Theoretical Advancements in Hawkes Processes and Their Practical Applications

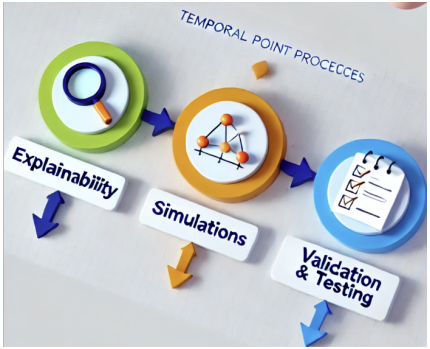
Xi Zhang
Florida Institute of Technology
November 19, 2024

Introduction

Temporal Point Processs (TPPs): mathematical models to describe the occurrence of events over time, well-suited for scenarios where the timing between events is critical

- Customer arrivals in retail
- Machine failure events in industrial
- Disease transmission in epidemiology
- Neural spike trains in neuroscience
- Trade executions in financial markets

Modeling with TPPs



- **Flexible Framework:** Model event dependencies, time-varying intensities, and incorporate external factors.
- **Generative Approach:**
  - Offers explainability
  - Facilitates simulations
  - Supports model validation and hypothesis testing
- **Impact:** Gain insights into the mechanisms driving event occurrences and develop predictive tools that enhance decision-making across diverse applications.

Hawkes Processes

Hawkes processes, introduced by Alan Hawkes in the early 1970s, are self-exciting models where each event increases the likelihood of future events (Hawkes, 1971).

- Applications:
  - Seismology: Each earthquake increases the likelihood of aftershocks.
  - Finance: Clustering of trades and market events over time.
  - Social Media: Information cascades and heavy-tailed engagement patterns.
- A suitable model for capturing event clustering, the rich-get-richer phenomenon, and heavy-tailed distributions.

Identified Gaps and Limitations

- **Model Evaluation and Selection:**
  - Goodness-of-Fit tests via the time-rescaling theorem are not applicable to terminating point process and right-censored data.
  - There is a need for broadly applicable methods for model evaluation and selection.
- **Hawkes Nonparametric Estimation:**
  - Current methods rely on complex estimation techniques.
  - Simple, intuitive and flexibility would be a valuable tool for many real-world scenarios.
- **Hawkes Process-based Prediction Model**
  - Sub-optimal prediction performance when adopt Hawkes process for prediction tasks.
  - Enhanced strategies are needed for real-world applicability and competitiveness.

Contributions

This doctoral work presents novel theoretical insights and strategies to address the above challenges, thereby advancing the theory and enhancing the practical applicability of Temporal Point Processes, with a particular focus on Hawkes processes.

Presentation Outlines

FLORIDA TECH

- Temporal Point Processes
- Generalized Time-Rescaling Theorem
- Nonparametric Kernel-Based Intensity Estimation for Hawkes Processes
- CASPER: Cascade Size Prediction for Self-Exciting Processes via Regression
- Summary

Contents

Definition: Temporal Point Process (TPP)

A **Temporal Point Process (TPP)** is a stochastic process where realizations represent the timings of discrete events along a timeline. The process can be defined by

- Conditional Intensity Function

$$\lambda^*(t) \triangleq \lambda(t|\mathcal{H}_{t^-}) = \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\{N[t, t+\Delta t) = 1|\mathcal{H}_{t^-}\}}{\Delta t} \tag{1}$$

- Next Event Time Distribution $f_{T_i}(t|\mathcal{H}_{t_{i-1}})$, $F_{T_i}(t|\mathcal{H}_{t_{i-1}})$ and $S_{T_i}(t|\mathcal{H}_{t_{i-1}})$
- Counting Process $N(t) \in \mathbb{Z}_+$, counts the number of events up to and includes time t

TPP Definition

The cumulative conditional intensity function $\Lambda^*(t) \triangleq \Lambda(t|\mathcal{H}_{t^-})$,

$$\Lambda^*(t) = \int_0^t \lambda^*(\tau)\, d\tau \tag{2}$$

which quantifies the cumulative expected number of events up to time t, conditioned on the process's history. Specifically, it satisfies $\Lambda^*(t) = \mathbb{E}[N(t)|\mathcal{H}_{t^-}]$

Relationships

$$f_{T_i}(t|\mathcal{H}_{t_{i-1}}) = \lambda^*(t) S_{T_i}(t|\mathcal{H}_{t_{i-1}}) \tag{3}$$

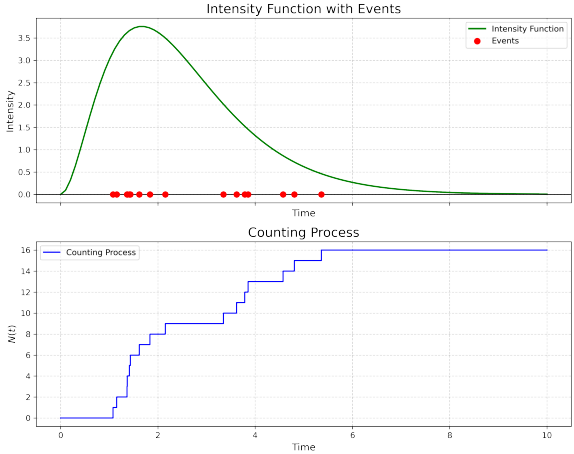
$$S_{T_i}(t|\mathcal{H}_{t_{i-1}}) = \exp\left(-\int_{t_{i-1}}^t \lambda^*(t) d\tau\right) \tag{4}$$

Example: Inhomogeneous Poisson Process

Poisson Process:

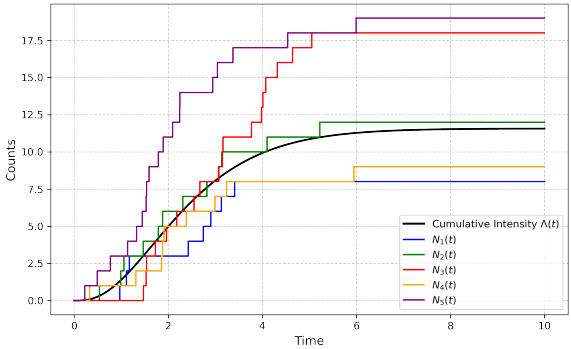
- Constant Homogeneous: $\lambda(t|\mathcal{H}_{t^-}) = \lambda$
- Time-Varying Inhomogeneous: $\lambda(t|\mathcal{H}_{t^-}) = \lambda(t)$

Example shows an Inhomogeneous Poisson process with intensity function $\lambda(t) = 10t^2e^{-1.2t}$.



