Learning to Select Exogenous Events for Marked Temporal Point Processes

Ping Zhang[*], **Rishabh Iyer**[*], **Ashish Tendulkar**[†], **Gaurav Aggarwal**[†] and **Abir De**[§]
[*]UT Dallas, [†]Google, § IIT Bombay.
{Ping.Zhang,rishabh.iyer}@utdallas.edu,
{ashishvt,gauravaggarwal}@google.com, abir@cse.iitb.ac.in

Abstract

Marked temporal point processes (MTPPs) have emerged as a powerful modeling tool for a wide variety of applications which are characterized using discrete events localized in continuous time. In this context, the events are of two types— *endogenous events* which occur due to the influence of the previous events and *exogenous events* which occur due to the effect of the externalities. However, in practice, the events do not come with endogenous or exogenous labels. To this end, our goal in this paper is to identify the set of exogenous events from a set of unlabeled events. To do so, we first formulate the parameter estimation problem in conjunction with exogenous event set selection problem and show that this problem is NP hard. Next, we prove that the underlying objective is a monotone and α-submodular set function, with respect to the candidate set of exogenous events. Such a characterization subsequently allows us to use a stochastic greedy algorithm which was originally proposed in [63] for submodular maximization. However, we show that it also admits an approximation guarantee for maximizing α-submodular set function, even when the learning algorithm provides an imperfect estimates of the trained parameters. Finally, our experiments with synthetic and real data show that our method performs better than the existing approaches built upon superposition of endogenous and exogenous MTPPs.

1 Introduction

In recent years, marked temporal point processes (MTPPs) have shown tremendous potential in modeling the arrival process of the asynchronous events in continuous time [65, 45, 64, 17, 22, 16, 51, 80, 33, 12, 59, 81, 73, 48, 75, 27, 52, 46]. They are extensively used in a wide variety of applications, *e.g.*, information diffusion in social networks [64, 75, 17, 73, 42, 15], location modeling [81], epidemic forecasting [48, 64], etc. At the outset, an MTPP event consists of two elements— its arrival time and the mark, where the latter encodes the category of the event, *e.g.*, the topic of a blog post, the sentiment associated with a tweet, disease of an individual, *etc*.

In the context of learning an MTPP, we observe two types of events— *endogenous events* which occur due to influence of the previously observed events and, *exogenous events* which are driven by the external sources or unobserved events, rather than being influenced by the previously observed events. For example, a post in social media may be influenced by some external news source, rather than the previous messages [26, 59, 12]; in epidemic process, an infection may be triggered by community transmission where the source is other than any previously known infections [49], etc. In such applications, the observed events appear unlabeled, *i.e.*, they are not tagged with their sources of influence. Consequently, the existing MTPP models [21, 15, 82, 76, 64] resort to the superposition of exogenous and endogenous MTPPs. However, such approaches view an event being exogenous or endogenous with equal likelihood and therefore, they do not focus on identifying which events are more likely to be exogenous than others. Consequently, they often remain oblivious to the presence of the exogenous events, which often preclude these models from realizing their full potential. While De et al. [12] tackle this problem in the particular context of opinion dynamics in social networks,

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

their formulation focus only on the mark dynamics and, skirt the effect of exogenous events on the arrival times. As a result, their framework can not be extended to a generic MTPP model, where the arrival time is often a key component.

1.1 Our contributions

In order to respond to the above limitations, we make the following contributions.

Learning in conjunction with selection of exogenous events. We design TPP-SELECT, a novel event selection algorithm which, for any given MTPP model, simultaneously filters the exogenous events from a stream of unlabeled events as well as train the model parameters over the remaining (endogenous) events (Section 3). More specifically, we seek to maximize the likelihood of all the events, jointly with respect to the set of exogenous events and the parameters of the associated MTPP model. Such a formulation induces a parameter estimation problem in conjunction with the combinatorial optimization problem for selecting the exogenous events.

Characterization of the objective. Subsequently, we show that our problem is NP-hard due to the involvement of the candidate exogenous events as an optimization variable. We tackle this problem using several technical innovations. First, we reformulate it as an instance of a set function maximization problem, where the objective corresponds to the value of the maximum likelihood, as a function of the candidate set of exogenous events. Then, based on this representation, we show that this objective satisfies monotonicity and α-submodularity— a notion of approximate submodularity [25, 43, 29] for linear MTPP and a wide variety of nonlinear MTPPs (Section 4). These results reveal an unexplored connection between approximate submodularity and temporal point process, that is of independent interest.

Approximation guarantee for greedy method in the face of imperfect training. Next, we show that these properties allow us to use an existing stochastic greedy algorithm [63]. While this algorithm is known to admit an approximation guarantee for submodular maximization, we show that it also enjoys an approximation guarantee for maximizing α-submodular functions. Moreover, we show that this algorithm also admits an approximation guarantee even when the training algorithm provides an imperfect estimation of the MTPP parameters.

Finally, we perform a comprehensive evaluation of our proposal on both synthetic and real data. Our experiments with synthetically generated sequences show that TPP-SELECT can accurately select the exogenous events from a stream of unlabeled events, even though it is completely unsupervised. Our experiments on real datasets show that, despite training on a subset (endogenous) of events, TPP-SELECT outperforms several baselines as well as the corresponding model which is trained on the entire dataset in the presence of exogenous events. Our code is available in https://github.com/noilreed/TPP-Select.

2 Related work

Apart from the works on MTPPs, our work is related to robust learning, subset selection, modeling external effects in different applications, active learning, etc.

Robust learning. Robust learning focuses on training machine learning models in the presence of outliers and noisy labels [9, 66, 3, 23, 85, 4, 57]. However, exogenous events are not outlier data, they are events which follow a different model, other than the central model— which, in our case, captures the endogenous event dynamics. These robust learning methods predominantly consider i.i.d. data and focus on designing models and training algorithms which would generalize well across the data including the outliers. In contrast, the events in our setup have strong temporal dependencies. Moreover, our objective is to select the exogenous events, so that the model can be trained on the remaining (endogenous) events to obtain more accurate parameter estimates

Data subset selection. In recent years, there is a flurry of work on data selection [78, 79, 77, 47, 2, 39, 50, 56, 7, 6, 36], which predominantly select a smaller sized subset of data to facilitate efficient training, by optimizing submodular functions and their variants. However, TPP-SELECT aims to filter the exogenous events from a sequence of events to help the underlying MTPP model achieve better predictive performance.

Modeling external effects. Our work is also connected to the methods that model the external effect in several applications, especially, in web or online social media [19, 59] that aim to model the effect of exogenous influence on the information diffusion in social media. However, they do not aim to select the exogenous messages from a stream of unlabeled messages. [34] aims to learn the effect of one user on the post of other users. Such a model can be used to quantify the external effect. The

proposal in [44] aims to learn a probabilistic model which captures latent influence present in data using mutually exciting processes combined with a random graph model. Furthermore it infers the latent network structure from the noisy observations. [74] models dynamics of event sequence in the particular application of purchase data modeling. Specifically, it combines three factors, namely user preferences, influence from other users and media advertisements. These proposals learn their model using EM algorithm combined with Gibbs sampling method. However, they do not provide an algorithm to explicitly demarcate between endogenous and exogenous events.

Optimization of submodular functions and their variants. Submodular functions have the ability to characterize a wide variety of set functions. Moreover, thanks to the existence of a simple greedy algorithm to maximize such functions, their use is widespread across several subset selection tasks [78, 79, 77, 47, 2]. Several popular set functions, *e.g.*, facility location, set cover, concave over modular, etc. are submodular [36, 2, 35]. In this context, there is a recent line of works which focus on maximizing different variants of submodular functions which include α-submodular functions and γ-weakly submodular functions [60, 53, 5, 11, 37]. In particular, our work is more focused around α-submodular functions.

Subset selection in conjunction with model learning. Existing works [7, 50, 36, 38, 13, 14, 56] have also considered subset selection in conjunction with model parameter estimation. However, in the context of applications as well as formulations, these works differ from us. Killamsetty et al. [38] consider simultaneous parameter estimation on training set and subset selection using validation error in order to facilitate efficient learning without introducing significant error. On the other hand, the works in [36, 50, 56, 7] focus on selecting coreset— a small representative subset of the entire data— together with parameter estimation. However, they consider i.i.d data and select subset for efficient training. In contrast, we partition sequence of events into exogenous and endogenous events for improved training of point process models. Finally, the works in [13, 14] consider simultaneous parameter estimation and subset selection. Similar to our work, they aim to find model parameters and a desired data subset simultaneously by minimizing the underlying loss functions with respect these variables. However, they focus on human assisted machine learning, whereas our work considers exogenous event selection in the context of marked temporal point processes. Thus the problem formulation, results and proofs are significantly different from these works.

Active learning. Active learning methods [78, 28, 68, 41, 24] aim to identify instances in the training data, in order to reduce the labelling cost. In contrast, our work aims to select samples which are not influenced by the previous events in a sequence of dependent observations. To that end, it trains the endogenous MTPP model on the rest of the events. More specifically, active learning methods aim to find a subset of instances so that the model trained over it gives accurate predictions for *all the samples* in the test set. In contrast, our work selects both exogenous and endogenous events and then trains two separate models $(\lambda_{\boldsymbol{\theta}}, m_{\boldsymbol{\theta}})$ and (μ, q) so that they perform well specifically on endogenous and exogenous events respectively.

3 Problem formulation

In this section, we formulate our problem of learning an MTPP in presence of exogenous events, starting with an overview of MTPPs.

3.1 Overview of marked temporal point processes

A marked temporal point process is a stochastic process which is characterized using a sequence of events $\mathcal{H} = \{e = (y, t)\}$, where $t \in \mathbb{R}^+$ is the time of occurrence and $y \in \mathcal{C} = \{c\}$ is mark of the event that has happened at time t [16, 15, 65, 10]. The dynamics of the arrival times can equivalently be represented as a counting process $N(t)$ that counts the number of events occurred until time t. In this context we also define $\mathcal{H}_t$ as the history of events happened until and excluding time t, *i.e.* $\mathcal{H}_t = \{e_i = (y_i, t_i) | e_i \in \mathcal{H}, t_{i-1} < t_i < t\}$. We specify the dynamics of the arrival times using a conditional intensity function $\lambda(t)$ which in turn encapsulates the probability that an event occurs in the infinitesimal time interval $[t, t + dt)$, described as follows [15, 73, 10]:

$$\lambda(t) = \Pr\left(\text{An event will occur between } t \text{ and } t + dt\right) = \mathbb{E}[dN(t) = 1], \tag{1}$$

where $dN(t) \in \{0, 1\}$ is the number of events arriving in the infinitesimal time-interval $[t, t + dt)$. On the other hand, we capture the behavior of marks using a probability distribution, *i.e.*

$$m(c) = \Pr\left(\text{Mark of an event is } c\right) = \Pr\left(y = c\right), \tag{2}$$

which is often modeled as a multinomial distribution over the set of marks $\mathcal{C}$.

3.2 Problem setup

Endogenous and exogenous events. In the context of MTPP, we observe two types of events, viz., (i) endogenous events that occur due to the influence of the previously observed events; and, (ii) exogenous events that are *not* influenced by the previously observed events but occur due to the influence of the externalities [21, 12]. The dynamics of the endogenous events are modeled using a parameterized conditional intensity function $\lambda(t) = \lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t)$ and conditional mark distribution $m(y) = m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t)$. The arrival time and the mark depend on the previously observed events $\mathcal{H}_t$,

$$\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t) = \mathbb{E}(dN(t) = 1 \,|\, \mathcal{H}_t), \quad m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t) = \Pr(y \,|\, \mathcal{H}_t). \tag{3}$$

Therefore, both $\lambda_{\boldsymbol{\theta}}$ and $m_{\boldsymbol{\theta}}$ are stochastic quantities. On the other hand, exogenous events are captured mostly using an independent MTPP in which, the intensity function or the mark distribution are modeled using a deterministic and mostly time invariant distributions. In our work, we model the intensity function and the mark distribution for exogenous events as Gamma and Beta noise,

$$\lambda(t) = \mu \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda), \qquad m(y) = q \sim \text{Beta}(\alpha_{m,y}, \beta_{m,y}), . \tag{4}$$

These distributions allows $\mu \in (0, \infty)$ and $q \in (0, 1)$. $\alpha_\bullet$ and $\beta_\bullet$ are hyperparameters. We further assume that both the arrival times and the marks of the exogenous events are *i.i.d.* random variables.

Use of superposition of MTPP models and their limitations. Suppose, if at all possible, the training observations $\mathcal{H}_T$ gathered during the time-window $[0, T)$ were already tagged with endogenous or exogenous labels. In such a case, one could first build the set of exogenous events $\mathcal{S}$ and the endogenous events $\mathcal{H}_T \backslash \mathcal{S}$ and then, use maximum likelihood estimation to learn $\{\alpha_\bullet, \beta_\bullet\}$ and $\boldsymbol{\theta}$ separately using the set of exogenous events $\mathcal{S}$ and the endogenous events $\mathcal{H}_T \backslash \mathcal{S}$, respectively.

In practice, the difference between the exogenous and the endogenous events is not evident in the training data. Specifically, these events are not tagged with *endogenous* or *exogenous* labels and therefore, it is not possible to construct the set of exogenous events $\mathcal{S}$, well in advance. Existing works [21, 15, 82, 75] tackle this problem by maximizing the likelihood corresponding to the superposition of exogenous and endogenous MTPPs. In such case, the intensity function becomes $\lambda(t) = \mu + \lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t)$ and the mark distribution incorporates the effect of exogenous events using additive offsets.

Maximizing the likelihood associated with the superposition of two MTPPs, often turns out to be suboptimal as shown in our experiments (Section 6). This is because, such an approach does not aim to find out the set of exogenous events $\mathcal{S}$ which maximizes likelihood function. In fact, due to the absence of any informed choice about $\mathcal{S}$, such a likelihood function assumes that all the partitions between exogenous and endogenous events $\{(\mathcal{S}, \mathcal{H}_T \backslash \mathcal{S})\}$ are equally likely (Appendix B.1). Such an assumption may not be true in general. We ameliorate this limitation by selecting partition $\{(\mathcal{S}, \mathcal{H}_T \backslash \mathcal{S})\}$ along with the model parameters, which induces the highest value of the likelihood, across all possible partitions and the hypothesis class.

