otherwise. Note that all our results pertain to MHPPs with conditional intensity given by Eq. (2) and obeying Assumption A.1 and Assumption A.2.

4.1. k^{th}-Generation Event Count Distribution

For $k \geq 0$, let $Z_k \triangleq N_k(\infty)$ indicate the total count of k^{th}-generation events on $\mathbb{R}^+$. To find the distribution of $N_k(t)$, we will start by finding the probability density function (PDF) of the event time occurrence in k-th generation.

Proposition 4.1 (PDF of k^{th}-generation occurrence times)**.** *For $k \geq 0$ and given the total count Z_k of k^{th}-generation events, the occurrence times of those Z_k events are i.i.d. with PDF*

$$f_k(t) \triangleq \frac{1}{\eta}(b * \xi^{*k})(t), \quad t \geq 0 \tag{3}$$

where $\xi(t) \triangleq \mathbb{E}_m\left\{\frac{\phi_m(t)}{\gamma_m}\right\}$ and ξ^{*k} is the k-fold convolution of ξ with itself, with the convention that $\xi^{*0} \triangleq \delta$.

Above, the expectation $\mathbb{E}_m\{\cdot\}$ with respect to the mark m is computed by utilizing the mark's PDF $g(\cdot)$, if the mark is continuous, or using the mark's probability mass function (PMF) $g(\cdot)$, if the mark is discrete. Eq. (3) reveals that k^{th}-generation event times are given as the sum of k i.i.d. random variables (RVs) with PDF $\xi(\cdot)$ plus an independent RV with PDF $\frac{1}{\eta}b(\cdot)$. For example, if $\frac{1}{\eta}b(t) = \xi(t) = \beta \mathrm{e}^{-\beta t}$, then, given the number Z_k of k^{th}-generation events, their event times are i.i.d. Erlang-distributed with parameters $(k+1, \beta)$.

This Proposition allows us now to identify the PGF of $N_k(t)$ at an arbitrary time $t \geq 0$, as presented in the following Theorem.

Theorem 4.2 (PGF of $N_k(t)$)**.** *For $k \geq 1$, the PGF of $N_k(t)$ is given as*

$$G_{N_k(t)}(w) = G_0\big(G^{\circ k}(G_k(w))\big) \tag{4}$$

where $G_0(w) \triangleq \mathrm{e}^{\eta(w-1)}$ is the PGF of a Poisson-distributed RV with parameter η and $G^{\circ k}(w)$ is the k-fold composition of $G(w) \triangleq \mathbb{E}_m\{e^{\gamma_m(w-1)}\}$ with itself, where $e^{\gamma_m(w-1)}$ is the PGF of a Poisson-distributed RV with parameter γ_m. Finally, $G_k(w) \triangleq 1 + F_k(t)(w-1)$ is the PGF of a Bernoulli-distributed RV with parameter $F_k(t)$, where $F_k(t) \triangleq \frac{1}{\eta}\left(u * b * \xi^{*k}\right)(t)$ is the cumulative distribution function (CDF) of the k^{th}-generation event occurrence times.

This Theorem provides a complete view of the distribution of k^{th}-generation event counts at any time t. The mean and variance of $N_k(t)$ can now be found by relying on well-known relations between an RV's PGF and its moments. The results are provided in the following Corollary.

Corollary 4.3 (Mean & Variance of $N_k(t)$)**.** *Define $\gamma \triangleq \mathbb{E}_m\{\gamma_m\}$ and $\nu \triangleq \mathbb{E}_m\{\gamma_m^2\}$. Let $\zeta(\cdot) \triangleq \gamma\xi(\cdot)$, where $\xi(\cdot)$ is defined as in Proposition 4.1. For $k \geq 1$, we have*

$$\mathbb{E}\{N_k(t)\} = \left(u * b * \zeta^{*k}\right)(t) \tag{5}$$

$$\mathrm{Var}(N_k(t)) = \mathbb{E}\{N_k(t)\} + (\mathbb{E}\{N_k(t)\})^2 \left(\frac{\nu}{\eta\gamma^2}\right) \sum_{j=0}^{k-1} \frac{1}{\gamma^j} \tag{6}$$

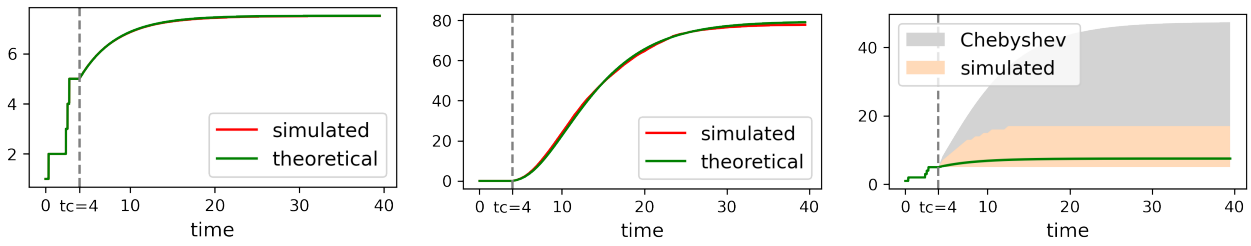
4.2. Process Event Count Moments

It turns out that, since $N(t) = \sum_{k\geq 0} N_k(t)$, in order to derive $\mathrm{Var}(N(t))$, one must first determine the covariance between different $N_k(t)$'s. This is what is done next.

Proposition 4.4. *For $p, q \geq 0$, the covariance between $N_p(t)$ and $N_q(t)$ is given as*

$$\mathrm{Cov}(N_p(t), N_q(t)) = \frac{\mathbb{E}\{N_{\max(p,q)}(t)\}}{\mathbb{E}\{N_{\min(p,q)}(t)\}} \mathrm{Var}\big(N_{\min(p,q)}(t)\big) \tag{7}$$

Now, equipped with the outcomes of Corollary 4.3 and Proposition 4.4, the mean and variance of $N(t)$ can be easily determined. The exact expressions are provided in our next result and play an important role in examining and understanding the behavior of MHPPs in terms of event counts over time (via the mean) and the associated uncertainty about these counts over time (via the variance).



(a) conditional mean (b) conditional variance (c) Chebyshev bounds

Figure 1. Comparison of theoretical versus simulation-stemming results for an MHPP, whose conditional intensity takes the form of Eq. (15) with $\alpha = 0.87$, $\beta = 1.27$ and $\kappa = 0.46$, while the marks are uniformly-distributed with support $[1, 2]$. A censoring time $t_c = 4$ is used resulting in 5 observed events $\mathcal{H}_{t_c} = \{(0.0, 1.6), (0.36, 1.52), (2.40, 1.66), (2.60, 1.23), (2.81, 1.76)\}$. The theoretical conditional mean and variance (green curves) are found based on Theorem 4.7, while the simulation results (red curves) are obtained empirically via $50,000$ cascades generated via Ogata's thinning algorithm. The Chebyshev-based 95% bounds (grey) are computed based on Eq. (14), while the simulated bounds (yellow) are based on the empirical 95% quantiles of the predicted counts.

Theorem 4.5 (Unconditional Mean & Variance of MHPP Event Counts)**.** *The mean and variance of $N(t)$ follows,*

$$\mathbb{E}\{N(t)\} = \sum_{k\geq 0} \mathbb{E}\{N_k(t)\} \tag{8}$$

$$\mathrm{Var}(N(t)) = \sum_{k\geq 0} \left(1 + \frac{2}{\mathbb{E}\{N_k(t)\}} \sum_{j=k+1}^{\infty} \mathbb{E}\{N_j(t)\}\right) \cdot \cdot \mathrm{Var}(N_k(t)) \tag{9}$$

where $\mathbb{E}\{N_k(t)\}$ and $\mathrm{Var}(N_k(t))$ are given by Eq. (5) and Eq. (6) of Corollary 4.3.

The results above provide the moments of MHPP with no observed history (unconditional moments), *i.e.*, $t_c = 0$. Hence, they cannot be directly leveraged to tackle real-world cascades popularity prediction problems. For this purpose, one would need to utilize moments that are conditioned on the cascades' past history instead ($t_c > 0$). The final, upcoming result of this work provides precisely such conditional means and variances and hinges on the next lemma.

Lemma 4.6. *Consider an MHPP $N(t)$ with conditional intensity as given in Eq. (2). Assume that we have observed the process' history $\mathcal{H}_{t_c} = \{(t_i, m_i)\}_{i:t_i\leq t_c}$ up to some censoring time $t_c > 0$ and that it consists of $N(t_c) \geq 0$ events. Then, the conditional count $N(t|\mathcal{H}_{t_c})$ for $t \geq t_c$ can be decomposed as $N(t|\mathcal{H}_{t_c}) = N(t_c) + \hat{N}(\Delta t)$, where $\Delta t \triangleq t - t_c \geq 0$ and $\hat{N}(\Delta t)$ is the count of a new MHPP, which (i) starts at time t_c ($\Delta t = 0$) with $\hat{N}(t_c) = 0$, (ii) features the same excitation function $\phi_m(\cdot)$ and mark distribution $g(\cdot)$ as the original process $N(t)$, and (iii) has a base intensity $\hat{b}(\cdot)$ given as*

$$\hat{b}(\Delta t) \triangleq b(\Delta t + t_c) + \sum_{(t_i,m_i)\in\mathcal{H}_{t_c}} \phi_{m_i}(\Delta t + t_c - t_i) \tag{10}$$

This Lemma, while seemingly straight-forward, to the best of our acknowledge, has not been adopted by existing works for finding conditional moments. Works like (O'Brien et al., 2020; Haimovich et al., 2020) could have used this Lemma to directly determine the conditional count moments for MHPPs, instead of following a long and unnecessary detour.