Example: Inhomogeneous Poisson Process

- This plot shows **five realizations** of the inhomogeneous Poisson process
- For Poisson processes, the cumulative intensity represents the mean counts over time.



Non-Terminating vs. Terminating Processes

Non-Terminating Process:

- Generates events indefinitely, with every realization a.s. containing an infinite number of events.
- $\Lambda^*(t) \to \infty$ as $t \to \infty$ for all histories.
- **Example:** A Poisson process with a constant rate, such as customer arrivals in a 24/7 store.

Terminating Process:

- Has a non-zero probability of ceasing to generate events after a finite number of events.
- $\Lambda^*(t)$ is bounded for at least one history.
- **Example:** An earthquake aftershock sequence that ceases after a period of inactivity.

Marked and Multi-variate Processes

Marked Point Processes:

- Each point t_i in the process is associated with a mark m_i that provides additional information.
- *Example:* Earthquake occurrences where each event is marked by its magnitude.

Multi-variate processes

- Consist of multiple, inter-dependent event streams or dimensions, allowing modeling of interactions between event types.
- *Example:* Financial transactions across different assets, where each asset's transactions can affect others.

Contents

GoF for Point Processes

- The Goodness-of-Fit (GoF) assessment evaluates how well a point process model aligns with observed data.
- Validating model fit is essential for ensuring that applied models capture real-world dynamics accurately.
- Guide model selection and refinement.

GoF for Point Processes

A popular approach for GoF uses the time rescaling theorem (Daley and Vere-Jones, 2003) to transform a point process into a unit-rate Poisson process.

This transformation enables the use of statistical tests and visualization techniques to verify if the rescaled event times follow a Poisson process.

- Probability-Probability Plot: A graphical tool used to compare theoretical vs empirical CDF.
- Kolmogorov-Smirnov Test: A statistical test adopted to generate confidence bands

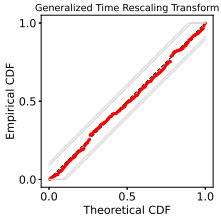


Figure: A straight line indicates a good fit. The inner and outer gray lines represent the 99% and 95% confidence bands, respectively.

Time-Rescaling Theorem

Theorem (Time-Rescaling Theorem)

Let $\{0 < t_1 < t_2 < \cdots\}$ be an unbounded, increasing sequence of time points in the half-line $(0, \infty)$, N^* a simple point process with history $\mathcal{H}$, and monotonic, continuous conditional cumulative intensity $\Lambda^*(t)$ such that $\Lambda^*(t) \to \infty$ a.s.. Then, with probability 1, the transformed sequence $\{v_i = \Lambda^*(t_i)\}$ is a realization of unit-rate Poisson process if and only if the original sequence $\{t_i\}$ is a realization from the point process defined by $\Lambda^*(t)$.

Limitations

- $\Lambda^*(t) \to \infty$ almost surely — Restricted to non-terminating processes
- Unbounded time sequence — Applicable only to complete, uncensored data

Generalized Time-Rescaling Theorem

To address the limitations, we present

Theorem

If $\{t_1, t_2, \ldots, t_n\}$ *is an incomplete realization, right-censored at time* t_c*, of a simple TPP characterized by a conditional intensity function* $\lambda(t|\mathcal{H}_{t^-})$ *with a continuous compensator* $\Lambda(t|\mathcal{H}_{t^-})$*, then* $\{v_i^g\}_{i=1}^n$*, defined as*

$$v_i^g \triangleq \ln\left(\frac{F_{T_i}(t_c|\mathcal{H}_{t_{i-1}})}{F_{T_i}(t_c|\mathcal{H}_{t_{i-1}}) - F_{T_i}(t_i|\mathcal{H}_{t_{i-1}})}\right) = \ln\left(\frac{1 - \mathrm{e}^{-\Lambda_{i-1}(t_c)}}{\mathrm{e}^{-\Lambda_{i-1}(t_i)} - \mathrm{e}^{-\Lambda_{i-1}(t_c)}}\right) \tag{5}$$

are i.i.d. Exp(1) *distributed. Here* $\Lambda_{i-1}(t) \triangleq \int_{t_{i-1}}^{t} \lambda(\tau|\mathcal{H}_{t_{i-1}}, \varnothing_{(t_{i-1},\tau)})d\tau$

Comparison

	Original	**Generalized**
Data	Complete increasing sequence	Incomplete, Right Censored data
$\Lambda^*(t)$	Continuous and unbounded a.s.	Continuous
i.i.d.	$\Lambda_{i-1}(t_i)$	$\ln\left(\frac{1-e^{-\Lambda_{i-1}(t_c)}}{e^{-\Lambda_{i-1}(t_i)}-e^{-\Lambda_{i-1}(t_c)}}\right)$

Table: Comparison of Time Rescaling Theorem and Generalized Time Rescaling Theorem

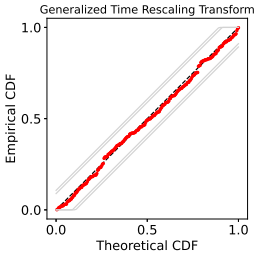
In case of $t_c \to \infty$ and $\Lambda^*(t)$ unbounded a.s., $v_i^g = v_i$.