3.3 Problem statement

We aim to select the exogenous events from a given stream of unlabeled events and simultaneously train these MTPP models (3) on the remaining events which are endogenous in nature.

Selecting exogenous events in conjunction with parameter estimation. Given a stream of events $\mathcal{H}_T$ recorded in the time interval $[0, T)$, we aim to pick $\mathcal{S} \subseteq \mathcal{H}_T$— the candidate set of k exogenous events, *i.e.*, $|\mathcal{S}| \leq k$— and concurrently, train the parameters $\boldsymbol{\theta}$ on the set of the endogenous events $\mathcal{H}_T \backslash \mathcal{S}$. Using the intensity functions and the mark distributions defined in Eq. (3)— (4), our goal is to solve the following regularized log-likelihood maximization problem:

$$\underset{\boldsymbol{\theta} \in \Theta, \mathcal{S} \subseteq \mathcal{H}_T}{\text{maximize}} \quad \mathcal{L}(\boldsymbol{\theta}; \mathcal{S}, \mathcal{H}_T), \qquad \text{subject to, } |\mathcal{S}| \leq k, \tag{5}$$

$$\text{with, } \mathcal{L}(\boldsymbol{\theta}; \mathcal{S}, \mathcal{H}_T) = \sum_{e_i \in \mathcal{H}_T \backslash \mathcal{S}} \Big[-\rho \, \|\boldsymbol{\theta}\|_2^2 + \log(\lambda_{\boldsymbol{\theta}}(t_i \,|\, \mathcal{H}_{t_i})) + \log m_{\boldsymbol{\theta}}(y_i \,|\, \mathcal{H}_{t_i}) \Big] - \int_0^T \lambda_{\boldsymbol{\theta}}(\tau \,|\, \mathcal{H}_\tau) \, d\tau$$
$$+ |\mathcal{S}| \mathbb{E}[\log \mu] - T \mathbb{E}[\mu] + |\mathcal{S}| \mathbb{E}[\log q]. \tag{6}$$

Here, ρ is the coefficient of the L_2 regularizer, k is a pre-specified integer with $k \leq |\mathcal{H}_T|$. The above problem poses a subset selection task in conjunction with likelihood maximization. Similar problems have also been addressed in [14, 13, 18] Next, we derive an equivalent representation of our optimization problem (6) as an instance of cardinality constrained set function maximization problem.

Reformulating (6) **as a set function optimization problem.** Given a fixed set of the exogenous events $\mathcal{S}$, the optimal value of the model parameters $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})$ becomes dependent only on the endogenous events $\mathcal{H}_T \backslash \mathcal{S}$. Therefore, if we define:

$$F(\mathcal{S}) = \mathcal{L}(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}); \mathcal{S}, \mathcal{H}_T) \tag{7}$$

then the optimization problem (6) becomes equivalent to solving the following optimization problem.

$$\underset{\mathcal{S} \subseteq \mathcal{H}_T}{\text{maximize}}\, F(\mathcal{S}), \quad \text{subject to, } |\mathcal{S}| \leq k \tag{8}$$

Hardness analysis. Given the optimal set of exogenous events $\mathcal{S}^*$, one can easily compute the optimal parameters $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}^*)$ for the problem (5) in polynomial time, if the objective (6) is concave with respect to $\boldsymbol{\theta}$ for any fixed $\mathcal{S}$. However, in general, both $\mathcal{S}^*$ and $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}^*)$ can not be solved in polynomial time. We formally state this result in the following proposition (Proven in Appendix B.2).

Proposition 1 *Solving the optimization problem in Eq.* (5) *is NP-hard.*

4 Our proposed framework for maximizing $F(\mathcal{S})$

In this section, we present our framework to maximize $F(\mathcal{S})$. More specifically, we first show that the set function $F(\mathcal{S})$ defined in Eq. (7) is monotone and α-submodular for a wide variety of MTPPs. Then, we present TPP-SELECT, an approximation algorithm based on a scalable stochastic greedy algorithm [55], which identifies the optimal set of exogenous events from a set of unlabeled events, for any given MTPP model. Finally, we show that such an algorithm enjoys an approximation guarantee even when the learning algorithm provides an imperfect estimate of the trainable parameter $\boldsymbol{\theta}$.

4.1 Characterization of $F(\mathcal{S})$

First, we show that $F(\mathcal{S})$ enjoys several useful properties for a wide variet of MTPPs, which allows us to leverage a stochastic greedy algorithm that offers approximation guarantee. Before formally stating those properties, we first present the formal definitions of monotonicity and α-submodularity.

Definition 2 *A function $F(\mathcal{S})$ is monotone non-decreasing if $F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \geq 0$ for all $e \in \mathcal{H}_T \backslash \mathcal{S}$. $F(\mathcal{S})$ is α-submodular if $F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \geq \alpha(F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}))$ for all $e \in \mathcal{H}_T \backslash \mathcal{T}$ and $\mathcal{S} \subseteq \mathcal{T} \subset \mathcal{H}_T$ [28, 83, 20].*

Monotonicity and α-submodularity of $F(\mathcal{S})$ with linear MTPPs. First, we consider a linear marked temporal point process which is built upon Hawkes process [15, 75, 64]. Here, we model the intensity function $\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t)$ and the mark distribution $m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t)$ for the endogenous events as a linear combination of multiple triggering kernels $\{\kappa_\bullet(t, t_i) > 0\}$, *i.e.*,

$$\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t) = \boldsymbol{\theta}_\lambda^\top \boldsymbol{\kappa}_\lambda(t), \qquad m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t) = \text{logistic}(y \cdot \boldsymbol{\theta}_m^\top \boldsymbol{\kappa}_m(t)) \tag{9}$$

where $\boldsymbol{\theta}_\bullet^\top \boldsymbol{\kappa}_\bullet(t) = \sum_{v=1}^d \theta_{\bullet,v} \sum_{e_i \in \mathcal{H}_t} \kappa_{\bullet,v}(t, t_i)$. Here, d is the dimension of $\boldsymbol{\kappa}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\lambda,v}, \boldsymbol{\theta}_{m,v} \,|\, v \in [d]\}$ is the set of trainable parameters. In general, $\kappa_{\bullet,v}(\cdot,\cdot)$ captures different forms of dependencies. For example, in the particular case of information diffusion in social networks, $\lambda_{\boldsymbol{\theta}}$ may correspond the rate of message of a user u with d neighbors and, $\kappa_{\bullet,v}(t, t_i) = e^{-\omega_\bullet (t - t_i)} \cdot \mathbf{1}(t_i \text{ is posted by user } v)$ which captures the scenario the effect of an previously occurred event e_i decays over time with the decay factor $\omega_\bullet > 0$. In the subsequent theorems, we formally state that $F(\mathcal{S})$ is a monotone and submodular set function (Proven in Appendix C.1).

Theorem 3 (Monotonicity of linear MTPP) *Assume that the endogenous events follow a linear MTPP* (9)*. Then the set function $F(\mathcal{S})$ defined in Eq.* (7) *is monotone in $\mathcal{S}$ if $\rho \geq \rho_{\min} = (0.5/e) \cdot \kappa_{\max,\lambda}^2 \cdot \exp\big(-2\mathbb{E}[\log(\mu \cdot q)]\big)$, where μ and q are the intensity and the mark distribution for the exogenous events and $\kappa_{\max,\lambda} = \max_{t \leq T} \|\boldsymbol{\kappa}_\lambda(t)\|_2^2$.*

We would like to hightlight that for several linear MTPP models, such an assumption on the regularization parameter ρ is quite loose. For examples, in the previously discussed example of information diffusion with exponential kernels, if we have $\kappa_{\max,\lambda} = 1$ and an exogenous event distribution $\mu, q \sim \text{Gamma}(1, 1)$, we have $\rho_{\min} < 0.1$.

Theorem 4 (α-submodularity of linear MTPP) *Given that the same conditions in Theorem 3 as well as $\left\|\int_0^\infty \boldsymbol{\kappa}_\lambda(t) dt\right\|_2 \leq K < \infty$, $\kappa_{\max,\bullet} = \max_{t \leq T} \|\boldsymbol{\kappa}_\bullet(t)\|_2$, $\kappa_{\min,\bullet} = \min_{v \in [d], t \leq T} |\kappa_{\bullet,v}(t)|$.*

If $K, \kappa_{\max,\bullet}, \kappa_{\min,\bullet} > 0$, then the set function $F(\mathcal{S})$ is α-submodular in $\mathcal{S}$ with $\alpha \geq \alpha_F$, where

$$\alpha_F = \frac{\log\left[\sqrt{2e\rho/\kappa_{\max,\lambda}^2}\right] + \mathbb{E}[\log(\mu \cdot q)]}{\frac{\kappa_{\max,\lambda}}{2\kappa_{\min,\lambda}} + \log\frac{4\kappa_{\max,\lambda}}{\kappa_{\min,\lambda}^2} + \log\left(\sqrt{\frac{2\rho\kappa_{\max,\lambda}}{\kappa_{\min,\lambda}}} + K\right) + \frac{\kappa_{\max,m}^2 \log 2}{2\rho} + \sqrt{\frac{\kappa_{\max,m}^2 \log 2}{\rho}} + \mathbb{E}[\log(\mu \cdot q)]}$$

In the previously discussed example of linear MTPP, which has an exponential kernel having decay factor ω, we have $\alpha_F^* = O(\log\rho/(\log\rho + be^{\omega T}))$, where b is a data dependent constant. Hence, by choosing a low decay factor ω, one can obtain a high value of α_F.

Proof sketches of Theorem 3 and 4: We prove Theorem 3 in two steps. In the first step, we show that $F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \geq \min_{\boldsymbol{\theta}_\lambda}[\rho\|\boldsymbol{\theta}_\lambda\|_2^2 - \log(\boldsymbol{\theta}_\lambda^\top \boldsymbol{\kappa}_\lambda(t))] + \exp\big(-2 \cdot \mathbb{E}[\log(\mu \cdot q)]\big)$. In the second step, we show that this lower bound is greater than $\mathbb{E}[\log\mu + \log q] + \log\left[\sqrt{2e\rho/\kappa_{\max}^2}\right]$ which proves the theorem. To prove Theorem 4, we aim to bound $[F(\mathcal{S} \cup \{e\}) - F(\mathcal{S})]/[F(\mathcal{T} \cup \{e\}) - F(\mathcal{T})]$. To bound the numerator, we use the result of Theorem 3. To bound the denominator, we first show that $F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}) \leq \rho\|\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))\|_2^2 - \log(\boldsymbol{\theta}_\lambda^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))^\top \boldsymbol{\kappa}_\lambda(t)) + \log[1 + \exp\big(-y \cdot \boldsymbol{\theta}_m^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))^\top \boldsymbol{\kappa}_m(t)\big)]$. These individual terms are finally bounded using minimum and maximum possible values of $\|\boldsymbol{\theta}\|_2$.

Monotonicity and α-submodularity of Nonlinear MTTPs. Next, we consider nonlinear MTPPs that do not have any fixed parameterized form, which are trained using deep learning methods [16, 84, 86]. Such models use a sequence encoder— either a recurrent neural network [16] or a transformer [84, 86]— to first derive an embedding $\boldsymbol{h}_i$ of the history of events $\mathcal{H}_{t_i}$ happened until time t_i and then use $\boldsymbol{h}_i$ to capture the generative process of the next event. Characterization of $F(\mathcal{S})$ for such a general MTPP is an extremely challenging task. Therefore, we focus on a class of nonlinear MTPPs which satisfy the following conditions on the intensity function and the mark distribution, *i.e.*,

$$\lambda_{\min} e^{-a\|\boldsymbol{\theta}\|_2} \leq \lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t) \leq \lambda_{\max} e^{a\|\boldsymbol{\theta}\|_2}, \qquad m_0(t) e^{-b\|\boldsymbol{\theta}\|_2} \leq m_{\boldsymbol{\theta}}(\cdot \,|\, \mathcal{H}_t) \leq 1. \tag{10}$$

We would like to highlight that many popular MTPP processes including recurrent marked temporal process (RMTPP) [16], self-attentive point process [84], transformer Hawkes process [86], neural Hawkes Process [51], etc. satisfy the above conditions. In fact, such an exponential bound is tight for the intensity function used in RMTPP [16], since its intensity function uses an exponential activation function in the last layer. In Appendix C.3, we discuss the representations of different MTPPs in the form defined in Eq. (10). Next, we show the monotonicity and submodularity of $F(\mathcal{S})$ in the following theorems (Proven in Appendix C.2)

Theorem 5 (Monotonicity of nonlinear MTPP) *Assume that the endogenous events follow a nonlinear MTPP with exponentially bounded intensity function, as described in Eq.* (10) *with* $\lambda_{\max} \leq \exp(-\mathbb{E}[\log(\mu \cdot q)])$. *Then the set function $F(\mathcal{S})$ defined in Eq.* (7) *is monotone in $\mathcal{S}$ if $\rho \geq 0.25\, a^2 / (\log[1/\lambda_{\max}] + \mathbb{E}[\log(\mu \cdot q)])$, where μ and q are the intensity and the mark distribution for the exogenous events.*

Theorem 6 (α-submodularity of nonlinear MTPP) *Given the same conditions in Theorem 5 and $\int_0^t \|\lambda_{\mathbf{0}}(t \,|\, \mathcal{H}_t)\|_2 dt \leq \Lambda_{\max} < \infty$, the set function $F(\mathcal{S})$ is α-submodular in $\mathcal{S}$ with $\alpha \geq \alpha_F$, where α_F is equal to*

$$\frac{\log\left[\exp\left(-a^2\mathbb{E}[\log(\mu \cdot q)]/4\right)/\lambda_{\max}\right]}{\Lambda_{\max} + (a+b)\sqrt{\frac{\log(1/\lambda_{\min} m_{\min}) + \Lambda_{\max}}{\rho}} + 2\log(1/\lambda_{\min} m_{\min}) + \mathbb{E}[\log(\mu \cdot q)]} \tag{11}$$

We would like to point out that $\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_T) = \lambda_0 > 0$ for the nonlinear MTPPs having exponential bounds. Indeed, in practice, we encounter several existing models [16, 86, 84] that choose activation functions, *e.g.*, Softplus, exponential function, having nonzero intensity functions even when the parameters $\boldsymbol{\theta} = \mathbf{0}$. However, for the linear MTPPs described in Eq. (9), $\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t) = 0$ for $\boldsymbol{\theta} = \mathbf{0}$. Therefore, the set of nonlinear MTPPs considered above does not cover the class of linear MTPPs (9) and therefore, the results and the proofs in Theorems 5- 6 are different from Theorems 3– 4. Appendix D characterizes other types of nonlinear MTPPs.