Theorem 4.7 (Conditional Mean & Variance of MHPP Event Counts)**.** *Consider the setting of Lemma 4.6, and let $\Delta t = t - t_c$. Then, for $t \geq t_c$, $\mathbb{E}\{N(t|\mathcal{H}_{t_c})\} = N(t_c) + \mathbb{E}\left\{\hat{N}(\Delta t)\right\}$ and $\mathrm{Var}(N(t|\mathcal{H}_{t_c})) = \mathrm{Var}\left(\hat{N}(\Delta t)\right)$.*

Note that $\mathbb{E}\left\{\hat{N}(\Delta t)\right\}$ and $\mathrm{Var}\left(\hat{N}(\Delta t)\right)$ can be computed based on the results of Theorem 4.5, since $\hat{N}(\Delta t)$ is an unconditional MHPP.

Figure 1 demonstrates that the expressions we derived for the conditional mean count and its associated variance in Theorem 4.7 strongly match the results obtained via time-consuming simulations.

Finally, let us remark that, even though all of our theoretical results presented in here pertain to the marked case, the corresponding results for unmarked Hawkes processes stem as special cases by using a Dirac delta distribution located at a fixed mark value in place of $g(\cdot)$.

5. CASPER

Motivated by the challenges faced by existing Hawkes process based generative models and by taking advantage of the closed-form expressions presented in Theorem 4.7, we propose Cascade Anytime Size Prediction via self-Exciting Regression model (CASPER), a MHPP-based predictive model for anytime popularity prediction. A notable characteristic of this model is that it can provide useful count predictions of an unfolding cascade based on its observed history without the need of assuming access to additional, fully-observed cascades generated by the same process.

Formulation. Consider a cascade that we have observed up until time t_c, during which $N(t_c)$ events have occurred, *i.e.*, $\mathcal{H}_{t_c} = \{(t_i, m_i)\}_{i=1,\dots,N(t_c)}$. To predict its future *size* (total count) at time $t > t_c$, we start by assuming that its underlying diffusion process is an MHPP with ground intensity function $\lambda^*(\cdot;\boldsymbol{\theta})$ and with unpredictable mark distribution $g(\cdot)$. The intensity function's parameters $\boldsymbol{\theta}$ constitute the model's parameters and their optimal values are determined via training as discussed next. The mark distribution $g(\cdot)$ is assumed to be known or empirically estimated from the observed mark values $\{m_i\}_{i=1,\dots,N(t_c)}$.

Learning. Let $N(t;\boldsymbol{\theta})$ denote the counting process in question and consider the event times $t_i < t_j \leq t_c$. Then, $\mathbb{E}\{N(t_j|\mathcal{H}_{t_i};\boldsymbol{\theta})\}$, the average event count at time t_j conditioned on $\mathcal{H}_{t_i}$ is the quantity generally adopted as the predicted popularity at time t_j given observations up to time t_i. Given that there are j events by time t_j, we define the squared loss $\ell_{ij}(\boldsymbol{\theta})$ for observations up to time t_i as

$$\ell_{ij}(\boldsymbol{\theta}) \triangleq \left(\mathbb{E}\{N(t_j|\mathcal{H}_{t_i};\boldsymbol{\theta})\} - j\right)^2 \tag{11}$$

By evaluating such loss terms for each ordered pair of observed event times (t_i, t_j), CASPER's overall loss function takes the form

$$L(\boldsymbol{\theta}|\mathcal{H}_{t_c}) \triangleq \frac{1}{|\mathcal{S}(t_c)|} \sum\sum_{(i,j)\in\mathcal{S}(t_c)} \ell_{ij}(\boldsymbol{\theta}) \tag{12}$$

where $\mathcal{S}(t_c) \triangleq \{(i,j) : 0 < t_i < t_j \leq t_c\}$ and $|\mathcal{S}(t_c)|$ is its cardinality. Note that, $\mathcal{S}(t_c)$ consists of $n(n-1)/2$ terms for $n \triangleq N(t_c)$ observed events.

The optimal model parameters $\boldsymbol{\theta}^*$ are thus found by minimizing the above overall loss function, that is, $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathcal{H}_{t_c})$, where $\boldsymbol{\Theta}$ is the feasible parameter set.

Prediction. Finally, a cascade's size at time $t > t_c$, given observations up to time t_c, is predicted as

$$N_{\text{pred}}(t) \triangleq \mathbb{E}\{N(t|\mathcal{H}_{t_c};\boldsymbol{\theta}^*)\} \tag{13}$$

If $N_{\text{true}}(t)$ is the true cascade size at time t, then the Chebyshev-based $100(1-\alpha)\%$ prediction intervals can be established as

$$|N_{\text{true}}(t) - N_{\text{pred}}(t)| < \sqrt{\frac{\text{Var}(N(t|\mathcal{H}_{t_c};\boldsymbol{\theta}^*))}{\alpha}} \tag{14}$$

These bounds are unsurprisingly very loose (as also witnessed in Figure 1) and, hence, are indicative, but of limited practical use. Finding tighter bounds could be a subject of future work.

Runtime Considerations. For n observed events, the loss function in Eq. (12) consists of $\mathcal{O}(n^2)$ terms for large n. When n is large, only a subset of them can be used to speed up training, without necessarily sacrificing prediction accuracy. Also, runtimes typically depend on the particular intensity functions employed. For reference, regarding the tweet prediction setup described in Section 6.1, CASPER's training[2] takes about 0.1207 seconds per (t_i, t_j) pair on a Windows 10 machine with an Intel® Core™ i7 $-$ 4720HQ CPU 2.60GHz processor and 16.0 GB of RAM.

6. Tweet Popularity Prediction

In this section, we employ CASPER for tweet popularity prediction tasks and report its prediction performances on both synthetic and real-world Twitter data.

6.1. CASPER Specifics

Here, we briefly describe the specific setting for CASPER that we used in order to conduct all of our experiments. For $i \geq 0$, let t_i be the i^{th} retweet time and let the associated mark m_i reflect the number of followers that the retweet's user has. Then, we choose the following conditional intensity for our modeling purposes:

$$\lambda^*(t, m; \alpha, \beta, \kappa) = \left(\alpha \sum_{i:t_i<t} m_i^\kappa \mathrm{e}^{-\beta(t-t_i)}\right) g(m) \tag{15}$$

where $\alpha > 0$ can be regarded as the "quality" of the tweet and $\beta > 0$ describes how fast a retweet's influence on other users fades away with time. The mark value m_i is interpreted as the strength of a user's influence. This strength is regulated by a parameter $0 < \kappa < 1$; the larger the value of κ, the more influence users with large number of followers exert on the rest of the social network. Moreover, the discrete mark distribution $g(m)$ is empirically estimated via a histogram of the observed mark values.

Using the results of Theorem 4.7, we obtain the following conditional mean count expression:

$$\mathbb{E}\{N(t_c + \Delta t|\mathcal{H}_{t_c})\} = N(t_c) + \\ + \begin{cases} \eta\beta\Delta t, & \gamma = 1 \\ \dfrac{\eta}{1-\gamma}\left(1 - \mathrm{e}^{-\beta(1-\gamma)\Delta t}\right), & \gamma \neq 1 \end{cases} \tag{16}$$

where $\eta \triangleq \frac{\alpha}{\beta}\sum_{t_i\leq t_c} m_i^\kappa \mathrm{e}^{-\beta(t_c-t_i)}$ and $\gamma \triangleq \frac{\alpha}{\beta}\sum_m m^\kappa g(m)$. Numerical details for finding this conditional mean and the associated conditional variance are provided in Appendix B.1.

The optimal parameter values $\boldsymbol{\theta}^* \triangleq \{\alpha, \beta, \kappa\}$ are found by minimizing the overall loss function of (12) using a projected gradient descent algorithm, which is detailed in Appendix B.2.

[2] Python 3.9.12 code for CASPER can be found at https://github.com/xizhang-cc/casper.

Table 1. The median and mean of prediction APEs% for different learning approaches on a synthetic dataset. Here, GT stands for the ground truth models, *i.e.*, the models used to generated the synthetic data, MSE represents models trained by minimizing the overall loss of Eq. (12) – CASPER's approach – and, finally, MLL stands for generative models trained by maximizing their likelihood.

	MEDIAN APEs%									MEAN APEs%								
Δt \ t_c	2			4			6			2			4			6		
	GT	MSE	MLL	GT	MSE	MLL	GT	MSE	MLL	GT	MSE	MLL	GT	MSE	MLL	GT	MSE	MLL
5	43.63	53.36	68.41	30.76	30.54	35.05	17.18	16.14	20.32	46.99	55.31	67.04	35.08	36.50	41.50	22.30	22.99	27.72
10	58.24	67.02	79.35	47.68	41.70	60.78	27.56	23.68	38.30	66.25	75.16	95.06	54.55	55.67	79.41	35.46	36.92	53.58
15	65.10	73.06	83.15	57.04	48.18	83.22	33.76	27.61	54.78	78.12	91.41	119.3	66.40	70.90	117.3	43.19	47.86	79.63
20	68.74	75.90	84.81	61.68	51.74	110.6	37.08	29.80	69.34	84.95	104.9	140.3	73.24	83.54	153.2	47.65	56.95	104.5

The learned γ parameter constitutes the estimated branching factor of the cascade. Note that a cascade will converge for $\gamma < 1$ and diverge otherwise. All cascades that we will consider will be of the former case.