GoF with the Generalized Theorem

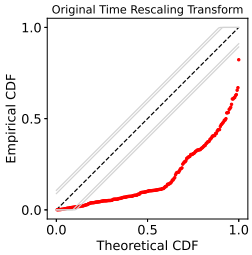
To assess the effectiveness of the generalized time-rescaling theorem in GoF, we consider two scenarios:

- **Case 1:** A **terminating** self-exciting Hawkes process with a long censoring time is employed to approximate the absence of censorship. This model is commonly used to capture the dynamics of event occurrences in social media.
- **Case 2:** A renewal process with Gamma-distributed inter-arrival times, but **short right-censoring** times. Such models frequently applied in neuroscience to model spike train data

GoF - Terminating Process



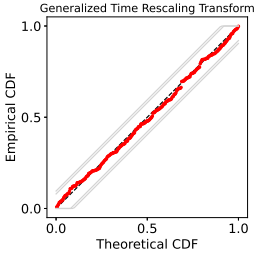
(a) Generalized



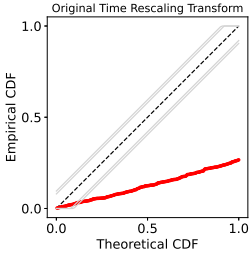
(b) Original

Figure: Scenario 1 (terminating Process) uses a Hawkes point process with conditional intensity function $\lambda^*(t) = 0.5e^{-4t} + \sum_{t_i:t_i<t} \left(e^{-2(t-t_i)}\right)$. The dataset consists of $1,000$ realizations, each censored at $t_c = 1000$.

GoF - Short Observation Time



(a) Generalized



(b) Original

Figure: Scenario 2 (right-censored data) uses a renewal process inter-arrival times $f_{T_i}(t|\mathcal{H}_{t_{i-1}}) = \frac{(t-t_{i-1})^{0.99}e^{-(t-t_{i-1})}}{\Gamma(1.99)}$. The dataset consists of $1,000$ realizations, each censored at $t_c = 1$.

Summary

- We generalize the time-rescaling theorem to handle terminating processes and incomplete observations.
- The generalized theorem applies to marked point processes.
- Experiments confirm the accuracy of this transform for model evaluation in cases where the original theorem inapplicable.
- The extension broadens the applicability of goodness-of-fit (GoF) tests to a wider range of processes and data.

Contents

Hawkes Process

A Hawkes process (Hawkes, 1971) is a self-exciting point process, featuring conditional intensity function:

$$\lambda^*(t) = \mu(t) + \sum_{t_i<t} \phi(t - t_i) \tag{6}$$

- $\mu(\cdot)$: **base or background intensity**, modeling exogenous influences.
- $\phi(\cdot)$: **Excitatory function**, modeling the self-exciting/endogenous influences.

Hawkes processes can capture temporal clustering, where events tend to "self-excite" and cluster in time, unlike the memoryless Poisson process.

Hawkes Self-Exciting Example

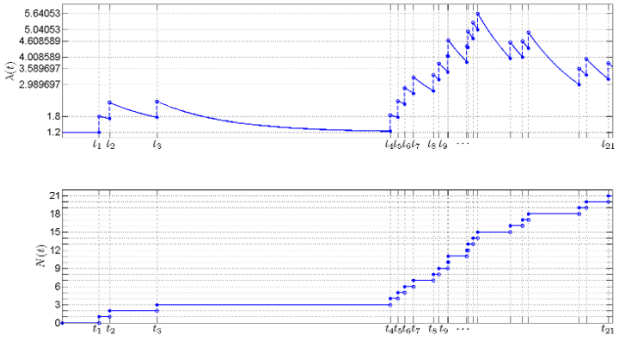


Figure: An illustrative example of a univariate Hawkes process with conditional intensity $\lambda^*(t) = 1.2 + 0.6 \sum_{t_i < t} \exp^{-0.8(t-t_i)}$.

Branching Process Equivalence

A Hawkes process can be viewed as a *branching process* with events generated through two mechanisms:

- *Immigrant Events (0-th Generation):* Occur independently, drawn from a Poisson process with base intensity $\mu(\cdot)$.
- *Offspring Events (1st Generation and Beyond):* Each immigrant event, along with existing offspring, produces new offspring that follow a Poisson process with intensity $\phi(t - t_i)$, based on the time since the parent event at t_i. This process continues recursively.

Illustration

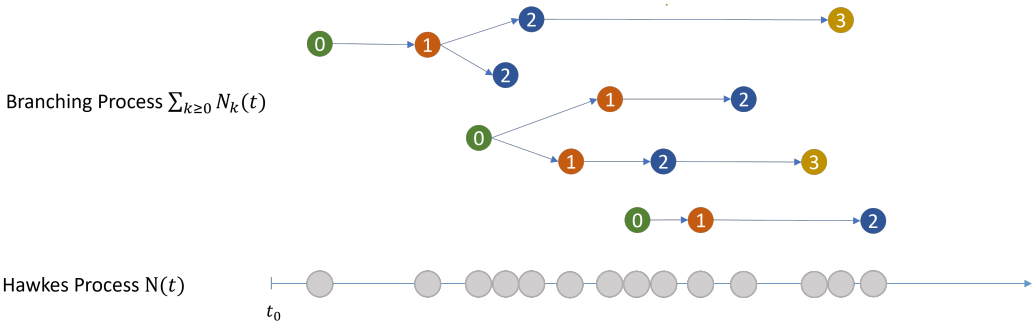


Figure: Illustration of branching process equivalence

Nonparametric Estimation: Objective

Dataset: R realizations $\mathcal{D} = \{\{t_i^r\}_{i=1}^{n_r}, t_c^r\}_{r=1}^{R}$, each censored at time t_c^r. Assume, without loss of generality, that all realizations start at time 0.