4.2 TPP-SELECT: Approximation algorithm for maximizing $F(\mathcal{S})$

Finally, we present TPP-SELECT, an approximation algorithm based on a stochastic greedy algorithm proposed in [55] in order to solve the optimization problem defined in (8).

Outline of TPP-SELECT. In Algorithm 1, we summarize TPP-SELECT to maximize $F(\mathcal{S})$. It is a randomized iterative algorithm based on the stochastic greedy algorithm proposed in [55], which keeps picking up one element per each of k iterations (refer to the for loop in line no. 2). In each iteration, given the current estimate of $\mathcal{S}$, Algorithm 1 picks up a candidate exogenous event v which maximizes $F(\mathcal{S} \cup \{v\})$. However, in contrast to the well known (deterministic) greedy algorithm [60], stochastic greedy algorithm computes the maximum only on a subset of elements— sampled from the current candidate set of elements.

Algorithm 1 Stochastic greedy algorithm [55]

Require: The set of observed events $\mathcal{H}_T$, the final time T, the regularization parameter ρ, the number of exogenous events k.

1: $\mathcal{S} \leftarrow \emptyset$
2: **for** $i \in [k]$ **do**
3: Randomly draw a subset $\mathcal{V}$ from $\mathcal{H}_T \backslash \mathcal{S}$
4: **for** $v \in \mathcal{V}$ **do**
5: $\boldsymbol{\theta} \leftarrow \textsc{Train}(\mathcal{L}; \mathcal{S} \cup \{v\})$
6: $F(\mathcal{S} \cup \{v\}) \leftarrow \mathcal{L}(\boldsymbol{\theta}; \mathcal{S}, \mathcal{H}_T)$
7: $\widehat{\boldsymbol{\theta}}(\mathcal{H}_T \backslash (\mathcal{S} \cup v)) = \boldsymbol{\theta}$
8: **end for**
9: $v^* \leftarrow \operatorname{argmax}_{v \in \mathcal{V}} F(\mathcal{S} \cup \{v\})$
10: $\mathcal{S} \leftarrow \mathcal{S} \cup \{v^*\}$
11: **end for**
12: **Return** $\mathcal{S}, \widehat{\boldsymbol{\theta}}(\mathcal{H}_T \backslash (\mathcal{S} \cup v^*))$

Advantage over deterministic greedy algorithm. The stochastic greedy algorithm is the randomized version of well known greedy algorithm [60]. If we put $\mathcal{V} = \mathcal{H}_T \backslash \mathcal{S}$, then the stochastic greedy algorithm becomes equivalent to the greedy algorithm. However, we need to train the model $|\mathcal{V}|$ times (due to the for loop in lines 4–8) per each iteration. As a result, the deterministic greedy algorithm would have been extremely inefficient, as it needs to greedily search for the optimal event over all the possible candidate events $\mathcal{V}$. Therefore, reducing the search space to a subset $\mathcal{V}$ would significantly improve the efficiency.

Approximation guarantees. Next, we provide the approximation guarantee of Algorithm 1 for solving the optimization problem defined in (8). Existing works [55, 37], provide approximation guarantees of Algorithm 1 when the exact value of $F(\mathcal{S})$ is available. However, in the context of our problem, a training algorithm may not provide the optimal value of model parameters $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})$, even when the MTPP model is linear. Hence, we mostly obtain an imperfect estimate of the model parameters $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})$. To this end, we derive the approximation guarantee of Algorithm 1 when the underlying training algorithm provides an imperfect estimate of the model parameters $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})$.

We would like to highlight that the following approximation guarantees apply to the cardinality constrained maximization problem of any monotone α-submodular function and therefore is of independent interest.

Theorem 7 *Assume that the training algorithm* $\textsc{Train}()$ *in Algorithm 1 provides imperfect estimates of the underlying model parameters, with* $\mathcal{L}(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}); \mathcal{S}; \mathcal{H}_T) - \mathcal{L}(\widehat{\boldsymbol{\theta}}(\mathcal{H}_T \backslash \mathcal{S}); \mathcal{S}; \mathcal{H}_T)) \leq \epsilon$ *for all* $\mathcal{S}$ *and that it runs with* $|\mathcal{V}| = O\left(\frac{|\mathcal{H}_T|}{k} \log(1/\delta)\right)$ *(cf. line number 3). Then,*

$$\mathbb{E}[F(\mathcal{S})] \geq (1 - \exp(-\alpha_F^*) - \alpha_F^* \cdot \delta) \cdot F(\mathcal{S}^*) - k\epsilon \tag{12}$$

where the expectation is taken over many draws of $\mathcal{V}$ *(cf. line number 3),* $\mathcal{S}^*$ *is the solution of the optimization problem* (8) *and* α_F^* *is the submodularity ratio, computed using Theorems 4 and 6.*

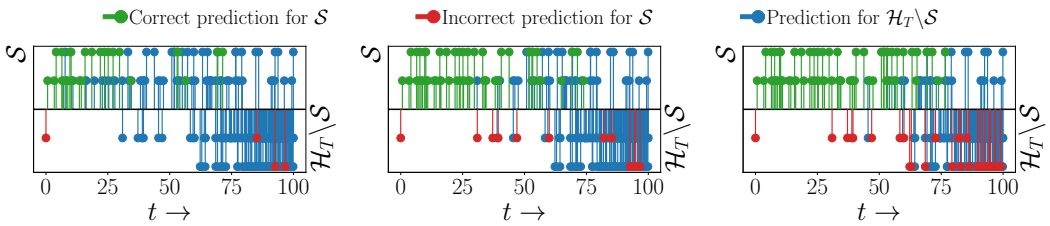
The above approximation guarantee in the presence of imperfect parameter estimates is related to several existing works [20, 62, 31, 30, 32, 71]. However, they do not focus on cardinality constrained maximization of α-submodular functions in general and not the well known stochastic greedy algorithm in particular. Note that, when we have a perfect training algorithm, *i.e.*, $\epsilon = 0$, Theorem 7 reduces to similar results in [55, 37].

5 Experiments with synthetic data

In this section, we create synthetic examples containing endogenous or exogenous label and evaluate TPP-SELECT from both qualitative and quantitative perspectives.

5.1 Experimental setup

Data generation. At the very outset, we generate data using linear MTPP defined in Eq. (9). More specifically, we set the number of marks $|\mathcal{C}| = 2$. In order to sample the endogenous events,



(a) iteration $i = \frac{1}{3}k$ (b) iteration $i = \frac{2}{3}k$ (c) iteration $i = k$

Figure 1: Snapshots of output of Algorithm 1 during one third, two third and final stages of executions. The exogenous intensity μ is tuned in such a way that $|\mathcal{S}| = 0.3\,|\mathcal{H}_T|$. Moreover, we also set $k = 0.3\,|\mathcal{H}_T|$. Top half (bottom half) of each figure indicates ground truth exogenous events $\mathcal{S}$ (endogenous events $\mathcal{H}_T \backslash \mathcal{S}$). Blue, green and red events indicate predicted endogenous events, correctly predicted exogenous events and wrongly predicted endogenous events. We observe that, the algorithm is able to predict the exogenous events with significant accuracy, as the number of wrongly predicted exogenous events is significantly lower than the correctly predicted exogenous events. The classification accuracy of final stage is 0.806.

we set $d = 2$ and draw unidimensional parameters $\theta_{\lambda,v}, \theta_{m,v} \sim \text{Unif}[0,1]$ for $v \in \{1,2\}$. We use the triggering kernel $\kappa_{\lambda,v}(t,t_i) = \exp(-\omega_\lambda(t-t_i)$ and $\kappa_{m,v} = \exp(-\omega_m(t-t_i))$, where $\omega_\lambda, \omega_m \in \text{Unif}[0,1]$ for $v \in [d]$. For exogenous messages, we sample μ and q using the exogenous distribution (4) with $\alpha_\lambda, \beta_\lambda, \alpha_m, \beta_m = 1$. Finally, we use Ogata's thinning algorithm [61] to generate both endogenous and exogenous events.

Baselines. We compare our method against several data selection based unsupervised robust learning methods: (i) EM algorithm [69] which first assigns a constant probability p which models the likelihood of an event e being exogenous and then estimates it along with model parameters $\boldsymbol{\theta}$ using expectation maximization; (ii) K-means [69] which uses K-means clustering to cluster the events in $K = 2$ groups and then label the smallest size cluster as $\mathcal{S}$, the exogenous event set; (iii) Principal Component Analysis (PCA) [70] which first finds a low-dimensional projection of the representation vector of the events, then sorts them in the decreasing order of the reconstruction error and finally, choose top-k events as the exogenous event set $\mathcal{S}$; and, (iv) Facility-Location [1] which maximizes a facility location function to select the set of exogenous events $\mathcal{S}$. For K-means, PCA, we directly used implementation available in Keras [8]. For Facility location, we used apricot library [67].

5.2 Results

We qualitatively show how TPP-SELECT keeps selecting the exogenous events throughout the progression of Algorithm 1. To that aim, we tune the exogenous intensity μ in such a way that $|\mathcal{S}| = 0.3\,|\mathcal{H}_T|$ and run Algorithm 1 with $k = 0.3\,|\mathcal{H}_T|$. Figure 1 summarizes the results, which shows that: (i) the algorithm is able to predict the exogenous events with remarkable accuracy, as the number of wrong predictions is significantly lower than the correct predictions; (ii) the performance of algorithm is strikingly accurate during the initial stage, as the number of wrong predictions of exogenous events is significantly low; and, (iii) even when TPP-SELECT makes wrong predictions about $e \in \mathcal{S}$, there always exists some correct ground truth exogenous event around it which indicates that TPP-SELECT was almost able to predict them correctly.

TPP-Select	EM	K-means	PCA	Fac-loc
0.806	0.554	0.620	0.562	0.299

Table 1: Classification accuracy of synthetic data in the figure 1 for different methods.

Next, we compare our method against the baselines. Table 1 summarizes the results, which shows that our method outperforms the baselines.

6 Experiments with real data

In this section, we provide a comprehensive evaluation of our proposal on several real world datasets, which shows that our method outperforms several data selection based robust learning methods, as well as the underlying base MTPP model which is trained over all observations including exogenous events. Appendix G contains additional experiments.

	TPP-SELECT	EM	K-means	PCA	Fac-Loc
Club	**0.32 ±0.02**	0.18 ±0.02	0.00 ±0.00	0.28 ±0.01	0.25 ±0.01
Election	**0.58 ±0.03**	0.31 ±0.01	0.00 ±0.00	0.19 ±0.00	0.44 ±0.03
Series	0.18 ±0.02	**0.27 ±0.02**	0.00 ±0.00	0.26 ±0.01	0.16 ±0.01
Verdict	**0.33 ±0.04**	0.27 ±0.01	0.00 ±0.00	0.18 ±0.00	0.10 ±0.01
BookOrder	**0.96 ±0.01**	0.41 ±0.03	0.00 ±0.00	0.92 ±0.01	0.46 ±0.01

Table 2: Event selection accuracy of all methods, *i.e.*, EM [69], K-means [69], PCA [70] and Facility-Location [1] across all datasets, when we inject $|\mathcal{S}| = 0.4|\mathcal{H}_T|$ exogenous events. Numbers in bold font (underline) indicate the best (second best) performer.

6.1 Experimental setup

Datasets. We considered five real world datasets, also summarized in Appendix F, which are (i) Club, (ii) Election, (iii) Series, (iv) Verdict and (v) BookOrder. Among these datasets, the first four datasets were gathered from the works [15, 82, 40], whereas, the last dataset were gathered from [16] [1].

Baselines. We compare our method against the four data selection based unsupervised robust learning methods, presented in Section 5.1.

6.2 Results

Comparison with baselines. We first compare the performance of TPP-SELECT against the baselines. To that aim, we inject 40% exogenous events and present to the event selection method for identifying these events. Here, we consider recurrent marked temporal point process (RMTPP) [16] as our base MTPP model $(\lambda_{\boldsymbol{\theta}}, m_{\boldsymbol{\theta}})$. Table 2 summarizes the results, which shows that (i) our method consistently outperforms the baselines in majority of the cases, as it aims to select the events in conjunction with the model parameters, whereas the baselines except for EM algorithm selects the events in a model-agnostic manner; (ii) except for Series and Verdict datasets, EM algorithm fares poorly across all other datasets, which we believe is due to the time-invariant modeling of probability of the exogenous events; (iii) PCA is the second best winner in majority of the datasets; and, (iv) K-means provides 0% accuracy, as the feature distance between events are poor predictors of exogenous or endogenous labels.