6.2. Evaluation Metric

Following the prior works of Zhao et al. (2015); Mishra et al. (2016); Chen & Tan (2018), we use the Absolute Percentage Error (APE) as a form of prediction performance measure for each cascade. Specifically, for a given cascade w and a prediction time t, its APE metric is defined as,

$$\text{APE}^w(t) = \frac{\left|N^w_{\text{pred}}(t) - N^w_{\text{true}}(t)\right|}{N^w_{\text{true}}(t)} \tag{17}$$

Smaller APE values reflect better prediction performances.

6.3. Synthetic Data

To compare our predictive learning approach (optimized to minimize the loss function in (12)) vis-à-vis the generative learning approach (optimized to maximize the likelihood function), we generate a synthetic dataset with the same intensity function as the one we assumed for the retweeting process in Eq. (15) with two different setups for the parameter values, mark distribution, and the follower number m_0 of the user that tweets to initiate the cascade: **(i)** $\boldsymbol{\theta}_1 = \{\alpha = 0.018, \beta = 1.8, \kappa = 0.54\}$, with $g(\cdot)$ being a discrete uniform distribution over the set $\{1, \ldots, 10,000\}$ and $m_0 = 9,000$ followers; and **(ii)** $\boldsymbol{\theta}_2 = \{\alpha = 1.7, \beta = 2.4, \kappa = 0.24\}$, with $g(\cdot)$ being a continuous Pareto distribution with shape parameter $c = 1.016$ and $m_0 = 100$ followers. Finally, we used Ogata's thinning algorithm (Ogata, 1981) to generate $10,000$ cascades per setup.

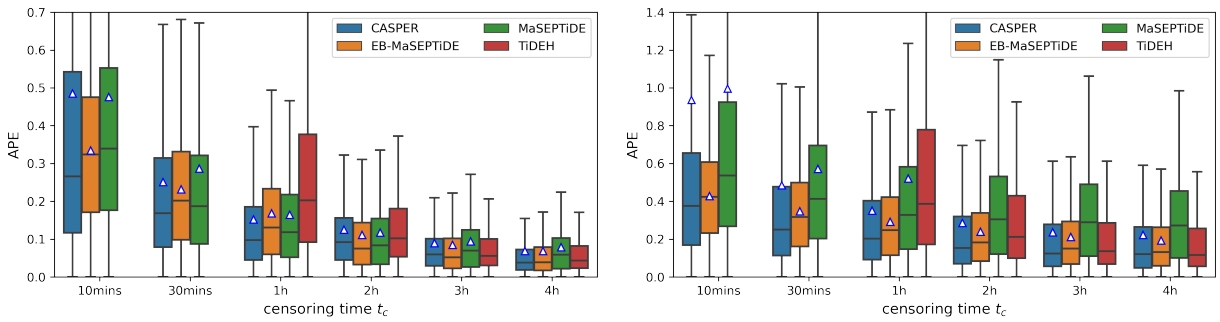
Table 1 presents the median and mean values of APEs% across $20,000$ cascades with varying $(t_c, \Delta t)$ for the two training approaches. For reference, the prediction performance of the ground truth models, *i.e.*, the models used to generate the cascades, is also reported. First of all, with consistently lower mean and median APE values, it is clear that CASPER's predictive learning approach (MSE) outperforms the generative learning approaches (MLL) in every scenario, that is, for short- and long-term predictions with either short or long censoring times. Secondly, CASPER's predictive performance (MSE) is highly competitive to the one of ground truth models (GT). It is worth noting that, except for the case when $t_c = 2$, for which the censoring time is very short (a challenging learning scenario), in both $t_c = 4$ and $t_c = 6$ cases, the CASPER-trained models report lower median APEs% values than the ground truth models.

6.4. Real-World Twitter Data

Here, we apply CASPER to real-world Twitter data for tweet popularity prediction and compare its prediction performance with existing point process based popularity models. Results from an additional comparison to CasFlow (Xu et al., 2021), a state-of-the-art deep learning approach, are presented in Appendix C.

6.4.1. SEISMIC DATA

Released by Zhao et al. (2015), SEISMIC is a widely adopted Twitter dataset for social media popularity prediction tasks (Mishra et al., 2016; Chen & Tan, 2018; Tan & Chen, 2021). It contains $166,076$ tweets, all of which have at least 50 retweets. For each tweet and corresponding retweets, all their posting times and relevant users' follower numbers are included. Following the setup of Zhao et al. (2015), we split the data into a training set with $71,815$ tweets and a test set with $94,254$ tweets. We then randomly select $40,000$ cascades from the test set to conduct our experiments. Notice that, unlike TiDeH and EB-MaSEPTide, CASPER does not need to utilize multiple cascades to learn to predict the future event count of a given cascade.



(a) Short-term prediction with $\Delta t = 4$ hours

(b) Long-term prediction with $\Delta t = 4$ days

Figure 2. Boxplots of short- and long-term prediction APE values between our and comparison models on SEISMIC data, with various censoring times. The Horizontal bars within the boxes indicate the median values, and the white triangles indicate the mean values. Some triangle indicated means are omitted as their values exceeds the y-axis limits in the plots

6.4.2. Comparison Algorithms

We consider three state-of-the-arts point process based models for comparison.

TiDeH (Kobayashi & Lambiotte, 2016) models the retweeting processes as Hawkes processes with consideration of the circadian nature of users activities. In model fitting, three shape parameters are optimized using an extra training set consisting of 100 fully observed cascades. The infectious rate of the target cascade is then estimated by the maximum likelihood method based on the observed events. A cascade's future popularity is calculated from the estimated retweet rate, which is evaluated by numerically solving TiDeH's self-consistent equation. As discussed in the TiDeH paper, this model works best with long observation periods, when there exist enough observed events (at least 2000) for model training.

MaSEPTiDE (Chen & Tan, 2018) is a Hawkes process based prediction model with time-varying background intensity, whose model parameters are learned generatively by maximizing the likelihood of the observed events. This model is the most similar to ours in term of amount of information used for model training, as it too does not need additional fully-observed cascades for model training. To predict, the future conditional mean intensity is estimated by numerically solving an integral equation and using a flexible parametric function.

EB-MaSEPTiDE (Tan & Chen, 2021) extends MaSEPTiDE by adopting an empirical Bayes approach to estimate MaSEPTiDE's parameters based on the observed events of the targeted tweet, but also on an extra set of fully-observed cascades. The fitted models were found to perform better in popularity prediction tasks, especially when making early-stage predictions (when only few events have been observed).

6.4.3. Prediction Performance

The prediction performances of CASPER and the comparison models on the SEISMIC data are reported in Figure 2. The APEs across cascades for short-term prediction ($\Delta t = 4$ hours) for various censoring times are presented in Figure 2a, while for long-term prediction ($\Delta t = 4$ days) are presented in Figure 2b. Overall, it is clear that CASPER consistently exhibits competitive, if not the best performance among all models.

A closer look at the comparisons between our model and MaSEPTiDE, the most similar approach to ours in terms of training information amount being used, shows that our model clearly outperform MaSEPTiDE across all scenarios.

Now, let's examine prediction performance per scenario, with special focus on comparisons with TiDeH in cases with relatively long observation periods ($t_c = 2\text{h}, 3\text{h}, 4\text{h}$) and on comparisons with EB-MaSEPTiDE in cases with shorter observation periods ($t_c = 10\text{mins}, 30\text{mins}, 1\text{h}$).

First, short-term prediction with long observation period is a much easier task compared to other scenarios, and as expected, all models report similar prediction performances as shown in the right side of Figure 2a. Secondly, regarding long-term prediction with long observation period, the right side of Figure 2b indicates that the performances of CASPER and EB-MaSEPTiDE are quite similar: CASPER exhibits a slightly lower median and EB-MaSEPTiDE shows a slightly lower mean. Rather surprisingly, TiDeH exhibits large mean APE values, although it has the lowest median value for $t_c = 4$ hours.

Finally, let us compare model performance for early stage predictions – the most interesting and, yet, the most challenging task in real-world settings. As very scarce information is available in terms of the observed history of the target cascade, EB-MaSEPTiDE resorts to using additional

fully-observed cascades for better model training. As shown on the left side of both Figure 2a and Figure 2b, compared to EB-MaSEPTiDE, CASPER consistently reports higher mean APE values, but lower median APE values for both short- and long-term predictions. This indicates that EB-MaSEPTiDE is more stable at providing relatively good estimations, while CASPER, although having a larger variance of prediction accuracy across cascades, offers better predictions for the majority of the cascades under consideration.

7. Conclusion

In this paper, by viewing a Hawkes process as an equivalent branching process, we present novel theoretical results that culminate in closed-form expressions for the conditional mean and variance for counting processes of general MHPPs. Secondly, by leveraging these results and motivated by the limitations of current Hawkes process based generative approaches, we introduced CASPER, a MHPP based predictive model for anytime popularity prediction. We showcased experiment results on synthetic data to demonstrate forecasting improvements gained by such a predictive approach. Moreover, experimental results with real-world (Twitter) data showcase that CASPER appreciably improves upon prior works in terms of prediction accuracy, especially for early-stage prediction.