Assumption: Data from a Hawkes process with intensity function:

$$\lambda^*(t) = \mu(t) + \sum_{t_i < t} \phi(t - t_i)$$

Objective: Estimate the components of the intensity function:

- Base Intensity Function $\mu(t)$
- Excitatory Function $\phi(t)$

Log-likelihood Minorizer

$$\ell = \sum_{r=1}^{R}\sum_{i=1}^{n_r} \log\left(\mu(t_i^r) + \sum_{j=1}^{i-1}\phi(t_i^r - t_j^r)\right) - \sum_{r=1}^{R}\int_0^{t_c^r}\mu(\tau)d\tau - \sum_{r=1}^{R}\sum_{i=1}^{n}\int_0^{t_c^r - t_i^r}\phi(\tau)d\tau$$

$$\geq \underbrace{\sum_{r=1}^{R}\sum_{i=1}^{n_r} z_{i0}^r \log\mu(t_i^r) - \sum_{r=1}^{R}\int_0^{t_c^r}\mu(\tau)d\tau +}_{\ell_{\text{base}}^{\text{LB}}(\mathcal{D},\{z_{i0}^r\};\mu(t))}$$

$$+ \underbrace{\sum_{r=1}^{R}\sum_{i=2}^{n_r}\sum_{j=1}^{i-1} z_{ij}^r \log\phi(t_i^r - t_j^r) - \sum_{r=1}^{R}\sum_{i=1}^{n_r}\int_0^{t_c^r - t_i^r}\phi(\tau)d\tau}_{\ell_{\text{exci}}^{\text{LB}}(\mathcal{D},\{z_{ij}^r\};\phi(t))} + \text{const.} \tag{7}$$

$\ell = \ell^{\text{LB}}_{\text{base}} + \ell^{\text{LB}}_{\text{exci}} + \text{const.}$ when

$$z^r_{i0} = \frac{\mu(t^r_i)}{\mu(t^r_i) + \sum_{k=0}^{i-1} \phi(t^r_i - t^r_k)} \tag{8}$$

$$z^r_{ij} = \frac{\phi(t^r_i - t^r_j)}{\mu(t^r_i) + \sum_{k=0}^{i-1} \phi(t^r_i - t^r_k)}, \quad 0 < j < i \tag{9}$$

The term z^r_{ij}, referred to as the **declustering (or parenthood) probability**, quantifies the likelihood that event j is the parent of event i. When $j = 0$, this probability indicates that event i originates from the base process, marking it as an immigrant.

Unconstrained Intensity Functions

Assume

- $\{t_{(p)}\}_{p=1}^{P} = \text{unique } \left(\{\{t_i^r\}_{i=1}^{n_r}, t_c^r\}_{r=1}^{R}\right)$
- $\{x_{(q)}\}_{q=1}^{Q} = \text{unique } \left(\{t_i^r - t_j^r\}_{1 \le j < i \le n_r} \cup \{t_c^r - t_k^r\}_k\}_{r=1}^{R}\right)$

If no constraints are imposed on the functions $\mu(\cdot)$ and $\phi(\cdot)$, both $\ell_{\text{base}}^{\text{LB}}$ and $\ell_{\text{exci}}^{\text{LB}}$ can become arbitrarily large with

- $\mu(t) = \sum_{p=1}^{P} w_p \delta(t - t_{(p)})$
- $\phi(t) = \sum_{q=1}^{Q} v_q \delta(t - x_{(q)})$

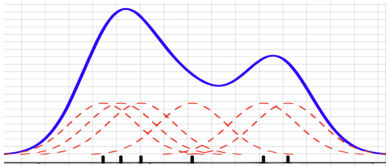
Impractical !

Kernel smoothing. The smoothed kernel-based intensity estimators:

$$\mu(t) = \sum_{p=1}^{P} w_p K_b(t - t_{(p)}), \tag{10}$$

$$\phi(t) = \sum_{q=1}^{Q} v_q K_b(t - x_{(q)}), \tag{11}$$

where $b > 0$ represents a bandwidth parameter and $K_b(x) \triangleq \frac{1}{b} K\left(\frac{x}{b}\right)$ denotes a scaled kernel function with interpolation kernel $K(\cdot)$.



Our Approach

An iterative algorithm executes three main tasks during each iteration:

1. Estimating the declustering probabilities, z_{ij}.
2. Determining the weights w_p and v_q based on the current declustering outcomes.
3. Deriving smoothed intensity estimators using kernel methods.

Weight Derivation via Empirical Likelihood

Inspired by Kernel Density Estimation (KDE), where weights for jumps in the cumulative distribution function (CDF) are derived by maximizing the empirical likelihood function, we proceed as follows:

- Construct empirical likelihood functions corresponding to the continuous likelihood functions $\ell_{\text{base}}^{\text{LB}}$ and $\ell_{\text{exci}}^{\text{LB}}$.
- Maximize these empirical log-likelihoods to determine the weights.
- Analyze the asymptotic properties of the resulting weights and algorithm.

Weights

We derive the following expressions for the weights:

$$w_p^k = \frac{\sum_{r=1}^{R} \sum_{i=1}^{n_r} z_{i0}^{rk} [\![t_i^r = t_{(p)}]\!]}{\sum_{r=1}^{R} [\![t_c^r \geq t_{(p)}]\!]} \tag{12}$$

$$v_q^k = \frac{\sum_{r=1}^{R} \sum_{i=2}^{n_r} \sum_{j=1}^{i-1} z_{ij}^{rk} [\![t_i^r - t_j^r = x_{(q)}]\!]}{\sum_{r=1}^{R} \sum_{i=1}^{n_r} [\![t_c^r - t_i^r \geq x_{(q)}]\!]} \tag{13}$$

Asymptotically Properties:

- w_p^k is an asymptotically consistent estimator of $\mu^k(t_{(p)})\Delta_p$.
- v_q^k is an asymptotically consistent estimator of $\phi^k(x_{(q)})\Delta_q$.
- The proposed algorithm asymptotically converges to the true intensity functions.

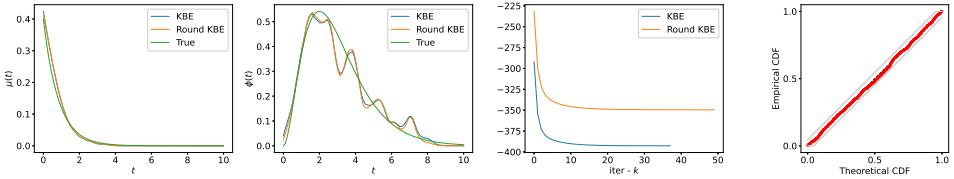
Aggregated Data Handling

Real-world data is often recorded in binned or aggregated form, where exact timing of individual events within a bin is unknown.