Comparison with the base MTPP model. Next, we investigate how well TPP-SELECT performs in comparison to the underlying base MTPP which models the event dynamics using the superposition between endogenous and exogenous MTPPs and thus, is trained on the entire sequence of events including the exogenous events. Given a sequence of observed events, we split it into training, validation and test sets. Here, the training set consists of first 60% events whereas, the validation and the test set consist of the subsequent 20% and 20% events respectively. We use TPP-SELECT on the training set $\mathcal{H}_T$, where it simultaneously selects the exogenous events $\mathcal{S}$ and trains the model parameters $(\lambda_{\boldsymbol{\theta}}, m_{\boldsymbol{\theta}})$ on the endogenous events $\mathcal{H}_T \backslash \mathcal{S}$. We estimate the hyperparmeters $\alpha_\bullet, \beta_\bullet, k$ which guide the generative mechanism of the exogenous events, using cross-validation on the validation set. Once we train the model, we use Ogata's thinning algorithm [61] to simulate future events to predict the unseen events in the test set $\mathcal{T}$. To this end, we evaluate the performance on in terms of mean absolute error (MAE), *i.e.*, $\frac{1}{|\mathcal{T}|}\sum_{e=(t,y)\in\mathcal{T}} \mathbb{E}[|t - \widehat{t}|]$ and mark prediction error (MPE), *i.e.*, $\frac{1}{|\mathcal{T}|}\sum_{e=(t,y)\in\mathcal{T}} \Pr(y \neq \widehat{y})$. Here, MAE and MPE measure the errors in predicting the arrival times and the marks of the future events.

Table 3 summarizes the results, which shows that TPP-SELECT performs better than the underlying base MTPP model across all datasets. Such observations imply that a model aware exogenous event filtering improves the predictive performance than learning a superposition of two models.

Effect of k. Finally, we investigate how k, the pre-specified number of exogenous events, impacts the performance of TPP-SELECT. Figure 2 summarizes the results, which show that (i) TPP-SELECT performs poorly when k is either too small or too large, and (ii) there exists a sweet spot of k, where TPP-SELECT shows the best peformance.

[1] The datasets are public and anonymized. We collect them from https://github.com/paramita1024/demarcation/tree/master/data and https://github.com/musically-ut/tf_rmtpp/tree/master/data/real. They are under MIT License. No dataset contains personally identifiable information or offensive content.

	Time prediction error in terms of MAE $\mathbb{E}(\lvert t - \widehat{t}\rvert)$				Mark prediction error $\Pr(y \neq \widehat{y})$			
	Base model: RMTPP		Base model: THP		Base model: RMTPP		Base model: THP	
	TPP-SELECT	Base MTPP	TPP-SELECT	Base MTPP	TPP-SELECT	Base MTPP	TPP-SELECT	Base MTPP
Club	**3.57±0.15**	3.67±0.05	**11.51±0.07**	11.93±0.08	0.60±0.11	**0.59±0.05**	**0.36±0.01**	0.42±0.01
Election	**8.50±0.02**	8.50±0.01	**30.79±0.02**	30.99±0.03	**0.43±0.02**	0.67±0.03	**0.38±0.00**	0.41±0.01
Series	**1.15±0.05**	1.71±0.07	**9.12±0.17**	9.99±0.20	**0.59±0.00**	0.69±0.01	**0.57±0.00**	0.58±0.00
Verdict	**2.27±0.00**	**2.27±0.00**	**7.88±0.01**	8.11±0.05	**0.57±0.00**	0.71±0.01	**0.57±0.00**	0.60±0.01
BookOrder	**0.07±0.00**	0.07±0.00	**0.09±0.00**	0.11±0.01	**0.45±0.00**	0.49±0.00	**0.37±0.01**	0.44±0.00

Table 3: Performance in terms of $\mathbb{E}(\lvert t - \widehat{t}\rvert)$, *i.e.*, the mean absolute error (MAE) for time prediction (left half) and the misclassification error $\Pr(y \neq \widehat{y})$ for the mark prediction (right half) for TPP-SELECT and the corresponding base MTPP model which is the superposition of endogenous and exogenous MTPP models. Here, this base MTPP model is learned over the entire sequence of observed events.

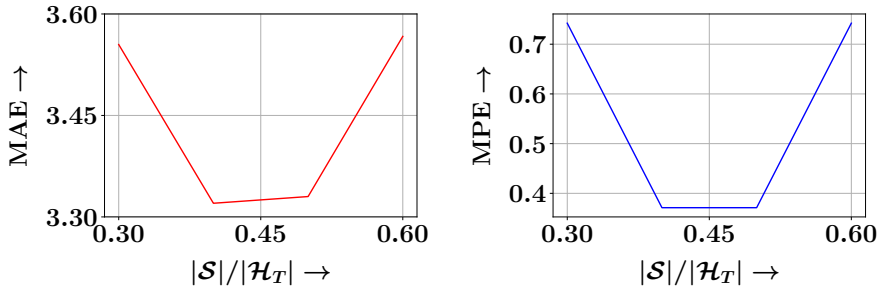


Figure 2: Variation of mean absolute error $\mathbb{E}(\lvert t - \widehat{t}\rvert)$ and mark prediction error (MPE) $\Pr(y \neq \hat{y})$ against k, the maximum number of exogenous events, on Club dataset. We observe that in both the cases when k is too small or high, then TPP-SELECT performs poorly. However, there exists a sweet spot of k, where TPP-SELECT shows its best performance.

7 Conclusion

In this paper, we aim to identify the set of exogenous events from a set of unlabeled events for learning marked temporal point process. To that goal, we have developed TPP-SELECT, a novel framework that selects the possible set of exogenous events from a set the observed events, under any given MTPP model. More specifically, we first cast this problem as an instance of exogenous event set selection problem, in conjunction with parameter estimation problem which is NP-hard, in general. However, we show that solving such a problem is equivalent to maximizing a monotone and α-submodular set function which leads us to develop TPP-SELECT— an approximation algorithm built upon a stochastic greedy algorithm which was originally proposed in [63]. In this work, we show that it also enjoys an approximation guarantee for α-submodular set function maximization, even when the learning algorithm provides an imperfect estimate of the trained parameters. Our experiments show that TPP-SELECT is able to outperform a wide variety of baselines based on data-selection from robust learning literature.

Our work opens several venues of future work. For example, our method can be easily applied to identify external influence in Twitter— such a situation can immediately beneficial in a wide variety of applications that include viral marketing, misinformation detection, etc. However, by deploying our method as-it-is can lead to privacy leakage of users. To counter this effect, one can design a differentially private framework to mitigate this effect, by building upon the work by [58]. In our work, we do not consider a multivariate point process. It would be interesting to develop exogenous event selection method for such processes. Finally, it would be valuable to extend our method to an online setting, where the task is to select the set of exogenous events in the face of an online stream of events.

Acknowledgments and Disclosure of Funding

The project was partially funded by Google Research Grant and DST Inspire Faculty Grant. This work is also supported by the National Science Foundation under Grant No. IIS-2106937. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, Google or DST.

References

[1] S. Ahmed and A. Atamtürk. Maximizing a class of submodular utility functions. *Mathematical programming*, 128(1):149–169, 2011.

[2] R. Bairi, R. Iyer, G. Ramakrishnan, and J. Bilmes. Summarization of multi-document topic hierarchies using submodular mixtures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 553–563, 2015.

[3] A. Bakshi and A. Prasad. Robust linear regression: Optimal rates in polynomial time. *arXiv preprint arXiv:2007.01394*, 2020.

[4] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2110–2119, 2017.

[5] A. A. Bian, J. M. Buhmann, A. Krause, and S. Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. *arXiv preprint arXiv:1703.02100*, 2017.

[6] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE transactions on information theory*, 59(10):6880–6892, 2013.

[7] T. Campbell and T. Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, pages 698–706, 2018.

[8] F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[9] C.-C. Chuang, S.-F. Su, J.-T. Jeng, and C.-C. Hsiao. Robust support vector regression networks for function approximation with outliers. *IEEE Transactions on Neural Networks*, 13(6):1322–1330, 2002.

[10] D. J. Daley, D. Vere-Jones, et al. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer New York, 2008.

[11] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv preprint arXiv:1102.3975*, 2011.

[12] A. De, S. Bhattacharya, and N. Ganguly. Demarcating endogenous and exogenous opinion diffusion process on social networks. In *Proceedings of the 2018 World Wide Web Conference*, pages 549–558, 2018.

[13] A. De, P. Koley, N. Ganguly, and M. Gomez-Rodriguez. Regression under human assistance. *AAAI*, 2020.

[14] A. De, N. Okati, A. Zarezade, and M. Gomez-Rodriguez. Classification under human assistance. *AAAI*, 2021.

[15] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. Gomez-Rodriguez. Learning and forecasting opinion dynamics in social networks. In *NIPS*, 2016.

[16] N. Du, H. Dai, R. Trivedi, U. Upadhyay, M. Gomez-Rodriguez, and L. Song. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*, 2016.

[17] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *KDD*, 2015.

[18] S. Durga, R. Iyer, G. Ramakrishnan, and A. De. Training data subset selection for regression with controlled generalization error. In *International Conference on Machine Learning*, pages 9202–9212. PMLR, 2021.

[19] S. Dutta, S. Masud, S. Chakrabarti, and T. Chakraborty. Deep exogenous and endogenous influence combination for social chatter intensity prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1999–2008, 2020.

[20] M. El Halabi and S. Jegelka. Optimal approximation for unconstrained non-submodular minimization. In *International Conference on Machine Learning*, pages 3961–3972. PMLR, 2020.

[21] M. Farajtabar, N. Du, M. G. Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. *Advances in neural information processing systems*, 27, 2014.

[22] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, and H. Zha. Fake news mitigation via point process based intervention. In *ICML*, 2017.

[23] P. Filzmoser and K. Nordhausen. Robust linear regression for high-dimensional data: An overview. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1524.

[24] Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

[25] K. Gatmiry and M. Gomez-Rodriguez. Non-submodular function maximization subject to a matroid constraint, with applications. *arXiv preprint arXiv:1811.07863*, 2018.

[26] P. Grabowicz, N. Ganguly, and K. Gummadi. Distinguishing between topical and non-topical information diffusion mechanisms in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, 2016.

[27] V. Gupta, S. Bedathur, S. Bhattacharya, and A. De. Learning temporal point processes with intermittent observations. In *International Conference on Artificial Intelligence and Statistics*, pages 3790–3798. PMLR, 2021.

[28] A. Hashemi, M. Ghasemi, H. Vikalo, and U. Topcu. Submodular observation selection and information gathering for quadratic models. *arXiv preprint arXiv:1905.09919*, 2019.

[29] H. Hassani, M. Soltanolkotabi, and A. Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems*, pages 5841–5851, 2017.

[30] A. Hassidim and Y. Singer. Submodular optimization under noise. 2017.

[31] A. Hassidim and Y. Singer. Optimization for approximate submodularity. pages 394–405. Curran Associates Inc., 2018.

[32] T. Horel and Y. Singer. Maximization of approximately submodular functions. 2016.

[33] H. Huang, H. Wang, and B. Mak. Recurrent poisson process unit for speech recognition. In *AAAI*, 2019.

[34] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274, 2013.

[35] R. K. Iyer. *Submodular optimization and machine learning: Theoretical results, unifying and scalable algorithms, and applications*. PhD thesis, 2015.

[36] V. Kaushal, R. Iyer, S. Kothawade, R. Mahadev, K. Doctor, and G. Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1289–1299. IEEE, 2019.

[37] R. Khanna, E. Elenberg, A. G. Dimakis, S. Negahban, and J. Ghosh. Scalable greedy feature selection via weak submodularity. *arXiv preprint arXiv:1703.02723*, 2017.

[38] K. Killamsetty, D. Subramanian, G. Ramakrishnan, and R. Iyer. Glister: A generalization based data selection framework for efficient and robust learning. *In AAAI*, 2021.

[39] K. Kirchhoff and J. Bilmes. Submodularity for data selection in machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 131–141, 2014.

[40] P. Koley, A. Saha, S. Bhattacharya, N. Ganguly, and A. De. Demarcating endogenous and exogenous opinion dynamics: An experimental design approach. *arXiv preprint arXiv:2102.05954*, 2021.

[41] K. Konyushkova, R. Sznitman, and P. Fua. Learning active learning from data. *arXiv preprint arXiv:1703.03365*, 2017.

[42] S. Kumar, X. Zhang, and J. Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *KDD*, 2019.

[43] B. Lehmann, D. Lehmann, and N. Nisan. Combinatorial auctions with decreasing marginal utilities. *Games and Economic Behavior*, 55(2):270–296, 2006.

[44] S. Linderman and R. Adams. Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421. PMLR, 2014.

[45] T. J. Liniger. *Multivariate hawkes processes*. PhD thesis, ETH Zurich, 2009.

[46] S. Liu and M. Hauskrecht. Event outlier detection in continuous time. In *International Conference on Machine Learning*, pages 6793–6803. PMLR, 2021.

[47] Y. Liu, R. Iyer, K. Kirchhoff, and J. Bilmes. Svitchboard ii and fisver i: High-quality limited-complexity corpora of conversational english speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[48] L. Lorch, A. De, S. Bhatt, W. Trouleau, U. Upadhyay, and M. Gomez-Rodriguez. Stochastic optimal control of epidemic processes in networks. *arXiv preprint arXiv:1810.13043*, 2018.

[49] L. Lorch, W. Trouleau, S. Tsirtsis, A. Szanto, B. Schölkopf, and M. Gomez-Rodriguez. A spatiotemporal epidemic model to quantify the effects of contact tracing, testing, and containment. *arXiv preprint arXiv:2004.07641*, 2020.

[50] M. Lucic, M. Faulkner, A. Krause, and D. Feldman. Training gaussian mixture models at scale via coresets. *The Journal of Machine Learning Research*, 18(1):5885–5909, 2017.

[51] H. Mei and J. M. Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *NIPS*, 2017.

[52] H. Mei, G. Qin, and J. Eisner. Imputing missing events in continuous-time event streams. In *International Conference on Machine Learning*, pages 4475–4485. PMLR, 2019.

[53] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization techniques*, pages 234–243. Springer, 1978.

[54] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrák, and A. Krause. Lazier than lazy greedy. *arXiv preprint arXiv:1409.7938*, 2014.

[55] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrák, and A. Krause. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[56] B. Mirzasoleiman, J. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. *In Proc. ICML*, 2020.