Possible directions of extending this work include finding tighter prediction bounds and employing other commonly-used excitation functions aside from the exponential function. In this work, we used the latter excitation function due to its mathematical simplicity and the fact that it allows for a simple expression of the conditional count moments. Foreseeably, one could, derive workable expressions for other excitation functions by taking advantage of prior work on sums of independent random variables, such as the one of Nadarajah (2008).

Acknowledgements

This work was supported in part by the U.S. Defense Advanced Research Projects Agency (DARPA) Grant No. FA8650-18-C-7823 under the *Computational Simulation of Online Social Behavior* (*SocialSim*) program of DARPA's Information Innovation Office. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, or the U.S. Government.

References

Bacry, E. and Muzy, J. First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62 (4):2184–2202, 2016. doi: 10.1109/TIT.2016.2533397.

Bao, P., Shen, H.-W., Huang, J., and Cheng, X.-Q. Popularity prediction in microblogging network: A case study on sina weibo. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pp. 177–178, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320382. doi: 10.1145/2487788.2487877.

Cao, Q., Shen, H., Cen, K., Ouyang, W., and Cheng, X. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pp. 1149–1158, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4918-5. doi: 10.1145/3132847.3132973.

Chen, F. and Tan, W. H. Marked self-exciting point process modelling of information diffusion on twitter. *The Annals of Applied Statistics*, 12(4):2175 – 2196, 2018. doi: 10.1214/18-AOAS1148.

Chen, X., Zhou, F., Zhang, K., Trajcevski, G., Zhong, T., and Zhang, F. Information diffusion prediction via recurrent cascades convolution. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 770–781, 2019. doi: 10.1109/ICDE.2019.00074.

Cui, L., Hawkes, A., and Yi, H. An elementary derivation of moments of hawkes processes. *Advances in Applied Probability*, 52(1):102–137, 2020. doi: 10.1017/apr.2019.53.

Dassios, A. and Zhao, H. A dynamic contagion process. *Advances in Applied Probability*, 43(3):814–846, 2011. doi: 10.1239/aap/1316792671.

Errais, E., Giesecke, K., and Goldberg, L. R. Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1(1):642–665, 2010. doi: 10.1137/090771272. URL https://doi.org/10.1137/090771272.

Gao, X., Cao, Z., Li, S., Yao, B., Chen, G., and Tang, S. Taxonomy and evaluation for microblog popularity prediction. *ACM Trans. Knowl. Discov. Data*, 13(2), March 2019. ISSN 1556-4681. doi: 10.1145/3301303.

Gupta, V., Jung, K., and Yoo, S.-C. Exploring the power of multimodal features for predicting the popularity of social media image in a tourist destination. *Multimodal Technologies and Interaction*, 4(3), 2020. ISSN 2414-4088. doi: 10.3390/mti4030064. URL https://www.mdpi.com/2414-4088/4/3/64.

Haimovich, D., Karamshuk, D., Leeper, T. J., Riabenko, E., and Vojnovic, M. Scalable prediction of information cascades over arbitrary time horizons, 2020.

Harris, T. E. *The Theory of Branching Process*. RAND Corporation, Santa Monica, CA, 1964.

Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 04 1971a. ISSN 0006-3444. doi: 10.1093/biomet/58.1.83.

Hawkes, A. G. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971b. ISSN 00359246. URL http://www.jstor.org/stable/2984686.

Hawkes, A. G. and Oakes, D. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974. doi: 10.2307/3212693.

Kobayashi, R. and Lambiotte, R. Tideh: Time-dependent hawkes process for predicting retweet dynamics. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), Mar. 2016. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14717.

Li, C., Ma, J., Guo, X., and Mei, Q. Deepcas: An end-to-end predictor of information cascades. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 577–586, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052643.

Liao, D., Xu, J., Li, G., Huang, W., Liu, W., and Li, J. Popularity prediction on online articles with deep fusion of temporal process and content features. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 200–207, Jul. 2019. doi: 10.1609/aaai.v33i01.3301200.

Ling, C., Tong, G., and Chen, M. Nestpp: Modeling thread dynamics in online discussion forums. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, pp. 251–260, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370981. doi: 10.1145/3372923.3404796.

Mishra, S., Rizoiu, M.-A., and Xie, L. Feature driven and point process approaches for popularity prediction. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pp. 1069–1078, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4073-1. doi: 10.1145/2983323.2983812.

Nadarajah, S. A review of results on sums of random variables. *Acta Applicandae Mathematicae*, 2008. ISSN 1572-9036. doi: 10.1007/s10440-008-9224-4. URL https://doi.org/10.1007/s10440-008-9224-4.

Oakes, D. The markovian self-exciting process. *Journal of Applied Probability*, 12(1):69–77, 1975. doi: 10.2307/3212408.

O'Brien, J. D., Aleta, A., Moreno, Y., and Gleeson, J. P. Quantifying uncertainty in a predictive model for popularity dynamics. *Phys. Rev. E*, 101:062311, Jun 2020. doi: 10.1103/PhysRevE.101.062311.

Ogata, Y. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981. doi: 10.1109/TIT.1981.1056305.

Singh, P. and Chand, S. Predicting the popularity of rumors in social media using machine learning. In Shukla, R. K., Agrawal, J., Sharma, S., Chaudhari, N. S., and Shukla, K. K. (eds.), *Social Networking and Computational Intelligence*, pp. 775–789, Singapore, 2020. Springer Singapore. ISBN 978-981-15-2071-6.

Szabo, G. and Huberman, B. A. Predicting the popularity of online content. *Communications of the ACM*, 53(8): 80–88, 2010. doi: 10.1145/1787234.1787254.

Tan, W. H. and Chen, F. Predicting the popularity of tweets using internal and external knowledge: an empirical bayes type approach. *AStA Advances in Statistical Analysis*, 2021. doi: 10.1007/s10182-021-00390-z.

Wang, Y., Ye, X., Zhou, H., Zha, H., and Song, L. Linking micro event history to macro prediction in point process models. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1375–1384, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL http://proceedings.mlr.press/v54/wang17f.html.

Wu, Q., Yang, C., Zhang, H., Gao, X., Weng, P., and Chen, G. Adversarial training model unifying feature driven and point process perspectives for event popularity prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 517–526, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271714.

Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X., Chu, S. M., and Zhu, H. On modeling and predicting individual paper citation count over time. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pp. 2676–2682. AAAI Press, 2016. ISBN 9781577357704. URL https://dl.acm.org/doi/10.5555/3060832.3060995.

Xu, X., Zhou, F., Zhang, K., Liu, S., and Trajcevski, G. Casflow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021. doi: 10.1109/TKDE.2021.3126475.

Yu, B., Chen, M., and Kwok, L. Toward predicting popularity of social marketing messages. In Salerno, J., Yang, S. J., Nau, D., and Chai, S.-K. (eds.), *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 317–324, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19656-0.

Yuan, N. J., Zhong, Y., Zhang, F., Xie, X., Lin, C.-Y., and Rui, Y. Who will reply to/retweet this tweet? the dynamics of intimacy from online social interactions. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pp. 3–12, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450337168. doi: 10.1145/2835776.2835800.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1513–1522, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783401.

Zhou, F., Xu, X., Trajcevski, G., and Zhang, K. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Comput. Surv.*, 54(2), March 2021. ISSN 0360-0300. doi: 10.1145/3433000.

A. Proofs

Occasionally, in the proofs that are presented here, we will make use of the following well-known facts regarding a Poisson process with intensity function $\lambda(t) : \mathbb{R}^+ \to \mathbb{R}^+$:

F.1 If $N(t)$ is the count of events generated by this process in $[0, t]$, then $N(t) \sim \text{Poisson}\Big(\int_{\tau=0}^{t} \lambda(\tau)d\tau\Big)$.

F.2 For some $T > 0$, given that $N(T) = n \geq 1$, the occurrence times $\{t_i\}_{i=1}^n$ of these n events are all i.i.d. distributed according to the PDF $f(t) \triangleq \frac{\lambda(t)}{\int_{\tau=0}^{T} \lambda(\tau)d\tau}$ with support $t \in [0, T]$.

Notice that, we derive and present our results for continuous mark distributions that have a PDF $g(m)$. Nevertheless, results for discrete mark distributions can easily be derived by following the general approaches presented in this work.

A.1. Proof of Proposition 4.1

Proof. We show this result via induction. The base case proceeds as follows: the 0^{th}-generation events, *i.e.*, the immigrants, arrive as a Poisson process with intensity $b(\cdot)$. By virtue of Fact F.2, consider a total Z_0 immigrants on $\mathbb{R}^+$, their occurrence times are i.i.d. with PDF $f_0(t) \triangleq \frac{b(t)}{\int_{\tau=0}^{\infty} b(\tau)d\tau} = \frac{1}{\eta}(b * \xi^{*0})(t)$ for $t \geq 0$, no matter what their observed total count Z_0 is.