Proposed Nonparametric Algorithm:

- Handles aggregated data without modifications.
- The derivation of weights w_p^k and v_q^k remains valid.
- Asymptotic analysis not applicable; can't assume $\Delta_p \to 0$ as $R \to \infty$.
- Experiments with aggregated data show strong results

Experiments - Synthetic Data



(a) base (b) excitatory (c) $-llg$ (d) GoF

Figure: The training dataset with exact timestamps has 1,262 events, yielding $P = 1262$ and $Q = 26045$ unique time points for the base and excitatory processes. Rounding to one decimal place reduces these to $P = 100$ and $Q = 321$. For both exact and rounded data, the optimal bandwidths, maximizing validation likelihood, are 1.0 for the base kernel and 0.3 for the excitatory kernel.

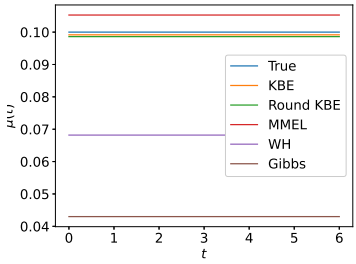
Experiments - Synthetic Data

FLORIDA TECH

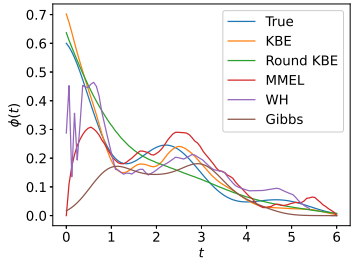
We compare our algorithm with the following methods:

- (Zhou et al., 2013) **MMEL:** EM & Solving ODE
- (Bacry and Muzy, 2016) **WH:** Moment matching, Wiener-Hopf equations estimation.
- (Zhou et al., 2020) **Gibbs:** EM & Bayesian inference with Gibb sampling

Experiments - Synthetic Data



(a)



(b)

Figure: Comparison with existing works on the Hawkes process, where the base intensity is constant at 0.1 and the excitatory function has a known support of $[0, 6]$. Panel (a) presents the estimated base intensity, while panel (b) illustrates the estimated excitatory function. The optimal bandwidth for our algorithm is 0.3 using exact data and 1.0 for rounded data.

Model Evaluation on Real-World Data

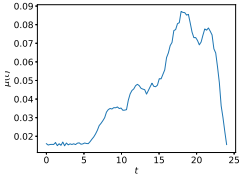
The NYPD Motor Vehicle Collision dataset (07/01/2017 – 08/31/2017)

- Accident timestamps recorded over single days.
- Data is split 50% for training and 50% for validation.

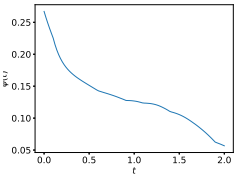
We investigate:

- (i) Base intensity: when accidents are likely.
- (ii) Excitatory intensity: if one accident increases the chance of others.

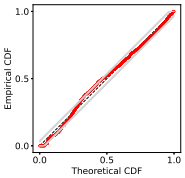
Real-World Case Study



(a) base



(b) excitatory



(c) GoF

Figure: Nonparametric estimation of the NYPD Motor Vehicle Collision data, modeled as a self-exciting Hawkes process. Panel (a) presents the estimated base intensity, while Panel (b) shows the estimated excitatory intensity. The optimal bandwidths, determined via validation, are 1.3 for the base intensity and 0.9 for the excitatory intensity. Panel (c) displays the PP plot, which assesses the model's goodness-of-fit on the validation dataset.

Summary

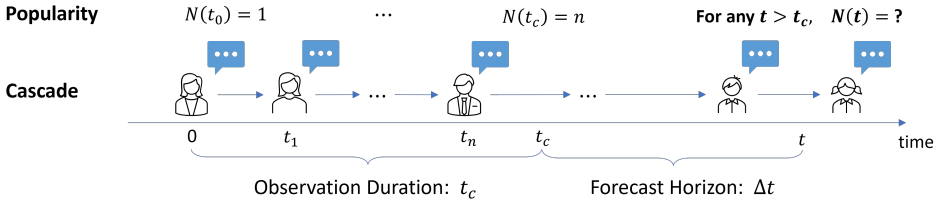
We introduce a nonparametric, kernel-based intensity estimation algorithm for Hawkes processes.

- Simple, intuitive and flexible
- Rigorously derived kernel weights.
- Established asymptotic properties.
- Capability to handle aggregated data.
- Natural extension to multi-variate Hawkes processes.

Contents

Popularity Prediction

Given an information cascade, whose initial progression has been observed, accurately predicting its total number of messages at any future time.



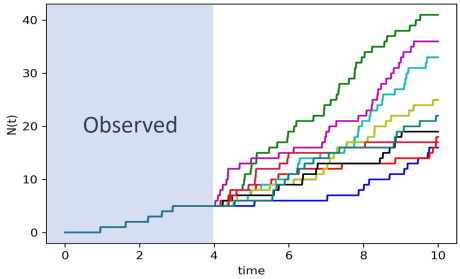
Point Process Modeling Approach

Point process models, especially the marked Hawkes process modes, have been widely adopted for popularity prediction.

- **Formulation**: Specify the cascade's diffusion dynamic as marked Hawkes process by specify its intensity function.
- **Learning**: Model parameters trained by likelihood maximization.
- **Prediction**: Adopt conditional mean counts of the learned model to predict.

Point Process Modeling Approach

- Conditional mean count at time t, is the mean value of the counting distribution at time t conditioned on the observed history.
- Unfortunately, no trackable solution for general marked Hawkes process, YET.



Point Process Modeling Approach

Compare to feature-based and deep learning discriminative models, the point process generative models:

- Light-weight training
- Inherently interpretable

BUT:

- lackluster prediction performances
- Costly simulation or problem specific estimation for prediction.