[57] B. Mirzasoleiman, K. Cao, and J. Leskovec. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020.

[58] M. Mitrovic, M. Bun, A. Krause, and A. Karbasi. Differentially private submodular maximization: Data summarization in disguise. In *International Conference on Machine Learning*, pages 2478–2487. PMLR, 2017.

[59] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41, 2012.

[60] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-i. *Mathematical programming*, 1978.

[61] Y. Ogata. On lewis' simulation method for point processes. *IEEE transactions on information theory*, 27(1):23–31, 1981.

[62] C. Qian, J.-C. Shi, Y. Yu, K. Tang, and Z.-H. Zhou. Subset selection under noise. pages 3560–3570, 2017.

[63] C. Qian, Y. Yu, and K. Tang. Approximation guarantees of stochastic greedy algorithms for subset selection. In *IJCAI*, pages 1478–1484, 2018.

[64] M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, and L. Xie. Sir-hawkes: on the relationship between epidemic models and hawkes point processes. In *The Web Confernce*, 2018.

[65] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck. Expecting to be hip: Hawkes intensity processes for social media popularity. In *WWW*, 2017.

[66] H. Sasaki, T. Takenouchi, R. Monti, and A. Hyvarinen. Robust contrastive learning and nonlinear ica in the presence of outliers. In *Conference on Uncertainty in Artificial Intelligence*, pages 659–668. PMLR, 2020.

[67] J. Schreiber, J. Bilmes, and W. S. Noble. apricot: Submodular selection for data summarization in python. *Journal of Machine Learning Research*, 21(161):1–6, 2020.

[68] O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[69] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[70] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2003.

[71] A. Singla, S. Tschiatschek, and A. Krause. Noisy submodular maximization via adaptive sampling with applications to crowdsourced image collection summarization. 2016.

[72] S. Soni, S. L. Ramirez, and J. Eisenstein. Detecting social influence in event cascades by comparing discriminative rankers. In *The 2019 ACM SIGKDD Workshop on Causal Discovery*, pages 78–99, 2019.

[73] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf, and M. Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. *PNAS*, 2019.

[74] Y. Tanaka, T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada. Inferring latent triggers of purchases with consideration of social effects and media advertisements. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 543–552, 2016.

[75] I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Modeling diffusion of competing products and conventions in social media. *arXiv preprint arXiv:1406.0516*, 2014.

[76] Y. Wang, G. Williams, E. Theodorou, and L. Song. Variational policy for guiding point processes. In *International Conference on Machine Learning*, pages 3684–3693. PMLR, 2017.

[77] K. Wei, R. Iyer, and J. Bilmes. Fast multi-stage submodular maximization. In *International conference on machine learning*, pages 1494–1502. PMLR, 2014.

[78] K. Wei, R. Iyer, and J. Bilmes. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pages 1954–1963, 2015.

[79] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes. Unsupervised submodular subset selection for speech data. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4107–4111. IEEE, 2014.

[80] S. Xiao, J. Yan, S. Chu, X. Yang, and H. Zha. Modeling the intensity function of point process via recurrent neural networks. In *AAAI*, 2017.

[81] Y. Yuan, F. E. Bachl, F. Lindgren, D. L. Borchers, J. B. Illian, S. T. Buckland, H. Rue, T. Gerrodette, et al. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11(4):2270–2297, 2017.

[82] A. Zarezade, A. De, U. Upadhyay, H. R. Rabiee, and M. Gomez-Rodriguez. Steering social activity: A stochastic optimal control point of view. *J. Mach. Learn. Res.*, 18:205–1, 2017.

[83] H. Zhang and Y. Vorobeychik. Submodular optimization with routing constraints. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[84] Q. Zhang, A. Lipani, O. Kirnap, and E. Yilmaz. Self-attentive hawkes processes. *ICML*, 2020.

[85] F. Zhu, W. Chen, and L. Chen. Robust linear subspace for image set retrieval. In *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, pages 390–394, 2020.

[86] S. Zuo, H. Jiang, Z. Li, T. Zhao, and H. Zha. Transformer hawkes process. In *ICML*, 2020.

Supplementary Material for Learning to Select Exogenous Events for Marked Temporal Point Processes

Table of Contents

A Broader impact **2**
- A.1 Positive potential impacts . . . 2
- A.2 Negative potential impacts . . . 2

B Discussion on problem setup **3**
- B.1 Benefits and limitations of our approach . . . 3
- B.2 Hardness analysis . . . 3

C Proofs of the technical results in Section 4 **4**
- C.1 Monotonicity and approximate submodularity for linear MTPP . . . 4
- C.2 Monotonicity and approximation guarantee for nonlinear MTPP . . . 7
- C.3 Representation of MTPPs using exponential bounds . . . 8
- C.4 Approximation guarantees of Algorithm 1 . . . 9

D Additional theoretical results **11**

E Additional experiments on synthetic data **12**
- E.1 Quantitative analysis . . . 12
- E.2 Scalability analysis . . . 12

F Additional details about the experiments with real data **13**
- F.1 Dataset details . . . 13
- F.2 Additional implementation details for base MTPP models . . . 13
- F.3 Compute resource . . . 13

G Additional experiments with real data **14**

A Broader impact

A.1 Positive potential impacts

At the very outset, our problem aims at selecting exogenous events from a stream of events. Such a problem has a wide variety of applications in real life,

Preventing misinformation flow. Misinformation is easily spread in social networks, which often has a negative impact on people's social life. For such a information cascade like fake news/rumor, our algorithm may be useful.

Viral marketing. While making marketing strategy, we can use our approach to get to know which user is more sensitive to exogenous events. These users can be used as seed users for improvement in viral marketing.

Improvement in cyber security. Cyber attacks cause tremendous damage to institutes and individuals. For example, DDoS attack sends a very large number of malicious requests in a short period of time with unpredictable pattern. Unlike real requests which are often relevant to historical events, these malicious requests usually shows exogenous pattern as they are history-irrelevant. In the context of cyber security, we can detect potential malicious events and prevent system from attacks using our model.

A.2 Negative potential impacts

Privacy concerns. Our model can select exogenous events from a series of events, and can parametrically estimate relationship between users. Currently, it does not protect privacy of users. In particular, the social relationship of users can be leaked through the parameters of the trained model. Therefore, it would be important to design privacy preserving event selection method using tools from differential privacy.

Bias introduced in our model: We have not tried anything to eliminate the unfairness and bias in training and prediction process. As a result, it can introduce bias into the algorithm. For example, in a cascade extracted from a community that is broadly represented by a specific group of persons, our model can label an event triggered by a user from any other group as exogenous. Therefore, it would be an important task to introduce fairness within the algorithm.

B Discussion on problem setup

B.1 Benefits and limitations of our approach

Benefits of our approach. Existing approaches [21, 15] model the arrival time dynamics of an event using superposition of exogenous and endogenous intensity functions, *i.e.*, $\mu + \lambda_{\boldsymbol{\theta}}$. We first put such an approach in context. More specifically, we first consider a partition $(\mathcal{S}, \mathcal{H}_T \backslash \mathcal{S})$ for endogenous and exogenous events. Then, we model the density of arrival times $e = (y, t)$ as,

$$p_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{H}_T \backslash \mathcal{S}) = \mu^{|\mathcal{S}|} \prod_{e \in \mathcal{H}_T \backslash \mathcal{S}} \lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t) \exp\left(-\mu T - \int_0^T \lambda_{\boldsymbol{\theta}}(\tau \,|\, \mathcal{H}_\tau) d\tau\right). \tag{13}$$

Now, if the prior probability of partition $(\mathcal{S}, \mathcal{H}_T \backslash \mathcal{S})$ is $p_{\text{prior}}(\mathcal{S})$, then the likelihood of (only) the arrival times of the observed events becomes

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathcal{H}_T) &= \sum_{\mathcal{S} \in 2^{\mathcal{H}_T}} p_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{H}_T \backslash \mathcal{S})\, p_{\text{prior}}(\mathcal{S}) \\ &= \sum_{\mathcal{S} \in 2^{\mathcal{H}_T}} \left[\prod_{e \in \mathcal{H}_T} \mu^{\mathbb{I}[e \in \mathcal{S}]} \left[\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t)\right]^{\mathbb{I}[e \notin \mathcal{S}]} p_{\text{prior}}(\mathcal{S}) \right] \times \exp\left(-\mu T - \int_0^T \lambda_{\boldsymbol{\theta}}(\tau \,|\, \mathcal{H}_\tau) d\tau\right) \end{aligned} \tag{14}$$

Now, if $p_{\text{prior}}(\mathcal{S})$ is constant, *i.e.*, $p_{\text{prior}}(\mathcal{S}) = p_{\text{prior}}$, then it immediately gives,

$$p_{\boldsymbol{\theta}}(\mathcal{H}_T) \propto \prod_{e \in \mathcal{H}_T} \left(\mu + \lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t)\right) \times \exp\left(-\mu T - \int_0^T \lambda_{\boldsymbol{\theta}}(\tau \,|\, \mathcal{H}_\tau) d\tau\right) \tag{15}$$

Hence, such an approach [21, 15] captures a particular case when every partition is equally likely, which may not be true in general. To avoid this problem, we aim to find $\mathcal{S}$ which maximizes the underlying likelihood function.

Furthermore, note that the exogenous distribution parameters μ and q in existing works [21, 15] are often treated as deterministic variables, which, however, cannot characterize the uncertainty of the exogenous influence. We tackle this challenge by modeling the uncertainty, *i.e.*,

$$\lambda(t) = \mu \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda), \qquad m(y) = q \sim \text{Beta}(\alpha_{m,y}, \beta_{m,y}). \tag{16}$$

Limitations of our approach. Apart from the limitations addressed in Appendix A and Conclusion in the main paper, another key technical limitation of our approach is lack of inherent scalability. The stochastic greedy algorithm does reduce the running time by comparing across a limited amount of data. However, such an approach can often affect the predictive performance. It would be interesting to consider similar setup like [38] to improve scalability of our method.

B.2 Hardness analysis

Proposition 1 *Solving the optimization problem in Eq.* (5) *is NP-hard.*

Proof. We consider the particular case of MTPP models where the intensity function is constant.

$$\lambda_{\boldsymbol{\theta}}(t \,|\, \mathcal{H}_t) = \lambda_0, \qquad m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t) = m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t). \tag{17}$$

Moreover, we consider the special case when $\mathcal{C} = \{+1, -1, 0\}$ and

$$m_{\boldsymbol{\theta}}(y \,|\, \mathcal{H}_t) = \begin{cases} e^{-(1 - y \cdot \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathcal{H}_t))^2} & \text{if, } y \in \{+1, -1\} \\ 1 - \sum_{y \in \{-1, +1\}} e^{-(1 - y \cdot \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathcal{H}_t))^2} & \text{if, } y = 0, \end{cases} \tag{18}$$

$$\lambda_0 = \exp(\mathbb{E}[\log(\mu \cdot q)]). \tag{19}$$

Here $\boldsymbol{\phi}(\mathcal{H}_t)$ is a representation vector for $\mathcal{H}_t$. We assume that (i) $\boldsymbol{\Phi} = [\boldsymbol{\phi}^\top(\mathcal{H}_{t_1}); \boldsymbol{\phi}^\top(\mathcal{H}_{t_N}); ..; \boldsymbol{\phi}^\top(\mathcal{H}_{t_N})]$ has full row rank N and (ii) $\mathcal{H}_T$ only contains marks $y \in \{+1, -1\}$ and (iii) $\rho = 0$. Then the log-likelihood (6) becomes:

$$\sum_{e_i \in \mathcal{H}_T \setminus \mathcal{S}} [\log(\lambda_{\boldsymbol{\theta}}(t_i \,|\, \mathcal{H}_{t_i})) + \log m_{\boldsymbol{\theta}}(y_i \,|\, \mathcal{H}_{t_i})] - \int_0^T \lambda_{\boldsymbol{\theta}}(\tau \,|\, \mathcal{H}_\tau)\, d\tau + |\mathcal{S}| \mathbb{E}[\log(\mu \cdot q)] - T\mathbb{E}[\mu]$$

$$= \sum_{e_i \in \mathcal{H}_T \setminus \mathcal{S}} \log m_{\boldsymbol{\theta}}(y_i \,|\, \mathcal{H}_{t_i}) - T\,\lambda_0 \; + |\mathcal{S}|\, \mathbb{E}[\log(\mu \cdot q)] + (|\mathcal{H}_T| - |\mathcal{S}|) \log \lambda_0 - T\,\mathbb{E}[\mu]$$

$$\overset{(i)}{=} - \sum_{e_i \in \mathcal{H}_T \setminus \mathcal{S}} (1 - y_i \cdot \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathcal{H}_{t_i}))^2 - T\,\lambda_0 \; - T\,\mathbb{E}[\mu] \tag{20}$$

The equality (i) is obtained by putting the particular instance described in Eqs. (18) — (19). The next steps of the proof follows directly from [14, Proof of Theorem 1], which is described below to make the proof self-contained. More specifically, we first define $\boldsymbol{\theta}_0 = \boldsymbol{\Phi}_R^{-1}(\boldsymbol{y} - \boldsymbol{r})$, where $\boldsymbol{y} = [y_1, .., y_N],\ \boldsymbol{r} \in \mathbb{R}^N$. where $\boldsymbol{\Phi}_R^{-1}$ is the right inverse of $\boldsymbol{\Phi}$ an $\boldsymbol{r}$ is an arbitrary real vector. A trite calculation shows that $\boldsymbol{\theta}_0^\top \boldsymbol{\phi}(\mathcal{H}_{t_i}) = y_i - r_i$. Finally, if we define a new variable $\boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{\theta}_0$, then we can write as follows:

$$\begin{aligned}(1 - y_i \cdot \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathcal{H}_{t_i}))^2 &= (1 - y_i \cdot \boldsymbol{\theta}'^\top \boldsymbol{\phi}(\mathcal{H}_{t_i}) - y_i^2 + y_i r_i)^2 \\ &= (r_i - \boldsymbol{\theta}'^\top \boldsymbol{\phi}(\mathcal{H}_{t_i}))^2 \qquad \text{(Since } y_i \in \pm 1) \end{aligned} \tag{21}$$

Hence, maximizing the objective defined Eq. (20) w.r.t. $\boldsymbol{\theta}$ and $\mathcal{S}$ is equivalent to minimizing the $\sum_{e_i \in \mathcal{H}_T} (r_i - \boldsymbol{\theta}'^\top \boldsymbol{\phi}(\mathcal{H}_{t_i}))^2$ which is known to be NP-hard [4].