Now, assume that, for $k \geq 1$, the occurrence times of the Z_{k-1} number of $(k-1)^{\text{th}}$-generation events are i.i.d. distributed with PDF $f_{k-1}(t) = \frac{1}{\eta}(b * \xi^{*(k-1)})(t)$. Each of these $(k-1)^{\text{th}}$-generation events will produce k^{th}-generation offspring events, which will be mutually independent with respect to each other. Let us consider a given $(k-1)^{\text{th}}$-generation event (p, m) with occurrence time p and associated mark m. Then, its offspring will stem from a Poisson process with intensity $\phi_m(t-p)$. Fact F.2 implies that the occurrence times of the offspring events will be i.i.d. and distributed with PDF

$$f_k(t|(p, m)) \triangleq \frac{\phi_m(t-p)}{\int_{\tau=p}^{\infty} \phi_m(\tau-p)d\tau} = \frac{1}{\gamma_m}\phi_m(t-p) \qquad t \geq p \tag{18}$$

Then, the PDF of the unconditional distribution of k^{th}-generation occurrence times can be computed as

$$f_k(t) = \int_{p=0}^{t} \int_{m\in\mathcal{M}} f_k(t|(p, m))f_{k-1}(p, m)dmdp \tag{19}$$

With unpredictable marks, we have that the joint distribution of occurrence times and associated marks follows the PDF $f_{k-1}(p, m) = f_{k-1}(p)g(m)$. Using this fact and defining $\xi(t) \triangleq \mathbb{E}_m\Big\{\frac{\phi_m(t)}{\gamma_m}\Big\}$, (19) becomes

$$\begin{aligned}(19) \Rightarrow \quad f_k(t) &= \int_{p=0}^{t} \left[\int_{m\in\mathcal{M}} \frac{\phi(t-p)}{\gamma_m} g(m)dm\right] f_{k-1}(p)dp = \int_{p=0}^{t} \xi(t-p)f_{k-1}(p)dp \\ &= (\xi * f_{k-1})(t) = \frac{1}{\eta}(b * \xi^{*k})(t) \end{aligned} \tag{3}$$

for $t \geq 0$. □

A.2. Proof of Theorem 4.2

Proof. Define $Z_k \triangleq N_k(\infty)$, *i.e.*, Z_k is the total count of k^{th} generation events. Then, by Fact F.1, $Z_0 \sim \text{Poisson}(\eta)$ and, therefore, Z_0's PGF is given as $G_0(w) \triangleq G_{Z_0}(w) = \mathrm{e}^{\eta(w-1)}$. Next, for $k \geq 1$, given that $Z_{k-1} = z_{k-1}$, consider any $(k-1)^{\text{th}}$-generation event (t_i^{k-1}, m_i^{k-1}), where $i = 1, \ldots, z_{k-1}$, and denote by $Z_{k,i}$ the count of k^{th}-generation offspring it gives rise to. Then, again by Fact F.1, $Z_{k,i}|m_i^{k-1} = m \sim \text{Poisson}(\gamma_m)$ for any $k \geq 1$ and, therefore,

$$\mathbb{P}\{Z_{k,i} = x\} = \int_{m\in\mathcal{M}} \mathbb{P}\big\{Z_{k,i} = x|m_i^{k-1} = m\big\}\, g(m)dm = \mathbb{E}_m\left\{\frac{(\gamma_m)^x \mathrm{e}^{-\gamma_m}}{x!}\right\} \tag{20}$$

for any $k \geq 1$ and $i = 1, \ldots, z_{k-1}$. Thus, $Z_{k,i}$'s PGF is computed as

$$G_{Z_{k,i}}(w) = \sum_{x \geq 0} \mathbb{P}\{Z_{k,i} = x\}\, w^x \overset{(20)}{=} \sum_{x \geq 0} w^x\, \mathbb{E}\left\{ \frac{(\gamma_m)^x \mathrm{e}^{-\gamma_m}}{x!} \right\} =$$

$$= \mathbb{E}\left\{ \mathrm{e}^{-\gamma_m} \sum_{x \geq 0} \frac{(w\gamma_m)^x}{x!} \right\} = \mathbb{E}\left\{ \mathrm{e}^{\gamma_m (w-1)} \right\} \tag{21}$$

for any $k \geq 1$ and $i = 1, \ldots, z_{k-1}$. Let $G(w) \triangleq G_{Z_{k,i}}(w)$. Since $Z_\ell = \sum_{i=1}^{Z_{\ell-1}} Z_{\ell,i}$ for $\ell \geq 1$, a standard result in the theory of branching processes (*e.g.*, see page 5 of (Harris, 1964)) implies that $G_{Z_\ell}(w) = G_{Z_{\ell-1}}(G_{Z_{\ell,i}}(w)) = G_{Z_{\ell-1}}(G(w))$ for $\ell \geq 1$. Hence, by straightforward induction, one obtains the PGF of the k^{th}-generation event count Z_k as

$$G_{Z_k}(w) = G_0\left(G^{\circ k}(w)\right) \qquad\qquad k \geq 1 \tag{22}$$

where $G^{\circ k}(\cdot)$ is the k-fold composition of $G(\cdot)$ with itself.

Finally, notice that $N_k(t) = \sum_{i=1}^{Z_k} u(t - t_i^k)$. Per our discussion in the proof of Proposition 4.1, given $Z_k = z_k$, the random variables $\{t_i^k\}_{i=1}^{z_k}$ are i.i.d. distributed. If $G_k(\cdot)$ is their common PGF, then

$$G_{N_k(t)}(w) = G_{Z_k}(G_k(w)) \qquad\qquad k \geq 1 \tag{23}$$

Furthermore, the random variables $\{u(t - t_i^k)\}_{i=1}^{z_k}$ are i.i.d. distributed as $\text{Bernoulli}(F_k(t))$, where $F_k(t)$ is the t_i^k's common CDF and, therefore, $G_k(w) \triangleq 1 + F_k(t)(w-1)$ is their common PGF. Combining (22) and (23) yields the final result of (4).

□

A.3. Proof of Corollary 4.3

Proof. For $k \geq 0$, given the PGF of $N_k(t)$, one can derive its mean and variance through a well-known result about PGFs (*e.g.*, see page 6 of (Harris, 1964)), which states that

$$\mathbb{E}\{N_k(t)\} = \lim_{w \uparrow 1} G'_{N_k(t)}(w) \tag{24}$$

$$\text{Var}(N_k(t)) = \lim_{w \uparrow 1} G''_{N_k(t)}(w) + \mathbb{E}\{N_k(t)\} - \left(\mathbb{E}\{N_k(t)\}\right)^2 \tag{25}$$

where $G'_{N_k(t)}(\cdot)$ and $G''_{N_k(t)}(\cdot)$ are the first- and second-order derivatives of $G_{N_k(t)}(\cdot)$, whose exact form is provided in (4) of Theorem 4.2. This form entails the PGFs $G_0(\cdot), G(\cdot)$ and $G_k(\cdot)$, which are also defined in Theorem 4.2. In what follows, we will rely on the following facts: $G'_o(1) = \eta$, $G''_o(1) = \eta^2$, $G(1) = 1$, $G'(1) = \gamma \triangleq \mathbb{E}\{\gamma_m\}$, $G''(1) = \nu \triangleq \mathbb{E}\{\gamma_m^2\}$, $G_k(1) = 1$, $G'_k(1) = F_k(t)$ and $G''_k(1) = 0$.

Derivation of $G'_{N_k(t)}(w)$ and $\mathbb{E}\{N_k(t)\}$. For $k \geq 1$, the derivative of $G'_{N_k(t)}(w)$ is computed as

$$(4) \Rightarrow \quad G'_{N_k(t)}(w) = G'_0\left(G^{\circ k}(G_k(w))\right) \frac{dG^{\circ k}(G_k(w))}{dw} \tag{26a}$$

and, hence,

$$(26a) \overset{w=1}{\Rightarrow} \quad G'_{N_k(t)}(1) = \eta \frac{dG^{\circ k}(G_k(w))}{dw}\bigg|_{w=1} \tag{26b}$$

By applying the differentiation chain rule, we obtain

$$\frac{dG^{\circ k}(G_k(w))}{dw} = G'_k(w) \prod_{i=1}^{k} G'\left(G^{\circ(k-i)}(G_k(w))\right) \tag{27a}$$

with the convention that $G^{\circ 0}(u) = u$. This leads to

$$(27a) \Rightarrow \quad \frac{dG^{\circ k}(G_k(w))}{dw}\bigg|_{w=1} = F_k(t)\gamma^k \tag{27b}$$

for $k \geq 1$. Combining the results obtained so far yields

$$(26b) \overset{(27b)}{\Rightarrow} \quad G'_{N_k(t)}(1) = \eta F_k(t)\gamma^k \tag{28}$$

which implies

$$(24) \overset{(28)}{\Rightarrow} \quad \mathbb{E}\{N_k(t)\} = \eta F_k(t)\gamma^k \tag{5}$$

Derivation of $G''_{N_k(t)}(w)$ and $\mathrm{Var}(N_k(t))$. For $k \geq 1$, the second order derivative of $G_{N_k(t)}(w)$ is computed via the chain rule as

$$(4) \Rightarrow \quad G''_{N_k(t)}(w) = G''_0\big(G^{\circ k}(G_k(w))\big)\left[\frac{dG^{\circ k}(G_k(w))}{dw}\right]^2 + \\ + G'_0\big(G^{\circ k}(G_k(w))\big)\frac{d^2G^{\circ k}(G_k(w))}{dw^k} \tag{29a}$$

and, thus,

$$(29a) \overset{w=1,(27b)}{\Rightarrow} \quad G''_{N_k(t)}(1) = \eta^2\left[F_k(t)\gamma^k\right]^2 + \eta\frac{d^2G^{\circ k}(G_k(w))}{dw^2}\bigg|_{w=1} \tag{29b}$$