Theoretical Objective

Derive closed-form expressions for conditional mean and variance of event counts.

With Conditional Mean:

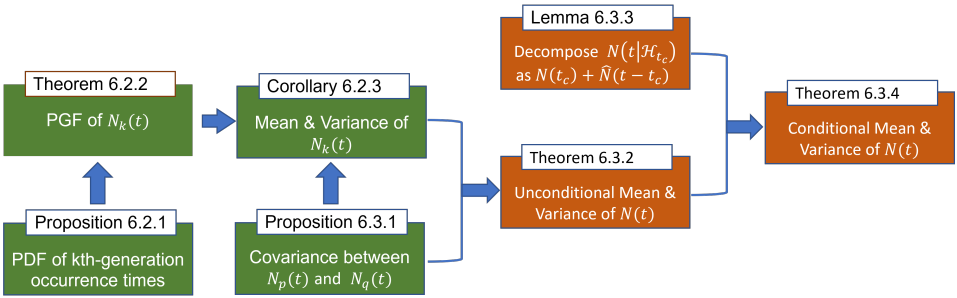
- Enable direct prediction from the derived expression.
- Support predictive learning of future event counts, in contract to learning via likelihood maximization.

With Conditional Variance:

- Establish Chebyshev-based prediction intervals.

Conditional Moments of Hawkes Process

A Hawkes process $N(t)$ can be equivalently viewed as a branching process, such that $N(t) = \sum_{k \geq 0} N_k(t)$, where $N_k(t)$ is the k-th generation counting process.



- **Formulation**: Specify the conditional intensity function.
- **Learning**: Minimize

$$L(\boldsymbol{\theta}|\mathcal{H}_{t_c}) \triangleq \frac{1}{|\mathcal{S}(t_c)|} \sum\sum_{(i,j)\in\mathcal{S}(t_c)} \left(\left(\mathbb{E}[N(t_j|\mathcal{H}_{t_i};\boldsymbol{\theta})] - j \right)^2 \right) \tag{14}$$

  where $\mathcal{S}(t_c) \triangleq \{(i,j) : 0 < t_i < t_j \leq t_c\}$ and $|\mathcal{S}(t_c)|$ is its cardinality. Here
  - $\mathbb{E}[N(t_j|\mathcal{H}_{t_i};\boldsymbol{\theta})]$:Predicted count at time t_j, given observation up to t_i
  - j: True count at time t_j
- **Prediction**: Adopt conditional mean counts of the learned model to predict.

Experiments - Synthetic Data

Objective: Evaluate Predictive vs. Generative Learning Approaches

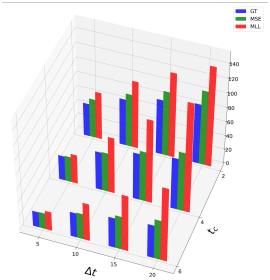


Figure: Comparison of prediction performance, the average APE% (lower better) on Synthetic Dataset

- GT: ground truth models
- MSE: models trained by minimizing the overall loss in Eq (14) – the predictive learning approach.
- MLL: models trained by maximizing likelihood – the generative learning approach.

Formulation:

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

- $\alpha > 0$: can be regarded as the "quality" of the tweet.
- $\beta > 0$: describes how fast a retweet's influence on other users fades away with time.
- The mark value m_i, number of followers of the poster, is interpreted as the strength of a user's influence. This strength is regulated by a parameter $0 < \kappa < 1$

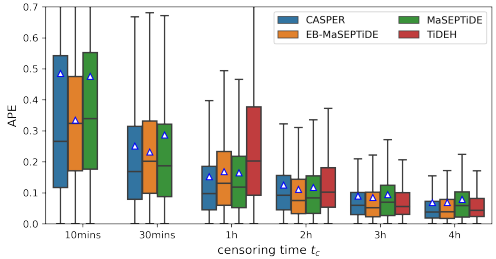
Comparison Models

FLORIDA TECH

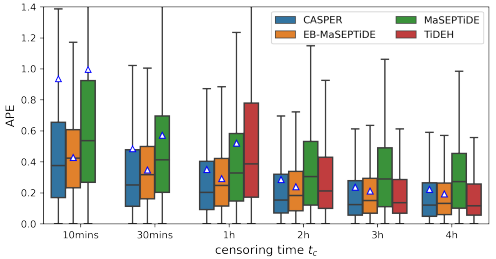
We compare our model with three state-of-the-art point process-based models:

- **TiDeH** Kobayashi and Lambiotte (2016): Models retweeting as a Hawkes process with circadian patterns, optimizing parameters on fully observed cascades and predicting future retweets through self-consistent equations.
- **MaSEPTiDE** Chen and Tan (2018): A Hawkes-based model with time-varying background intensity, trained via likelihood maximization without needing fully observed cascades.
- **EB-MaSEPTiDE** Tan and Chen (2021): Extends MaSEPTiDE using empirical Bayes to enhance early-stage predictions, incorporating additional fully observed cascades for parameter estimation.

Comparison Models



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



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

Figure: Boxplots of short- and long-term prediction APE values between our and comparison models on Twitter data, with various censoring times and prediction intervals. 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

Summary

- For Marked Hawkes Point Process (MHPP) with arbitrary, Lebesgue-integrable conditional intensity function and unpredictable marks, we derive closed-form expressions for the conditional mean and variance of its counting process.
- For anytime popularity prediction, we propose CAscade Size Prediction for self-Exciting processes via Regression (CASPER), a Hawkes process based predictive model, which is optimized to minimize the prediction error directly, rather than to maximize the generative likelihood value.

Contents

Contributions Overview

- **Generalized Time-Rescaling Theorem:**
  - Extend the original theorem to accommodate terminating processes and incomplete observations, enhancing model validation across a wider range of applications.
- **Nonparametric Estimation of Hawkes Processes:**
  - Introduce a simple and effective kernel-based intensity estimation algorithm tailored for Hawkes processes.
- **Predictive Modeling with Hawkes Processes:**
  - Derive closed-form expressions for the first- and second-order moments of counting processes and propose a discriminative learning approach for improved predictions.