C Proofs of the technical results in Section 4

Unless otherwise stated, we denote $\|\cdot\|$ as $\|\cdot\|_2$ and $\boldsymbol{a} \preceq \boldsymbol{b}$ means pointwise inequality, *i.e.*, $a_i \geq b_i$ for all $i \leq \dim(\boldsymbol{a})$

C.1 Monotonicity and approximate submodularity for linear MTPP

Theorem 3 *Assume that the endogenous events follow a linear MTPP* (9)*. Then the set function* $F(\mathcal{S})$ *defined in Eq.* (7) *is monotone in* $\mathcal{S}$ *if* $\rho \geq \rho_{\min} = (0.5/e) \cdot \kappa_{\max,\lambda}^2 \cdot \exp\big(-2\,\mathbb{E}[\log(\mu \cdot q)]\big)$*, where* μ *and* q *are the intensity and the mark distribution for the exogenous events and* $\kappa_{\max,\lambda} = \max_{t \leq T} \|\boldsymbol{\kappa}_\lambda(t)\|_2^2$*.*

Proof. From Lemma 9 in Appendix D, we have that: for $e = (y, t) \notin \mathcal{S}$,

$$\begin{aligned} F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) &\geq \rho\, \|\boldsymbol{\theta}^*(\mathcal{H}_T \setminus \mathcal{S})\|^2 - \log \lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T \setminus \mathcal{S})}(t \,|\, \mathcal{H}_t) + \mathbb{E}[\log(\mu \cdot q)] \\ &\geq \min_{\boldsymbol{\theta}_\lambda} \left[\rho\, \|\boldsymbol{\theta}_\lambda\|_2^2 - \log(\boldsymbol{\theta}_\lambda^\top \boldsymbol{\kappa}_\lambda(t))\right] + \mathbb{E}[\log(\mu \cdot q)] \qquad (22) \\ &\overset{(i)}{\geq} \min_{\boldsymbol{\theta}_\lambda} \left[\rho \frac{\left(\boldsymbol{\theta}_\lambda^\top \boldsymbol{\kappa}_\lambda(t)\right)^2}{\|\boldsymbol{\kappa}_\lambda(t)\|^2} - \log(\boldsymbol{\theta}_\lambda^\top \boldsymbol{\kappa}_\lambda(t))\right] + \mathbb{E}[\log(\mu \cdot q)] \\ &\geq \min_x \left[\frac{\rho x^2}{\kappa_{\max,\lambda}^2} - \log(x)\right] + \mathbb{E}[\log(\mu \cdot q)] \\ &\overset{(ii)}{\geq} \mathbb{E}[\log(\mu \cdot q)] + \log\left[\sqrt{\frac{2e\rho}{\kappa_{\max,\lambda}^2}}\right] \\ &> 0, \qquad (23) \end{aligned}$$

where inequality (i) is due to Cauchy-Schwartz inequality $\left\|\boldsymbol{a}^\top \boldsymbol{b}\right\|_2 \leq \|\boldsymbol{a}\|_2 \cdot \|\boldsymbol{b}\|_2$; inequality (ii) is due to Proposition 11 and the last inequality holds true by assumption on ρ.

Theorem 4 *Given that the same conditions in Theorem 3 as well as* $\left\|\int_0^\infty \boldsymbol{\kappa}_\lambda(t) dt\right\| \leq K < \infty$*,* $\kappa_{\max,\bullet} = \max_{t \leq T} \|\boldsymbol{\kappa}_\bullet(t)\|$*,* $\kappa_{\min,\bullet} = \min_{v \in [d], t \leq T} |\kappa_{\bullet,v}(t)|$*. If* $K, \kappa_{\max,\bullet}, \kappa_{\min,\bullet} > 0$*, then the set*

function $F(\mathcal{S})$ is α-submodular in $\mathcal{S}$ with

$$\alpha \geq \alpha_F = \frac{\log\left[\sqrt{2e\rho/\kappa^2_{\max,\lambda}}\right] + \mathbb{E}[\log(\mu \cdot q)]}{\frac{\kappa_{\max,\lambda}}{2\kappa_{\min,\lambda}} + \log\frac{4\kappa_{\max,\lambda}}{\kappa^2_{\min,\lambda}} + \log\left(\sqrt{\frac{2\rho\kappa_{\max,\lambda}}{\kappa_{\min,\lambda}}} + K\right) + \frac{\kappa^2_{\max,m}\log 2}{2\rho} + \sqrt{\frac{\kappa^2_{\max,m}\log 2}{\rho}} + \mathbb{E}[\log(\mu \cdot q)]} \tag{24}$$

Proof. We prove the theorem in two steps. In the first step, we bound $F(\mathcal{S}\cup e) - F(\mathcal{S})$ using the proof of previous theorem, *i.e.*, Theorem 3. Hence, $F(\mathcal{S}\cup e) - F(\mathcal{S}) \geq \mathbb{E}[\log(\mu\cdot q)] + \log\left[\sqrt{2e\rho/\kappa^2_{\max,\lambda}}\right]$. Next we bound $F(\mathcal{T}\cup\{e\}) - F(\mathcal{T})$.

$$\begin{aligned}
&F(\mathcal{T}\cup\{e\}) - F(\mathcal{T}) \\
&\overset{(i)}{\leq} \rho\,\|\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2 - \log\lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(t\,|\,\mathcal{H}_t) \\
&\qquad - \log m_{\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(y\,|\,\mathcal{H}_t) + \mathbb{E}[\log(\mu\cdot q)] \\
&= \rho\,\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2 - \log\left(\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t)\right) \\
&\quad + \rho\,\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2 + \log\left[1 + \exp\left(-y\cdot\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t)\right)\right] \\
&\qquad + \mathbb{E}[\log(\mu\cdot q)]
\end{aligned} \tag{25}$$

Here, inequality (i) is due to Lemma 10. Bounding $F(\mathcal{T}\cup\{e\}) - F(\mathcal{T})$ requires bounding the two terms in Eq. (25), *i.e.*, $\rho\,\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2 - \log(\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t))$ and, $\rho\,\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2 + \log[1 + \exp\left(-y\cdot\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t)\right)]$.

Bounding the first term *i.e.*, $\rho\,\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2 - \log(\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t))$:

We note that

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}_\lambda}\Bigg[\rho\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\|\boldsymbol{\theta}_\lambda\|^2 \\
&\qquad - \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\log(\boldsymbol{\theta}^\top_\lambda\boldsymbol{\kappa}_\lambda(t_i)) + \int_0^T\boldsymbol{\theta}^\top_\lambda\boldsymbol{\kappa}_\lambda(\tau)\,d\tau\Bigg]\Bigg|_{\boldsymbol{\theta}_\lambda=\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))} = 0,
\end{aligned} \tag{26}$$

as the corresponding maximum value occurs at $\boldsymbol{\theta}_\lambda = \boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))$. Hence, we have that:

$$\begin{aligned}
&2\rho\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})) \\
&\quad - \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\frac{\boldsymbol{\kappa}_\lambda(t_i)}{\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t_i)} + \int_0^T\boldsymbol{\kappa}_\lambda(\tau)\,d\tau = 0
\end{aligned} \tag{27}$$

As, $\boldsymbol{\kappa}_\lambda(\cdot) \succeq 0$ ($\succeq$ indicates pointwise inequality), the above equation implies that:

$$\begin{aligned}
&2\rho\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})) \\
&\qquad \preceq \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\frac{\boldsymbol{\kappa}_\lambda(t_i)}{\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t_i)} \\
&\qquad \preceq \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\frac{\boldsymbol{\kappa}_\lambda(t_i)}{\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|_1\,\boldsymbol{\kappa}_\lambda(t_i)} \\
&\qquad \preceq \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\frac{\boldsymbol{\kappa}_\lambda}{\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|_1\min\limits_{v\in[d],e_i,e_j\in\mathcal{H}_T}\kappa_v(t_i,t_j)} \\
&\qquad \overset{(i)}{\preceq} \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\frac{\boldsymbol{\kappa}_\lambda}{\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|_2\min\limits_{v\in[d],e_i,e_j\in\mathcal{H}_T}\kappa_v(t_i,t_j)}
\end{aligned} \tag{28}$$

In each of the above, we consider pointwise inequality (or equality) by $\succeq$. (i) is due to the fact that $\|\boldsymbol{\theta}_\lambda\|_1 \geq \|\boldsymbol{\theta}_\lambda\|_2$. Eq. (28) further implies that,

$$
\begin{aligned}
&2\rho\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|_2\,\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\min_{v\in[d],e_i,e_j\in\mathcal{H}_T}\kappa_v(t_i,t_j)\preceq\sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\boldsymbol{\kappa}_\lambda\\
&\implies 2\rho\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|_2^2\min_{v\in[d],e_i,e_j\in\mathcal{H}_T}\kappa_v(t_i,t_j)\leq\left\|\sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\boldsymbol{\kappa}_\lambda\right\|\\
&\qquad\qquad\leq|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\,\kappa_{\max,\lambda}\\
&\implies\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|\leq\theta_{\max,\lambda}=\sqrt{\frac{\kappa_{\max,\lambda}}{2\rho\kappa_{\min,\lambda}}}
\end{aligned}
\tag{29}
$$

Eqs. (27) and (29) suggest that,

$$
\begin{aligned}
&2\rho\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\,\theta_{\max,\lambda}+K\geq\frac{\kappa_{\min,\lambda}\,|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|}{\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|\,\kappa_{\max,\lambda}}\\
&\implies\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|\geq\theta_{\min,\lambda}=\frac{\kappa_{\min,\lambda}}{\kappa_{\max,\lambda}\left(2\rho\sqrt{\frac{\kappa_{\max,\lambda}}{2\rho\kappa_{\min,\lambda}}}+K\right)}
\end{aligned}
\tag{30}
$$

Hence, we bound $\rho\,\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2-\log(\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t))$ as follows:

$$
\begin{aligned}
&\rho\,\|\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2-\log(\boldsymbol{\theta}^*_\lambda(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_\lambda(t))\\
&\quad\leq\rho\,\theta^2_{\max,\lambda}-\log[\|\boldsymbol{\theta}_\lambda\|_1\,\kappa_{\min,\lambda}]\\
&\quad\overset{(i)}{\leq}\rho\,\theta^2_{\max,\lambda}-\log[\|\boldsymbol{\theta}_\lambda\|_2\,\kappa_{\min,\lambda}]\\
&\quad\leq\rho\,\theta^2_{\max,\lambda}-\log\left[\theta_{\min,\lambda}\,\kappa_{\min,\lambda}\right]\\
&\quad=\frac{\kappa_{\max,\lambda}}{2\kappa_{\min,\lambda}}+\log\left(\frac{\kappa_{\max,\lambda}}{\kappa^2_{\min,\lambda}}\right)+\log\left(K+\sqrt{\frac{2\rho\kappa_{\max,\lambda}}{\kappa_{\min,\lambda}}}\right).
\end{aligned}
\tag{31}
$$

Here inequality (i) is due to the fact that $\|\boldsymbol{\theta}_\lambda\|_1\geq\|\boldsymbol{\theta}_\lambda\|_2$.

Bounding the second term *i.e.*, $\rho\,\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2+\log[1+\exp\left(-y\,\cdot\,\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t)\right)]$:

First, we note that

$$
\begin{aligned}
&\rho|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\,\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2\\
&\quad\leq\rho|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\,\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2\\
&\qquad+\sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\log[1+\exp\left(-y_i\,\cdot\,\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t_i)\right)]\\
&\quad\overset{(i)}{\leq}|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\log 2\\
&\quad\implies\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|\leq\sqrt{\frac{\log 2}{\rho}},
\end{aligned}
\tag{32}
$$

where (i) is due to the fact that:

$$
\begin{aligned}
&\rho|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\,\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\|^2\\
&\qquad+\sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\log[1+\exp\left(-y_i\,\cdot\,\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t_i)\right)]
\end{aligned}
\tag{33}
$$

$$
\leq\rho\times\mathbf{0}+\sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})}\log[1+\exp\left(-y_i\,\cdot\,\mathbf{0}^\top\boldsymbol{\kappa}_m(t_i)\right)]=|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})|\cdot\log 2
\tag{34}
$$

which is because, $\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))$ is the minimum of Eq. (33).

Hence, we have that:

$$\rho \left\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2 + \log\left[1+\exp\left(-y\,\cdot\,\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t)\right)\right]$$
$$\leq \rho \left\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2 + \log 2 + \boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t) + (\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))^\top\boldsymbol{\kappa}_m(t))^2/2$$
$$\leq \rho \left\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2 + \log 2 + \left\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|\kappa_{\max,m} + \left\|\boldsymbol{\theta}^*_m(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2\kappa^2_{\max,m}/2$$
$$\leq 2\log 2 + \kappa^2_{\max,m}\log 2/(2\rho) + \kappa_{\max,m}\sqrt{\log 2/\rho} \tag{35}$$

First we add the bounds in Eqs. (31) and (35) and then put it into Eq. (25) to obtain the required bound on $F(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})) - F(\mathcal{T})$. The above result together with the bound on $F(\mathcal{S}\cup e) - F(\mathcal{S})$ gives us the required bound on α.