Also, after applying once again the chain rule and performing some algebraic manipulations, one obtains that

$$(27a) \Rightarrow \quad \frac{d^2G^{\circ k}(G_k(w))}{dw^2} = \frac{dG^{\circ k}(G_k(w))}{dw}\cdot \\ \cdot\left[\frac{G''_k(w)}{G'_k(w)} + \sum_{i=1}^{k}\frac{G''\big(G^{\circ(k-i)}(G_k(w))\big)}{G'\big(G^{\circ(k-i)}(G_k(w))\big)}\frac{dG^{\circ(k-i)}(G_k(w))}{dw}\right] \tag{30a}$$

and, thus,

$$(30a) \overset{(27b)}{\Rightarrow} \quad \frac{d^2G^{\circ k}(G_k(w))}{dw^2}\bigg|_{w=1} = \nu F_k(t)\gamma^{k-1}\sum_{i=1}^{k}\frac{dG^{\circ(k-i)}(G_k(w))}{dw}\bigg|_{w=1} \tag{30b}$$

Finally, once again, using the chain rule, one obtains that

$$\frac{dG^{\circ(k-i)}(G_k(w))}{dw} = G'_k(w)\prod_{j=1}^{k-i}G'\Big(G^{\circ(k-i-j)}(G_k(w))\Big) \qquad 1 \leq i \leq k \tag{31a}$$

with the convention that $\prod_{j=1}^{0}(\cdot) = 1$. This yields

$$(31a) \Rightarrow \quad \frac{dG^{\circ(k-i)}(G_k(w))}{dw}\bigg|_{w=1} = F_k(t)\gamma^{k-i} \qquad 1 \leq i \leq k \tag{31b}$$

Putting everything together yields

$$(29b) \overset{(30b),(31b)}{\Rightarrow} \quad G''_{N_k(t)}(1) = \left[\eta F_k(t)\gamma^k\right]^2 + \eta\nu F_k^2(t)\gamma^{k-1}\sum_{i=0}^{k-1}\gamma^i = \\ = \left(1 + \frac{\nu}{\eta\gamma^2}\sum_{j=0}^{k-1}\frac{1}{\gamma^j}\right)\left(\mathbb{E}\{N_k(t)\}\right)^2 \tag{32}$$

which finally yields

$$\mathrm{Var}(N_k(t)) = \mathbb{E}\{N_k(t)\} + \left(\mathbb{E}\{N_k(t)\}\right)^2\left(\frac{\nu}{\eta\gamma^2}\right)\sum_{j=0}^{k-1}\frac{1}{\gamma^j} \tag{6}$$

□

Proof. The results are derived by taking the first- and second-order derivative of $G_{N_k(t)}(w)$ at 1^-. Let's start by $G_{(k)}(w)$, the k-fold composition of $G(w)$. Define $\gamma \triangleq G'(1) = \int_m \gamma_m g(m) dm$ and $\nu \triangleq G''(1) = \int_m (\gamma_m)^2 g(m) dm$, by composition rules, we have

$$G'_{(k)}(1) = \gamma^k \quad \text{and} \quad G''_{(k)}(1) = \nu\gamma^{k-1} \sum_{j=0}^{k-1} \gamma^j$$

It is also trivial to show $G'_0(1) = \eta$, and $G''_0(1) = \eta^2$ for $G_0(w)$, and $G'_{\delta_k(t)}(1) = F_k(t)$, and $G''_{\delta_k(t)}(1) = 0$ for $G_{\delta_k(t)}(w)$. Hence,

$$G'_{N_k(t)}(1) = G'_0(1) G'_{(k)}(1) G'_{\delta_k^t}(s) = \eta\gamma^k F_k(t)$$

and

$$\begin{aligned} G''_{N_k(t)}(1) =& G''_0(1) \left(G'_{(k)}(1) G'_{\delta_k(t)}(1)\right)^2 + G'_0(1) G''_{(k)}(1) \left(G'_{\delta_k(t)}(1)\right)^2 + G'_0(1) G'_{(k)}(1) G''_{\delta_k(t)}(1) \\ =& \eta^2 \left(\gamma^k F_k(t)\right)^2 + \eta\nu\gamma^{k-1} \sum_{j=0}^{k-1} \gamma^j \left(F_k(t)\right)^2 \end{aligned}$$

And the mean and variance are derived by $\mathbb{E}\{N_k(t)\} = G'_{N_k(t)}(1)$ and $\mathrm{Var}(N_k(t)) = G''_{N_k(t)}(1) + G'_{N_k(t)}(1) - \left(G'_{N_k(t)}(1)\right)^2$. □

A.4. Proof of Proposition 4.4

Proof. For some $p \geq 0$, consider the i^{th} event of the p^{th} generation (t_i^p, m_i^p) of the process and let $\widetilde{N}_i(t)$ indicate the number of offspring this event causes and that belong to the q^{th} generation of the process by time $t \geq t_i^p$, where $q > p$. Due to the particular generative structure of an MHPP, one can easily argue that $\widetilde{N}_i(t)$ has the same distribution as the event count $N_{q-p}(t)$ of the $(q-p)^{\text{th}}$ generation at time t of a similar process, which has been started at time t_i^p instead of at time $t = 0$ and that has a base intensity of $\delta(\cdot)$ instead of $b(\cdot)$. This is the key observation for this result.

Moreover, let $f_p(\cdot | t_i^p \leq t)$ be the conditional PDF of t_i^p given that its associated event occurs before time t. Then, $f_p(x | t_i^p \leq t) = \frac{f_p(x)}{(u * f_p)(t)}$ for $0 \leq x \leq t$ (and equals 0, if otherwise), where $f_p(\cdot)$ is t_i^p's unconditional PDF, which, Then, we have that

$$\begin{aligned} \mathbb{E}\Big\{\widetilde{N}_i(t)\Big\} &= \mathbb{E}\{N_{q-p}(t - t_i^p)\} = \mathbb{E}_{t_i^p}\{\mathbb{E}\{N_{q-p}(t - t_i^p) | \, t_i^p\}\} = \int_{\mathbb{R}} \mathbb{E}\{N_{q-p}(t-x) | \, t_i^p = x\} f_p(x | t_i^p \leq t) dx \overset{(3)}{=} \\ &= \frac{1}{(u * f_p)(t)} \int_{\mathbb{R}} \mathbb{E}\{N_{q-p}(t-x) | \, t_i^p = x\} f_p(x) dx \end{aligned} \tag{33}$$

By virtue of Proposition 4.1, $f_p(\cdot)$ is given by (3) in the main paper as

$$f_p(x) \overset{(3)}{=} \frac{1}{\eta}(b * \xi^{*p})(x) \qquad\qquad x \geq 0 \tag{34}$$

Furthermore, based on our key observation and (5) of Corollary 4.3, $\mathbb{E}\{N_{q-p}(t-x) | \, t_i^p = x\}$ is given as

$$\mathbb{E}\{N_{q-p}(t-x) | \, t_i^p = x\} \overset{(5)}{=} (u * \zeta^{*(q-p)})(t-x) \qquad\qquad t \geq x > 0 \tag{35}$$

As a reminder, $\zeta(\cdot) \triangleq \gamma\xi(\cdot)$. Thence, we obtain

$$(33) \overset{(34),(35)}{\Rightarrow} \quad \mathbb{E}\Big\{\widetilde{N}_i(t)\Big\} = \frac{(u * b * \xi^{*q})(t)}{(u * b * \xi^{*p})(t)} = \frac{\mathbb{E}\{N_q(t)\}}{\mathbb{E}\{N_p(t)\}} \tag{36}$$

Based on our earlier statements, it holds that $N_q(t) = \sum_{i=1}^{N_p(t)} \widetilde{N}_i(t)$ and, therefore,

$$\mathbb{E}\{N_q(t) | N_p(t)\} = N_p(t) \, \mathbb{E}\Big\{\widetilde{N}_i(t)\Big\} \overset{(36)}{=} N_p(t) \frac{\mathbb{E}\{N_q(t)\}}{\mathbb{E}\{N_p(t)\}} \tag{37}$$

Since one has that

$$\mathbb{E}\{N_q(t)N_p(t)\} = \mathbb{E}\{N_q(t)|N_p(t)\}\,\mathbb{E}\{N_p(t)\} \tag{38}$$

and

$$\mathbb{E}\big\{N_p^2(t)\big\} = \mathrm{Var}(N_p(t)) + \big(\mathbb{E}\{N_p(t)\}\big)^2 \tag{39}$$

we readily obtain that

$$\mathrm{Cov}(N_p(t), N_q(t)) \overset{(37),(38),(39)}{=} \frac{\mathbb{E}\{N_q(t)\}}{\mathbb{E}\{N_p(t)\}}\,\mathrm{Var}(N_p(t)) \tag{40}$$

which holds for $0 \leq p \leq q$. Interchanging the roles of p and q in (40) yields (7) of Proposition 4.4.