Through these contributions, we aim to advance the theoretical understanding of Hawkes processes and broaden their applicability to diverse real-world challenges.

Related Publications

- **Xi Zhang**, A. Aravamudan, G.C. Anagnostopoulos. *A Non-parametric Kernel-based Intensity Estimation Algorithm for Hawkes Processes*. Submitted to AISTATS 2025. Acceptance rate: ∼30%. **Under review**.
- **Xi Zhang**, A. Aravamudan, G.C. Anagnostopoulos. *A Generalized Time Rescaling Theorem for Temporal Point Processes*. Submitted to Neural Computation. Impact Factor: 3.28. **Under review**.
- A. Aravamudan, **Xi Zhang**, G.C. Anagnostopoulos. *Anytime user engagement prediction in information cascades*. AAAI 2023. Acceptance rate: 19.6%. DOI.
- **Xi Zhang**, A. Aravamudan, G.C. Anagnostopoulos. *Anytime information cascade popularity prediction via self-exciting processes*. ICML 2022. Acceptance rate: 21.9%. URL.
- A. Aravamudan, **Xi Zhang**, J. Song, S.M. Fiore, G.C. Anagnostopoulos. *Influence dynamics among narratives. Social, Cultural, and Behavioral Modeling*, Springer, 2021. Acceptance rate: 57%. DOI.

Acknowledgment

Xi Zhang acknowledges partial support from National Aeronautics and Space Administration Grant No. 80NSSC23K0500, Defense Threat Reduction Agency Grant No. HDTRA1-22-C-0005, Defense Advanced Research Projects Agency Grant No. FA8650-18-C-7823 and Nation-al Science Foundation Grant No. CNS-1200552

Thank You!

References

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

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

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition.

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

Kobayashi, R. and Lambiotte, R. (2016). Tideh: Time-dependent hawkes process for predicting retweet dynamics. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1).

References

Tan, W. H. and Chen, F. (2021). Predicting the popularity of tweets using internal and external knowledge: an empirical bayes type approach. *AStA Advances in Statistical Analysis*.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. (2020). Efficient inference for nonparametric hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional hawkes processes. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1301–III–1309. JMLR.org.

Poisson Process

A Poisson process is characterized by a history-independent intensity function, meaning the event rate of future events is unaffected by past occurrences.

Homogeneous Poisson Process:

- Feature a constant rate, $\lambda^*(t) = \lambda$.
- Ideal for modeling events that occur at a steady average rate over time.
- *Example:* Telephone calls arriving at a switchboard at a steady average rate.

Inhomogeneous Poisson Process:

- Feature a time-varying rate, $\lambda^*(t) = \lambda(t)$.
- Suitable for modeling events with rates that vary based on external factors or time.
- *Example:* Traffic flow with higher activity during peak hours or special events.

Next Event Time Distribution

Consider $i-1$ events have been observed at times $t_1, \ldots, t_{i-1}$.

- T_i, the random variable (RV) indicating the i-th event time, is a mixture of a continuous RV on (t_{i-1}, ∞) and a discrete RV at $\{\infty\}$

$$f_{T_i}(t|\mathcal{H}_{t_{i-1}}) \triangleq \frac{\mathbb{P}\{T_i \in [t, t+dt), T_i < \infty | \mathcal{H}_{t_{i-1}}\}}{dt}, \quad t \in (t_{i-1}, \infty) \tag{16}$$

$$p_{T_i}(\infty|\mathcal{H}_{t_{i-1}}) \triangleq \mathbb{P}\{T_i = \infty | \mathcal{H}_{t_{i-1}}\} = 1 - \lim_{t\to\infty} \int_{t_{i-1}}^{t} f_{T_i}(s|\mathcal{H}_{t_{i-1}})\, ds \tag{17}$$

- Function $f_{T_i}(t|\mathcal{H}_{t_{i-1}})$ fully characterizes the distribution of the i-th event.

Next Event Time Distribution

For $t \in (t_{i-1}, \infty)$

$$F_{T_i}(t|\mathcal{H}_{t_{i-1}}) \triangleq \mathbb{P}\{T_i \leq t, T_i < \infty | \mathcal{H}_{t_{i-1}}\} = \int_{t_{i-1}}^{t} f_{T_i}(s|\mathcal{H}_{t_{i-1}}) ds \tag{18}$$

$$S_{T_i}(t|\mathcal{H}_{t_{i-1}}) \triangleq \mathbb{P}\{T_i \notin (t_{i-1}, t) | \mathcal{H}_{t_{i-1}}\} = 1 - F_{T_i}(t|\mathcal{H}_{t_{i-1}}) \tag{19}$$

Define $\pi_i \triangleq \lim_{t\to\infty} \int_{t_{i-1}}^{t} f_{T_i}(s|\mathcal{H}_{t_{i-1}})\, ds$, notice, as $t \to \infty$, $F_{T_i}(t|\mathcal{H}_{t_{i-1}}) \to \pi_i$ and $S_{T_i}(t|\mathcal{H}_{t_{i-1}}) \to 1 - \pi_i$.

Conditional Intensity Function

The conditional intensity function $\lambda^*(t) \triangleq \lambda(t|\mathcal{H}_{t^-})$

$$\lambda^*(t) \triangleq \lim_{\Delta t \downarrow 0} \frac{\mathbb{P}\{N[t, t+\Delta t) = 1|\mathcal{H}_{t^-}\}}{\Delta t} \tag{20}$$

which represents the instantaneous rate of an event occurring at time t given the history right before t.