C.2 Monotonicity and approximation guarantee for nonlinear MTPP

Theorem 5 *Assume that the endogenous events follow a nonlinear MTPP with exponentially bounded intensity function, as described in Eq.* (10) *with* $\lambda_{\max} \leq \exp(-\mathbb{E}[\log(\mu\cdot q)])$. *Then the set function* $F(\mathcal{S})$ *defined in Eq.* (7) *is monotone in* $\mathcal{S}$ *if* $\rho \geq 0.25\,a^2/\left(\log\left[1/\lambda_{\max}\right] + \mathbb{E}[\log(\mu\cdot q)]\right)$, *where* μ *and* q *are the intensity and the mark distribution for the exogenous events.*

Proof. From Lemma 9 in Appendix D, we have that: for $e = (y,t) \notin \mathcal{S}$,

$$\begin{aligned}
F(\mathcal{S}\cup\{e\}) - F(\mathcal{S}) &\geq \rho\left\|\boldsymbol{\theta}^*(\mathcal{H}_T\backslash\mathcal{S})\right\|^2 - \log\lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T\backslash\mathcal{S})}(t\,|\,\mathcal{H}_t) + \mathbb{E}[\log(\mu\cdot q)] \\
&\geq \min_{\boldsymbol{\theta}}[\rho\left\|\boldsymbol{\theta}\right\|^2 - \log\lambda_{\max} - a\left\|\boldsymbol{\theta}\right\|] + \mathbb{E}[\log(\mu\cdot q)] \\
&\geq -\frac{a^2}{4\rho} + \left[\log\left(\frac{1}{\lambda_{\max}}\right) + \mathbb{E}[\log\mu + \log q]\right] \\
&> 0
\end{aligned} \tag{36}$$

The last inequality follows from the assumption of ρ. This readily proves the theorem.

Theorem 6 *Given the same conditions in Theorem 5 and* $\int_0^t \left\|\lambda_{\mathbf{0}}(t\,|\,\mathcal{H}_t)\right\|_2 dt \leq \Lambda_{\max} < \infty$, *the set function* $F(\mathcal{S})$ *is* α*-submodular in* $\mathcal{S}$ *with*

$$\alpha \geq \alpha_F = \frac{\log\left[\exp\left(-a^2\mathbb{E}[\log(\mu\cdot q)]/4\right)/\lambda_{\max}\right]}{\Lambda_{\max} + (a+b)\sqrt{\dfrac{\log(1/\lambda_{\min}m_{\min}) + \Lambda_{\max}}{\rho}} + 2\log(1/\lambda_{\min}m_{\min}) + \mathbb{E}[\log(\mu\cdot q)]} \tag{37}$$

Proof. In the first step, we bound $F(\mathcal{S}\cup e) - F(\mathcal{S})$ using the proof of previous theorem, *i.e.*, Theorem 5. Hence, $F(\mathcal{S}\cup e) - F(\mathcal{S}) \geq \log\left[\exp\left(-a^2\mathbb{E}[\log(\mu\cdot q)]/4\right)/\lambda_0\right]$. Next we bound $F(\mathcal{T}\cup\{e\}) - F(\mathcal{T})$. We observe that:

$$\begin{aligned}
&F(\mathcal{T}\cup\{e\}) - F(\mathcal{T}) \\
&\overset{(i)}{\leq} \rho\left\|\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2 - \log\lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(t\,|\,\mathcal{H}_t) \\
&\qquad - \log m_{\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(y\,|\,\mathcal{H}_t) + \mathbb{E}[\log(\mu\cdot q)] \\
&\overset{(ii)}{\leq} \rho\left\|\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2 + (a+b)\left\|\boldsymbol{\theta}^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\| - \log(\lambda_{\min}m_{\min}) + \mathbb{E}[\log(\mu\cdot q)]
\end{aligned} \tag{38}$$

Here, inequality (i) is due to Lemma 10 and inequality (ii) comes from the fact that $\lambda \geq \lambda_0\exp(-a\left\|\boldsymbol{\theta}\right\|)$ and $m \geq \exp(-b\left\|\boldsymbol{\theta}\right\|)$. In order to bound Eq. (38), we bound $\left\|\boldsymbol{\theta}\right\|$. To this aim, we have:

$$\begin{aligned}
&\rho\left\|\theta^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2\left|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})\right| \\
&\leq \rho\left\|\theta^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))\right\|^2\left|\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})\right| \\
&- \sum_{e_i\in\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\})} \log\left[\lambda_{\theta^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(t_i\,|\,\mathcal{H}_{t_i})\cdot m_{\theta^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(y_i\,|\,\mathcal{H}_{t_i})\right] \\
&\qquad + \int_0^T \lambda_{\theta^*(\mathcal{H}_T\backslash(\mathcal{T}\cup\{e\}))}(\tau\,|\,\mathcal{H}_\tau)\,d\tau
\end{aligned} \tag{39}$$

which is less than the value of the objective given by $\boldsymbol{\theta} = \mathbf{0}$. Thus,

$$\rho \left\|\theta^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))\right\|^2 |\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\})| \leq |\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\})| \log(1/\lambda_{\min} m_{\min}) + \Lambda_{\max}$$
$$\implies \left\|\theta^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))\right\|^2 \leq \frac{\log(1/\lambda_{\min} m_{\min}) + \Lambda_{\max}}{\rho} \tag{40}$$

To this end, we have:

$$F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}) \leq \Lambda_{\max} + (a+b)\sqrt{\frac{\log(1/\lambda_{\min} m_{\min}) + \Lambda_{\max}}{\rho}} + 2\log(1/\lambda_{\min} m_{\min}) + \mathbb{E}[\log(\mu \cdot q)] \tag{41}$$

The above result together with the bound on $F(\mathcal{S} \cup e) - F(\mathcal{S})$ gives us the required bound.

C.3 Representation of MTPPs using exponential bounds

Here, we discuss that under some mild conditions, popular MTPP models [16, 86] can admit exponential bounds when the hidden states $\boldsymbol{h}_i$ are bounded, *i.e.*, $||\boldsymbol{h}_i|| < h_{\max}$. Furthermore, we assume that $t_i - t_{i-1} \leq \delta_{\max}$, $t_0 > t_{\min} > 0$.

Exponential boundedness of RMTPP. RMTPP [16] considers exponential intensity function, *i.e.*,

$$\lambda_{\boldsymbol{\theta}}(t_{i+1} \,|\, \mathcal{H}_{t_i}) = \exp(\boldsymbol{w}^\top \boldsymbol{h}_i + \nu(t - t_i) + v). \tag{42}$$

Then we have:

$$\exp(-(h_{\max} + \delta_{\max} + 1)\|[\boldsymbol{w}, \nu, v]\|_2) \leq \lambda_{\boldsymbol{\theta}}(t_{i+1} \,|\, \mathcal{H}_{t_i}) \leq \exp((h_{\max} + \delta_{\max} + 1)\|[\boldsymbol{w}, \nu, v]\|_2) \tag{43}$$

Exponential boundedness of THP. In Transformer Hawkes Process [86], a softplus activation function is used for intensity modeling. More specifically, we have:

$$\lambda_{\boldsymbol{\theta}}(t_{i+1} \,|\, \mathcal{H}_{t_i}) = \sum_{j=1}^{J} \text{Softplus}\left(\alpha_j \frac{t - t_i}{t_i} + \boldsymbol{w}_j^\top \boldsymbol{h}_i + v_j\right) \tag{44}$$

In the following, we formally prove the exponential bounds for THP.

Proposition 8 *Given a Transformer Hawkes Process [86] with the intensity function* (44)*. If in Eq.* (44)*, we have* $u_j = \alpha_j \frac{t - t_i}{t_i} + \boldsymbol{w}_j^\top \boldsymbol{h}_i + v_j \geq u_{\min} > 0$*; the hidden states* $\boldsymbol{h}_i$ *are bounded,* i.e.*,* $||\boldsymbol{h}_i|| < h_{\max}$*; and,* $t_i - t_{i-1} \leq \delta_{\max}$*,* $t_0 > t_{\min} > 0$*, then we have:*

$$\lambda_0 \exp(-a||\boldsymbol{\theta}||_2) \leq \lambda_{\boldsymbol{\theta}}(t_{i+1} \,|\, \mathcal{H}_{t_i}) \leq \lambda_0 \exp(a||\boldsymbol{\theta}||_2) \tag{45}$$

where, $\lambda_0 = J$*,* $a = \sqrt{D}\,(2 + \log 2/u_{\min})(\delta_{\max}/t_{\min} + h_{\max} + 1)$*.*

Proof. Define $\omega = 2 + \log 2/u_{\min}$ Due to the fact that $1 - 1/x \leq \log(x) \leq x - 1$, we have:

$$\begin{aligned}
&\frac{\exp(u_j)}{1 + \exp(u_j)} \leq \text{Softplus}(u_j) \leq \exp(u_j) \\
&\overset{(i)}{\implies} 0.5 \cdot \exp(-2|u_j|) \leq \text{Softplus}(u_j) \leq \exp(2|u_j|), \\
&\overset{(ii)}{\implies} \exp(-\omega|u_j|) \leq \text{Softplus}(u_j) \leq \exp(\omega|u_j|), \\
&\implies \sum_{j=1}^{J} \exp(-\omega|u_j|) \leq \lambda_{\boldsymbol{\theta}}(t_{i+1} \,|\, \mathcal{H}_{t_i}) \leq \sum_{j=1}^{J} \exp(\omega|u_j|),
\end{aligned} \tag{46}$$

where (i) is due to the fact that

$$\exp(-|u_j|)(\exp(u_j) + 1) \leq 2 \implies \frac{\exp(-|u_j|)}{2} \leq \frac{1}{1 + \exp(u_j)} \text{ and } \exp(-|u_j|) \leq \exp(u_j),$$

(ii) is due to the fact that

$$\exp(-2|u_j| - (|u_j|/u_{\min})\log 2) \leq \frac{1}{2}\exp(-2|u_j|) \tag{47}$$

Now we have the following bounds.

$$\begin{aligned}\sum_{j=1}^{J}\exp(\omega|u_j|) &\leq \sum_{j=1}^{J}\exp\left(\omega\left|\alpha_j\frac{t-t_i}{t_i}+\boldsymbol{w}_j^\top\boldsymbol{h}_i+v_j\right|\right)\\ &\overset{(i)}{\leq} J\exp\left(\omega\sum_{j=1}^{J}\|[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\|_1\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)\right)\\ &\overset{(ii)}{\leq} J\exp\left(\omega\left\|\odot_{j=1}^{J}[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\right\|_1\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)\right)\\ &\overset{(iii)}{\leq} J\exp\left(\sqrt{D}\omega\left\|\odot_{j=1}^{J}[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\right\|_2\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)\right). \qquad (48)\end{aligned}$$

Here D is the total number of parameters, *i.e.*, $D = J(\dim(\boldsymbol{w})+2)$; $\odot$ indicates concatenation; inequality (i) is due to that $|\boldsymbol{a}\cdot\boldsymbol{b}| < ||\boldsymbol{a}||_1\, b_{\max}$, inequality (ii) is due to that $\sum_i ||\boldsymbol{a}_i||_1 = ||\odot_i \boldsymbol{a}_i||_1$ and inequality (iii) is due to the fact that $||\boldsymbol{a}||_1 \leq \sqrt{\dim(\boldsymbol{a})}||\boldsymbol{a}||_2$. Moreover, we have:

$$\begin{aligned}\sum_{j=1}^{J}\exp(-\omega|u_j|) &\overset{(i)}{\geq} J\exp\left(-\omega\sum_{j=1}^{J}\left|\alpha_j\frac{t-t_i}{t_i}+\boldsymbol{w}_j^\top\boldsymbol{h}_i+v_j\right|/J\right)\\ &\overset{(ii)}{\geq} J\exp\left(-\omega\sum_{j=1}^{J}\|[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\|_1\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)/J\right)\\ &\overset{(iii)}{\geq} J\exp\left(-\omega\left\|\odot_{j=1}^{J}[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\right\|_1\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)/J\right)\\ &\overset{(iv)}{\geq} J\exp\left(-\sqrt{D}\,\omega\left\|\odot_{j=1}^{J}[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\right\|_2\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)/J\right),\\ &\geq J\exp\left(-\sqrt{D}\,\omega\left\|\odot_{j=1}^{J}[\boldsymbol{w}_j,\ \alpha_j,\ v_j]\right\|_2\left(\delta_{\max}/t_{\min}+h_{\max}+1\right)\right), \qquad (49)\end{aligned}$$

where (i) is due to A.M $\geq$ G.M, inequality (ii) is due to that $|\boldsymbol{a}\cdot\boldsymbol{b}| < ||\boldsymbol{a}||_1\, b_{\max}$, inequality (iii) is due to that $\sum_i ||\boldsymbol{a}_i||_1 = ||\odot_i \boldsymbol{a}_i||_1$ and inequality (iv) is due to the fact that $||\boldsymbol{a}||_1 \leq \sqrt{\dim(\boldsymbol{a})}||\boldsymbol{a}||_2$. Eqs. (48) and (49) give us required parameters for the exponential bounds, *i.e.*, λ_0 and a in Eq. (45).

C.4 Approximation guarantees of Algorithm 1

We first restate the Theorem 7.