□

A.5. Proof of Theorem 4.5

Proof. In light of Proposition 4.4, the proof of this result is straightforward. Obviously, $N(t) = \sum_{p\geq 0} N_p(t)$, where $N_p(t)$ is the count of p^{th}-generation events up to time t. Taking expectations yields (8). Next, based on the same fact, we have that

$$\mathrm{Var}(N(t)) = \mathrm{Cov}\left(\sum_{p\geq 0} N_p(t), \sum_{q\geq 0} N_q(t)\right) = \sum_{p\geq 0}\sum_{q\geq 0} \mathrm{Cov}(N_p(t), N_q(t)) \tag{41}$$

By virtue of (7) of Proposition 4.4, after some manipulations of the sums involved in (41) and noting that $\mathrm{Cov}(\cdot,\cdot) \equiv \mathrm{Var}(\cdot)$, one arrives at the expression for the variance of $N(t)$ given in (9). □

A.6. Proof of Lemma 4.6

Proof. Let us assume an MHPP with ground intensity $\lambda(t|\mathcal{H}_{t^-}) \equiv \lambda^*(t)$ as given by

$$\lambda(t|\mathcal{H}_{t^-}) = b(t) + \sum_{i:t_i<t} \phi_{m_i}(t - t_i) \tag{1}$$

and assume that we have observed the process' history $\mathcal{H}_{t_c} = \{(t_i, m_i)\}_{i:t_i\leq t_c}$ up to some censoring time $t_c > 0$, which consists of $N(t_c)$ events. Then, for $t > t_c$ we can express (1) as

$$\lambda(t|\mathcal{H}_{t^-}) = \underbrace{b(t) + \sum_{i:t_i\leq t_c} \phi_{m_i}(t - t_i)}_{\equiv\lambda(t|\mathcal{H}_{t_c})} + \sum_{i:t_c<t_i<t} \phi_{m_i}(t - t_i) \tag{42}$$

From this decomposition, if we concern ourselves only with events occurring past t_c, having observed all $N(t_c)$ prior events in $[0, t_c]$, we can view the relevant generating process as another MHPP, which starts at time t_c and features a conditional event time intensity given by (42). Note that, under these circumstances, $\lambda(t|\mathcal{H}_{t_c^-})$ is a non-stochastic intensity and serves as the base intensity $\hat{b}(\cdot)$ of the newly-defined MHPP. In particular, since the newly-defined process is active for $t > t_c$, let $\Delta t \triangleq t - t_c$; then, $\hat{b}(\Delta t) = \lambda(\Delta t + t_c|\mathcal{H}_{t_c})$ and is given by (10). Furthermore, the latter process features the same excitation function $\phi_m(\cdot)$ as the original process. Also, it is easy to discern that, if the original process features independent marks, then so does the newly-defined one; both processes are endowed with the same mark distribution $g(\cdot)$. Finally, it is obvious that the event count $N(t) = N(\Delta t + t_c)$ of the original process is going to be given as the sum of $N(t_c)$, plus the event count $\hat{N}(\Delta t)$ of the newly-defined process, *i.e.*, $N(\Delta t + t_c) = N(t_c) + \hat{N}(\Delta t)$. □

B. Tweet Popularity Prediction with CASPER

B.1. Calculating Conditional Mean and Variance

Here we present the numerical details involved in getting the conditional mean count shown in Eq. (16) for MHPP with conditional intensity Eq. (15), *i.e.*

$$\lambda^*(t, m) = \alpha \sum_{i:t_i<t} m_i^\kappa \mathrm{e}^{-\beta(t-t_i)} g(m) \tag{15}$$

For deriving the conditional moments, we first find the new base function $\hat{b}(\cdot)$ following Eq. (10) as shown in Lemma 4.6,

$$\hat{b}(\tau) = \sum_{t_i \le t_c} \alpha m_i^\kappa \mathrm{e}^{-\beta(\tau + t_c - t_i)} \tag{43}$$

Now, following Theorem 4.7, we calculate the following terms

$$\eta = \int_{s=0}^{\infty} \hat{b}(s) ds = \frac{\alpha}{\beta} \sum_{t_i \le t_c} m_i^\kappa \mathrm{e}^{-\beta(t_c - t_i)} \tag{44}$$

$$\gamma_m = \int_{s=0}^{\infty} \alpha m^\kappa \mathrm{e}^{-\beta s} ds = \frac{\alpha m^\kappa}{\beta} \tag{45}$$

$$\gamma = \sum_m \gamma_m g(m) = \frac{\alpha}{\beta} \sum_m m^\kappa g(m) = \frac{\alpha f_1(\kappa)}{\beta}, \text{ where } f_1(\kappa) \triangleq \sum_m m^\kappa g(m) \tag{46}$$

$$\nu = \sum_m \gamma_m^2 g(m) = \frac{\alpha^2}{\beta^2} \sum_m m^{2\kappa} g(m) = \frac{\alpha^2 f_2(\kappa)}{\beta^2}, \text{ where } f_2(\kappa) \triangleq \sum_m m^{2\kappa} g(m) \tag{47}$$

And the ζ function,

$$\zeta(\tau) = \gamma \sum m \frac{1}{\gamma_m} \phi_m(\tau) g(m) = \gamma\beta \sum_m \mathrm{e}^{-\beta\tau} g(m) = \gamma\beta \mathrm{e}^{-\beta\tau} \tag{48}$$

For mathematical simplicity, we do not explicitly show α and instead refer to its relation with γ, *i.e.* our parameter set id $\boldsymbol{\theta} = \{\gamma, \beta, \kappa\}$. The α value can be retrieved by $\alpha = \gamma\beta / f_1(\kappa)$. Accordingly,

$$\eta = \frac{\gamma}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^\kappa \mathrm{e}^{-\beta(t_c - t_i)} \tag{49}$$

$$\nu = \frac{\gamma^2 f_2(\kappa)}{f_1^2(\kappa)} \tag{50}$$

and

$$\hat{b}(\tau) = \frac{\gamma\beta}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^\kappa \mathrm{e}^{-\beta(\tau + t_c - t_i)} \tag{51}$$

The k-fold convolution of the function ζ follows

$$\zeta^{(*k)}(\tau) = \frac{(\gamma\beta)^k \tau^{k-1}}{(k-1)!} \mathrm{e}^{-\beta\tau} \tag{52}$$

and

$$\left(u * \zeta^{(*k)}\right)(\tau) = \gamma^k \left(1 - \sum_{j=0}^{k-1} \frac{(\beta\tau)^j}{j!} \mathrm{e}^{-\beta\tau}\right) \tag{53}$$

Also,

$$\hat{b}(\tau) = \frac{\gamma\beta}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \mathrm{e}^{-\beta(\tau + t_c - t_i)} = \frac{1}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \zeta(\tau + t_c - t_i) u(\tau) \tag{54}$$

hence

$$\left(\hat{b} * \zeta^{*k}\right)(\tau) = \frac{1}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \left(\zeta^{*k}(\tau) * \zeta(\tau + t_c - t_i) u(\tau)\right) \tag{55}$$

$$= \frac{1}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \mathrm{e}^{-\beta(t_c - t_i)} \zeta^{*(k+1)}(\tau) \tag{56}$$

Now, combining all of the above expressions to derive the conditional mean count,

$$\mathbb{E}\left\{\hat{N}_k(\tau)\right\} = \left(u * \hat{b} * \zeta^{*k}\right)(\tau) \tag{57}$$

$$= \frac{1}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \mathrm{e}^{-\beta(t_c - t_i)} \left(u * \zeta^{*(k+1)}\right)(\tau) \tag{58}$$

$$= \frac{1}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \mathrm{e}^{-\beta(t_c - t_i)} \gamma^{k+1} \left(1 - \sum_{j=0}^{k} \frac{(\beta\tau)^j}{j!} \mathrm{e}^{-\beta\tau}\right) \tag{59}$$

$$= \gamma^k \eta \left(1 - \sum_{j=0}^{k} \frac{(\beta\tau)^j}{j!} \mathrm{e}^{-\beta\tau}\right) \tag{60}$$

The variance $\mathrm{Var}\left(\hat{N}_k(\tau)\right)$ directly follows Eq. (6),

$$\mathrm{Var}\left(\hat{N}_k(\tau)\right) = \mathbb{E}\left\{\hat{N}_k(\tau)\right\} + \left(\mathbb{E}\left\{\hat{N}_k(\tau)\right\}\right)^2 \left(\frac{\nu}{\eta\gamma^2}\right) \sum_{j=0}^{k-1} \frac{1}{\gamma^j}$$

For the exponential triggering function, we can obtain a closed form solution for the mean as,

$$\sum_{k=1}^{\infty} \zeta^{(*k)}(\tau) = \gamma\beta \mathrm{e}^{-\beta(1-\gamma)\tau} \tag{61}$$

$$\left(u * \sum_{k=1}^{\infty} \zeta^{(*k)}\right)(\tau) = \frac{\gamma}{1-\gamma} \left(1 - \mathrm{e}^{-\beta(1-\gamma)\tau}\right) \tag{62}$$

And thus, for $\gamma \neq 1$,

$$\mathbb{E}\left\{\hat{N}(\tau)\right\} = \left(u * \hat{b} * \sum_{k=0}^{\infty} \zeta^{(*k)}\right)(\tau)$$

$$= \frac{1}{f_1(\kappa)} \sum_{t_i \le t_c} m_i^{\kappa} \mathrm{e}^{-\beta(t_c - t_i)} \left(u * \sum_{k=1}^{\infty} \zeta^{*k}(\tau)\right)$$

$$= \frac{\eta}{1-\gamma} \left(1 - \mathrm{e}^{-\beta(1-\gamma)\tau}\right)$$

When $\gamma = 1$, it can be easily shown that $\mathbb{E}\left\{\hat{N}(\tau)\right\} = \eta\beta\tau$, and therefore, we obtain Eq. (16).