The cumulative conditional intensity function $\Lambda^*(t) \triangleq \Lambda(t|\mathcal{H}_{t^-})$,

$$\Lambda^*(t) = \int_0^t \lambda(\tau|\mathcal{H}_{t^-})\, d\tau, \tag{21}$$

which quantifies the cumulative expected number of events up to time t, conditioned on the process's history. Specifically, it satisfies $\Lambda^*(t) = \mathbb{E}[N(t)|\mathcal{H}_{t^-}]$

Relationships

$$f_{T_i}(t|\mathcal{H}_{t_{i-1}}) = \lambda^*(t) S_{T_i}(t|\mathcal{H}_{t_{i-1}}) \tag{22}$$

$$S_{T_i}(t|\mathcal{H}_{t_{i-1}}) = \exp\left(-\int_{t_{i-1}}^{t} \lambda^*(t) d\tau\right) \tag{23}$$

Recall $\pi_i \triangleq \lim_{t\to\infty} \int_{t_{i-1}}^{t} f_{T_i}(s|\mathcal{H}_{t_{i-1}})\, ds$

- $\pi_i = 1$: The i-th event occurs a.s., $f_{T_i}(t|\mathcal{H}_{t_{i-1}})$ is a proper probability density function (PDF), and $\int_{t_{i-1}}^{t} \lambda^*(t) d\tau$ is unbounded.
- $\pi_i < 1$: The i-th event has a nonzero probability of not occurring, $f_{T_i}(t|\mathcal{H}_{t_{i-1}})$ is an improper PDF, and $\int_{t_{i-1}}^{t} \lambda^*(t) d\tau$ is bounded.

Non-Terminating Processes

A non-terminating process keeps generates events over an infinite time horizon, meaning every realization a.s. contains an infinite number of events.

- For all events $i \geq 1$, the i-th event's distribution function is proper.
- $\Lambda^*(t) \to \infty$ as $t \to \infty$ for all possible histories. An unbounded $\Lambda^*(t)$ indicates indefinite event generation, with no history resulting in a finite $\Lambda^*(t)$.

Terminating Processes

A terminating process may have realizations with a finite number of events, and a.s. has a finite number of events in some cases.

- There exists at least one $i \geq 1$ for which the i-th event's distribution function is improper.
- $\Lambda^*(t)$ is bounded for at least one possible history, indicating that the process may stop after a finite number of events.

Marked Temporal Point Process

Definition: A marked temporal point process extends a standard temporal point process by assigning each event time a stochastic *mark* from a specified *mark space* $\mathcal{M}$.

- Can be viewed as a point process on $(\mathbb{R}^+ \times \mathcal{M})$ with conditional intensity $\lambda^*(t, m)$, where $\mathbb{R}^+$ denotes non-negative real numbers.
- The marginal temporal component, or *ground process*, has ground intensity $\lambda^*(t)$ and counting process $N_g = \{t_i\}$.

Example:

- Earthquake modeling, where each event has a magnitude and location as marks.
- Social media retweet modeling, each retweet event includes the follower count of the user posting it.

Types of Marks:

- **Independent Marks:** Given $N_g = \{t_i\}$, the marks $\{m_i\}$ are mutually independent and depend only on their corresponding times t_i.
- **Unpredictable Marks:** Each mark's distribution at t_i is independent of previous times and marks (t_j, m_j) where $t_j < t_i$.

Under conditions of independent or unpredictable marks, the conditional intensity $\lambda^*(t, m)$ factorizes into the form

$$\lambda^*(t, m) = \lambda^*(t) g(m|t), \tag{24}$$

where $\lambda^*(t)$ is the ground intensity and $g(m|t)$ is the mark distribution given time t.

Branching Process Equivalence

A Hawkes process can be viewed as a *branching process* with events generated through two mechanisms:

- *Immigrant Events (0-th Generation):* Occur independently, drawn from a Poisson process with base intensity $\mu(\cdot)$.
- *Offspring Events (1st Generation and Beyond):* Each immigrant event, along with existing offspring, produces new offspring that follow a Poisson process with intensity $\phi(t - t_i)$, based on the time since the parent event at t_i. This process continues recursively.

Weights Interpretation

Through declustering, each Hawkes realization is decomposed into :

- One realization from base Poisson process, starting at $t = 0$ and censored at $t = t_c$.
- n_r realization from excitatory Poisson processes, each starting at t_i and censored at t_c.

The weights can be interpreted as:

- w_p^k from Eq. (12) represents the mean event count in $(t_{(p-1)}, t_{(p)})$ across the R base Poisson realizations.
- v_q^k from Eq. (13) represents the mean event count in $(x_{(q-1)}, x_{(q)})$ across $\sum_{r=1}^{R} n^r$ excitatory Poisson realizations.

Existing Approaches

FLORIDA TECH

Discriminative: design to predict

Categorized into:

- Feature-based: Explainable, but requires laborious manual feature extraction.
- Deep learning: : Lean effective features, but are opaque to model interpretation.

Satisfactory prediction performance, but

- Require an abundance of data and computing effort to support model training and hyper-parameter tuning
- For anytime popularity prediction, require training a distinct model for every desired $(t_c, \Delta t)$.

Generative: design to describe

Advantageous models:

- Hawkes point process models
  - self-exciting property explains well the "rich-get-richer" phenomena.

Light-weight training and inherently interpretable, but

- lackluster prediction performances
- Costly simulation or problem specific estimation for prediction.

Contribution 1: Generalized Time-Rescaling Theorem

We extend the time-rescaling theorem, traditionally limited to non-terminating processes with complete observations, to accommodate terminating processes and incomplete observations.

This extension enables more robust model validation across diverse contexts, particularly for Hawkes processes modeling real-world event dynamics.

Contribution 2: Nonparametric Estimation of Hawkes

We propose a nonparametric, kernel-based intensity estimation algorithm specifically designed for Hawkes processes.

The algorithm iteratively updates declustering probabilities and estimates the intensity function using rigorously derived kernel weights and established asymptotic properties, offering simplicity and flexibility for various real-world applications.

Contribution 3: Adopt Hawkes Process to Predict

FLORIDA TECH

We analyze the first- and second-order moments of counting processes associated with Hawkes processes, providing a closed-form expression for conditional mean counts.

This facilitates predictive learning of future event counts, resulting in models that outperform those based on Hawkes processes fitted via maximum likelihood estimation.