For the variance, unfortunately, we cannot find a closed form solution. Hence, we estimate the variance by truncating the infinite sum,

$$\mathrm{Var}\Big(\hat{N}(\tau)\Big) \approx \sum_{k=0}^{K_{\max}} \left(1 + \frac{2}{\mathbb{E}\Big\{\hat{N}_k(\tau)\Big\}} \left(\mathbb{E}\Big\{\hat{N}(\tau)\Big\} - \sum_{j=0}^{k} \mathbb{E}\Big\{\hat{N}_j(t)\Big\}\right)\right) \cdot \mathrm{Var}\Big(\hat{N}_k(\tau)\Big) \tag{63}$$

B.2. Training Details

As noted in the paper, our model is optimized by minimizing the proposed objective function Eq. (12) using a projected gradient descent algorithm. In this section, we provide the objective function for the retweeting process, and obtain its gradients with respect to the model parameters.

Let $\mathcal{S}(t_c) \triangleq i, j : 0 < t_i < t_j \leq t_c$ and $n \triangleq N(t_c)$

$$\begin{aligned} L(\boldsymbol{\theta}|\mathcal{H}_{t_c}) &= \frac{2}{n(n+1)} \sum_{(i,j)\in\mathcal{S}(t_c)} \sum \left(\mathbb{E}\Big\{\widetilde{N}(t_j|\mathcal{H}_{t_i})\Big\} - N(t_j)\right)^2 \\ &= \frac{2}{n(n+1)} \sum_{(i,j)\in\mathcal{S}(t_c)} \sum \left(p_{ij}(\gamma,\beta) q_i(\kappa,\beta) - N(t_j)\right)^2 \end{aligned} \tag{64}$$

where,

$$p_{ij}(\gamma,\beta) = \left(\frac{\gamma}{1-\gamma}\left(1 - \mathrm{e}^{-\beta(1-\gamma)(t_j - t_i)}\right)\right)^{[\![\gamma\neq 1]\!]} (\gamma\beta\tau)^{[\![\gamma=1]\!]} \tag{65}$$

$$q_i(\kappa,\beta) = \frac{1}{f_1(\kappa)} \left(\sum_{t_v \leq t_i} m_i^{\kappa} \mathrm{e}^{-\beta(t_i - t_v)}\right) \tag{66}$$

And its gradients with respect to the model parameters are,

$$\frac{\partial f(\gamma,\beta,\kappa)}{\partial \gamma} = \frac{2}{n(n+1)} \sum_{(i,j)\in\mathcal{S}(t_c)} \sum 2\Big(p_{ij}(\gamma,\beta) q_i(\kappa,\beta) - (j-i)\Big) q_i(\kappa,\beta) \frac{\partial p_{ij}(\gamma,\beta)}{\partial \gamma} \tag{67}$$

$$\frac{\partial f(\gamma,\beta,\kappa)}{\partial \kappa} = \frac{2}{n(n+1)} \sum_{(i,j)\in\mathcal{S}(t_c)} \sum 2\Big(p_{ij}(\gamma,\beta) q_i(\kappa,\beta) - (j-i)\Big) p_{ij}(\gamma,\beta) \frac{\partial q_i(\kappa,\beta)}{\partial \kappa} \tag{68}$$

$$\frac{\partial f(\gamma,\beta,\kappa)}{\partial \beta} = \frac{2}{n(n+1)} \sum_{(i,j)\in\mathcal{S}(t_c)} \sum 2\Big(p_{ij}(\gamma,\beta) q_i(\kappa,\beta) - (j-i)\Big) \left(p_{ij}(\gamma,\beta) \frac{\partial q_i(\kappa,\beta)}{\partial \beta} + q_i(\kappa,\beta) \frac{\partial p_{ij}(\gamma,\beta)}{\partial \beta}\right) \tag{69}$$

$$\tag{70}$$

where,

$$\frac{\partial p_{ij}(\gamma,\beta)}{\partial \gamma} = \left(\frac{1}{(1-\gamma)^2}\left(1 - \mathrm{e}^{-\beta(1-\gamma)(t_j - t_i)}\right) - \frac{\gamma}{1-\gamma}\beta(t_j - t_i)\mathrm{e}^{-\beta(1-\gamma)(t_j - t_i)}\right)^{[\![\gamma\neq 1]\!]} (\beta\tau)^{[\![\gamma=1]\!]} \tag{71}$$

$$\frac{\partial p_{ij}(\gamma,\beta)}{\partial \beta} = \Big(\gamma(t_j - t_i)\mathrm{e}^{-\beta(1-\gamma)(t_j - t_i)}\Big)^{[\![\gamma\neq 1]\!]} (\gamma\tau)^{[\![\gamma=1]\!]} \tag{72}$$

$$\frac{\partial q_i(\kappa,\beta)}{\partial \kappa} = -\frac{f_1'(\kappa)}{f_1^2(\kappa)} \left(\sum_{t_v \leq t_i} m_i^{\kappa} \mathrm{e}^{-\beta(t_i - t_v)}\right) + \frac{1}{f_1(\kappa)} \left(\sum_{t_v \leq t_i} (\ln m_i)\, m_i^{\kappa} \mathrm{e}^{-\beta(t_i - t_v)}\right) \tag{73}$$

$$\frac{\partial q_i(\kappa,\beta)}{\partial \beta} = -\frac{1}{f_1(\kappa)} \left(\sum_{t_v \leq t_i} m_i^{\kappa} (t_i - t_v) \mathrm{e}^{-\beta(t_i - t_v)}\right) \tag{74}$$

C. Comparison with CasFlow

Deep learning based prediction models, such as (Cao et al., 2017; Li et al., 2017; Chen et al., 2019; Xu et al., 2021), have gained their popularity in recent year. However, they all require (at the minimum, among all exploited features) the cascade graph structure, *i.e.*, the explicit retweeting paths, to construct their networks' input layers. Such information, however, is not always available. Twitter, for example, does not disclose such information. Furthermore, these models need excessively large number of fully-observed cascades for model training and validation, which makes the comparison with point process based models, unreasonable.

Regardless, due to the increasing popularity of deep learning approaches, we compare our model with CasFlow (Xu et al., 2021), the state-of-the-art deep learning model, for Weibo message popularity prediction. Sina Weibo is the largest microblogging platform in China. Each tweet and its retweets form a retweeting cascade, and the explicit retweeting path can be retrieved.

Following the same setup of CasFlow, we first filter out cascades who has less than 10 observed events, and focus on tweets posted between 8 a.m. and 6 p.m., leaving each tweet at least 6 hours to reap retweets. The filtered data is then randomly split it into training set (70%), validation set (15%), and test set (15%). Again, following the setup of CasFlow, which observe the Weibo cascades for 30 minutes and 1 hour, and predict their retweet counts at 24 hours, we have $(t_c = 0.5 \text{ hour}, \Delta t = 23.5 \text{ hour})$, and $(t_c = 1 \text{ hour}, \Delta t = 23 \text{ hour})$.

In the case of $(t_c = 0.5 \text{ hour}, \Delta t = 23.5 \text{ hour})$, after filtering and splitting, we end up with 21463 training, 4599 validation, and 4599 test cascades. In case of $(t_c = 1 \text{ hour}, \Delta t = 23 \text{ hour})$, we end up with 29908 training, 6409 validation, and 6408 test cascades. As CASPER does not need any full-observed cascades for train and validation, only the test set is used in producing the prediction results of CASPER. The train and validation sets are used solely in training the CasFlow model.

Further, CasFlow build a global network graph from all queried Weibo cascades. We estimate the number of followers (mark values) with the count of edge from this global graph. Unlike CasFlow, which use the retweeting path information of each cascade, CASPER only use the time stamps, together with the follower numbers extracted above for model train and predict.

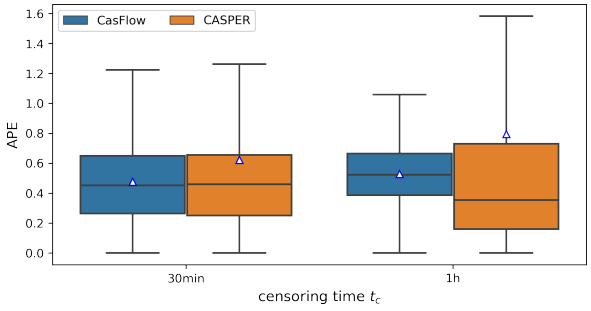


Figure 3. Boxplots of APE values between CASPER and CasFlow model on Weibo data for the above two setups. The Horizontal bars within the boxes indicate the median values, and the white triangles indicate the mean values.

We report the Absolute Percentage Error (APE) values across test cascades in Figure 3. As showed, for $t_c = 30$min, CasFlow reports slightly lower median and lower mean APE values. In the case of $t_c = 1$hour, our CASPER reports much lower median APE values, but with higher mean and larger variance.

Again, let's emphasis, the comparison between our model and deep learning model CasFlow, is not rigid due to the extreme imbalance of data used in model training and prediction, and the huge difference in computational cost.