Theorem 7 *Assume that the training algorithm* TRAIN() *in Algorithm 1 provides imperfect estimates of the underlying model parameters, with* $\mathcal{L}(\boldsymbol{\theta}^*(\mathcal{H}_T\backslash\mathcal{S});\mathcal{S};\mathcal{H}_T) - \mathcal{L}(\widehat{\boldsymbol{\theta}}(\mathcal{H}_T\backslash\mathcal{S});\mathcal{S};\mathcal{H}_T)) \leq \epsilon$ *for all* $\mathcal{S}$ *and that it runs with* $|\mathcal{V}| = O\left(\frac{|\mathcal{H}_T|}{k}\log(1/\delta)\right)$ *(cf. line number 3). Then,*

$$\mathbb{E}[F(\mathcal{S})] \geq (1-\exp(-\alpha_F^*)-\alpha_F^*\cdot\delta)\cdot F(\mathcal{S}^*) - k\epsilon \qquad (50)$$

where the expectation is taken over many draws of $\mathcal{V}$ *(cf. line number 3),* $\mathcal{S}^*$ *is the solution of the optimization problem* (8) *and* α_F^* *is the submodularity ratio, computed using Theorems 4 and 5.*

Proof Through the rest of this proof, we denote $F(\mathcal{S}) = \mathcal{L}(\boldsymbol{\theta}^*(\mathcal{H}_T\backslash\mathcal{S});\mathcal{S};\mathcal{H}_T)$ and $\hat{F}(\mathcal{S}) = \mathcal{L}(\widehat{\boldsymbol{\theta}}(\mathcal{H}_T\backslash\mathcal{S});\mathcal{S};\mathcal{H}_T))$. So essentially, we assume we do not have access to F but rather an approximation $\hat{F}$. In other words, we pick $i \in \text{argmax}_{j\in\mathcal{V}\backslash\mathcal{S}_i}\hat{F}(j|\mathcal{S}_i)$. Note that for any set $\mathcal{S}$, $|F(\mathcal{S}) - \hat{F}(\mathcal{S})| \leq \epsilon$. Hence, it implies that $F(\mathcal{S}) - \epsilon \leq \hat{F}(\mathcal{S}) \leq F(\mathcal{S}) + \epsilon$. Given this, we have that $F(i|\mathcal{S}) - 2\epsilon \leq \hat{F}(i|\mathcal{S}) = \hat{F}(\mathcal{S}\cup i) - \hat{F}(\mathcal{S}) \leq F(i|\mathcal{S}) + 2\epsilon$. Now given a $\hat{i}$ such that $\hat{F}(\hat{i}|\mathcal{S}) \geq \hat{F}(i|\mathcal{S}), \forall i \notin \mathcal{S}$, we have that $F(\hat{i}|\mathcal{S}) - 2\epsilon \geq F(i|\mathcal{S}) + 2\epsilon$ which implies that:

$$F(\hat{i}|\mathcal{S}) \geq F(i|\mathcal{S}) + 4\epsilon \qquad (51)$$

Next, we study the bound for the stochastic greedy algorithm (Algorithm 1). Following Lemma 2 from [54], at every round of the stochastic greedy algorithm, and combining with Equ. (51), we have:

$$\mathbb{E}(F(a_{i+1}|\mathcal{S}_i)) \geq \frac{1-\delta}{k} \sum_{a \in \mathcal{S}^* \setminus \mathcal{S}_i} [F(a|\mathcal{S}_i) - 4\epsilon] \tag{52}$$

Here a_{i+1} is the element picked at the ith round of stochastic greedy. Also, $\mathcal{S}_i$ is the set obtained by stochastic greedy in the ith step of greedy. Following from α-submodularity, we have that:

$$\sum_{a \in \mathcal{S}^* \setminus \mathcal{S}_i} F(a|\mathcal{S}_i) \geq \alpha F(\mathcal{S}^*|\mathcal{S}_i) \geq \alpha(F(\mathcal{S}^*) - F(\mathcal{S}_i)) \tag{53}$$

This is because while picking the best item in the stochastic greedy algorithm, we have access only to $\hat{F}$ and not F. This implies,

$$\mathbb{E}[F(\mathcal{S}_{i+1}) - F(\mathcal{S}_i)] \geq [(1-\delta)\alpha - 4k\epsilon/\text{OPT}]/k\mathbb{E}[F(\mathcal{S}^*) - F(\mathcal{S}_i)] \tag{54}$$

Here, we denote $F(\mathcal{S}^*) = \text{OPT}$. Next, define $\Gamma_i = \mathbb{E}[f(\mathcal{S}^*) - f(\mathcal{S}_i)]$. We then have the following recursion:

$$\Gamma_i - \Gamma_{i+1} \geq [(1-\delta)\alpha - 4k\epsilon/\text{OPT}]/k\Gamma_i \tag{55}$$

which implies that $\Gamma_{i+1} \leq \{1 - [(1-\delta)\alpha - 4k\epsilon/\text{OPT}]/k\}$. Using induction, we get:

$$\mathbb{E}[F(\mathcal{S}_k)] \geq (1 - (1 - \frac{(1-\delta)\alpha}{k} - 4\epsilon/\text{OPT})^k)F(\mathcal{S}^*) \tag{56}$$

Since $\delta, \epsilon \approx 0$, this gives us an approximation factor of $1 - e^{-\alpha} - \alpha\delta - k\epsilon/OPT$.

Combining everything, this proves the result:

$$\mathbb{E}[F(\mathcal{S})] \geq (1 - \exp(-\alpha_F^*) - \alpha_F^* \cdot \delta) \cdot F(\mathcal{S}^*) - k\epsilon \tag{57}$$

■

D Additional theoretical results

Lemma 9 *Given that $F(\mathcal{S})$ is defined in Eq. (7), then for any $e = (y, t) \in \mathcal{H}_T \backslash \mathcal{S}$, we have:*

$$F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \geq \rho \left\| \boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}) \right\|^2 - \log \lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})}(t \,|\, \mathcal{H}_t) + \mathbb{E}[\log(\mu \cdot q)]. \tag{58}$$

Here $\lambda_{\boldsymbol{\theta}}$ and $m_{\boldsymbol{\theta}}$ can take any functional form.

Proof. Let us first define:

$$\mathcal{L}_\lambda\left(\boldsymbol{\theta}, \mathcal{S}; \mathcal{H}_T\right) = -\rho \left|\mathcal{H}_T \backslash \mathcal{S}\right| \left\|\boldsymbol{\theta}\right\|^2 + \sum_{e_i \in \mathcal{H}_T \backslash \mathcal{S}} \log \lambda_{\boldsymbol{\theta}}(t_i \,|\, \mathcal{H}_{t_i}) - \int_0^T \lambda_{\boldsymbol{\theta}}(\tau \,|\, \mathcal{H}_\tau)\, d\tau \tag{59}$$

$$\mathcal{L}_p\left(\boldsymbol{\theta}, \mathcal{S}; \mathcal{H}_T\right) = \sum_{e_i \in \mathcal{H}_T \backslash \mathcal{S}} \log m_{\boldsymbol{\theta}}(y_i \,|\, \mathcal{H}_{t_i}), \tag{60}$$

Hence, we have that $\mathcal{L}(\bullet) = \mathcal{L}_\lambda\left(\bullet\right) + \mathcal{L}_p\left(\bullet\right)$

$$\begin{aligned}
&F(\mathcal{S} \cup \{e\}) - F(\mathcal{S}) \\
&= \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{S} \cup \{e\})), \mathcal{S} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S}; \mathcal{H}_T\right) + \mathbb{E}[\log(\mu \cdot q)] \\
&= \underbrace{\mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{S} \cup \{e\})), \mathcal{S} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S} \cup \{e\}; \mathcal{H}_T\right)}_{\geq 0} \\
&\qquad + \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S}; \mathcal{H}_T\right) + \mathbb{E}[\log(\mu \cdot q)] \\
&\overset{(i)}{\geq} \Big[\mathcal{L}_\lambda\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}_\lambda\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S}; \mathcal{H}_T\right) \\
&\qquad + \mathcal{L}_p\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}_p\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S}), \mathcal{S}; \mathcal{H}_T\right)\Big] + \mathbb{E}[\log(\mu \cdot q)] \\
&= \rho \left\|\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})\right\|^2 - \log \lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})}(t \,|\, \mathcal{H}_t) - \log m_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})})(y \,|\, \mathcal{H}_t) + \mathbb{E}[\log(\mu \cdot q)] \\
&\overset{(ii)}{\geq} \rho \left\|\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})\right\|^2 - \log \lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{S})}(t \,|\, \mathcal{H}_t) + \mathbb{E}[\log(\mu \cdot q)],
\end{aligned} \tag{61}$$

where (i) is due to the fact that $\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{S} \cup \{e\})) = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}, \mathcal{S} \cup \{e\}; \mathcal{H}_T\right)$; (ii) is due to the fact that $m_{\boldsymbol{\theta}}$ is a probability distribution.

Lemma 10 *Given that $F(\cdot)$ is defined in Eq. (7), then for any $e = (y, t) \in \mathcal{H}_T \backslash \mathcal{T}$, we have:*

$$\begin{aligned}
F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}) \leq\ & \rho \left\|\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))\right\|^2 - \log \lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))}(t \,|\, \mathcal{H}_t) \\
& - \log m_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))}(y \,|\, \mathcal{H}_t) + \mathbb{E}[\log(\mu \cdot q)]
\end{aligned} \tag{62}$$

Here $\lambda_{\boldsymbol{\theta}}$ and $m_{\boldsymbol{\theta}}$ can take any functional form.

Proof. We note that:

$$\begin{aligned}
&F(\mathcal{T} \cup \{e\}) - F(\mathcal{T}) \\
&= \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\})), \mathcal{T} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{T}), \mathcal{T}; \mathcal{H}_T\right) + \mathbb{E}[\log(\mu \cdot q)] \\
&= \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\})), \mathcal{T} \cup \{e\}; \mathcal{H}_T\right) - \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\})), \mathcal{T}; \mathcal{H}_T\right) \\
&\quad + \underbrace{\mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\})), \mathcal{T}; \mathcal{H}_T\right) - \mathcal{L}\left(\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{T}), \mathcal{T}; \mathcal{H}_T\right)}_{\leq 0} + \mathbb{E}[\log(\mu \cdot q)] \\
&\overset{(i)}{\leq} \rho \left\|\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))\right\|^2 \\
&\qquad - \log \lambda_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))}(t \,|\, \mathcal{H}_t) - \log m_{\boldsymbol{\theta}^*(\mathcal{H}_T \backslash (\mathcal{T} \cup \{e\}))}(y \,|\, \mathcal{H}_t) + \mathbb{E}[\log(\mu \cdot q)].
\end{aligned} \tag{63}$$

Here, inequality (i) is due to

$$\boldsymbol{\theta}^*(\mathcal{H}_T \backslash \mathcal{T}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}, \mathcal{T}; \mathcal{H}_T\right).$$

Proposition 11 $\min_x (ax^2 - \log x) = \frac{1}{2} + \log(\sqrt{2a})$.

Proof. Differentiating with respect to x, we have: $x = 1/\sqrt{2a}$, which readily gives the result.

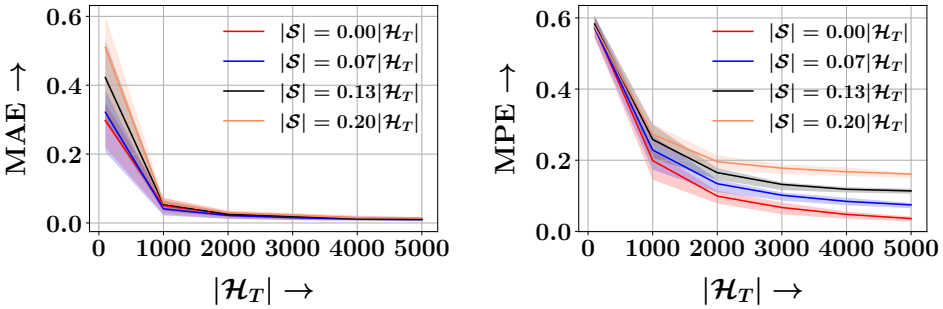


Figure 3: Variation of mean absolute error $\mathbb{E}(|t - \widehat{t}|)$ and mark prediction error (MPE) $\Pr(y \neq \hat{y})$ against the number of observations $|\mathcal{H}_T|$ on synthetic data.

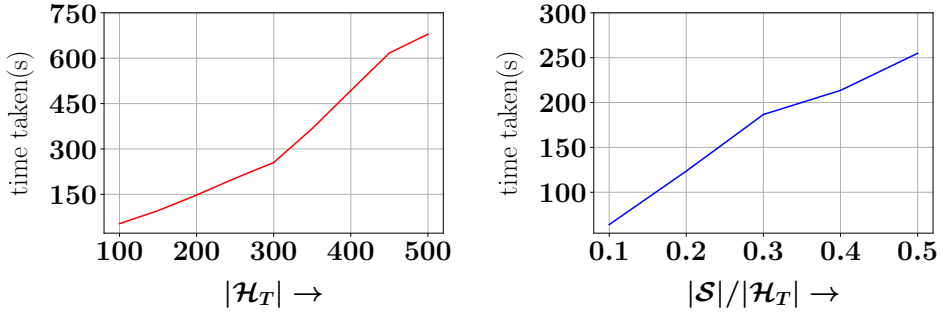


Figure 4: Scalability analysis. Variation of time taken against $|\mathcal{H}_T|$ and $|\mathcal{S}|/|\mathcal{H}_T|$. In the left figure, we set $|\mathcal{S}|/|\mathcal{H}_T| = 0.3$ and in the right figure, we set $|\mathcal{H}_T| = 200$.

E Additional experiments on synthetic data

E.1 Quantitative analysis

Here, we investigate the variation of performance against the number of observations $|\mathcal{H}_T|$ on synthetic data for various values of k, the prespecified number of exogenous events. Figure 3 summarizes the results, which shows that, (i) as the number of observed events are increasing, the prediction error decreases; and, (ii) as k increases, the prediction error increases.

E.2 Scalability analysis

Next, we analyze the time taken by our method against $|\mathcal{H}_T|$— the total number of observations and $|\mathcal{S}|/|\mathcal{H}_T|$, the prespecified fraction of exogenous events. Figure 4 summarizes the figure which shows that, for a sequence with $|\mathcal{H}_T| = 500$, our method takes less than 13 minute to converge in single AMD Ryzen 7 3700X CPU with 32GB RAM machine.

Dataset	$\lvert\mathcal{H}(T)\rvert$	$\lvert\mathcal{C}\rvert$	Class imbalance factor
Club	9409	3	3.88
Election	1584	3	2.62
Series	16318	3	1.83
Verdict	10691	3	2.13
BookOrder	15000	2	1.04

Table 4: Datasets statistics. Class imbalance factor indicates the ratio of highest to lowest occurring marks.

F Additional details about the experiments with real data

F.1 Dataset details

We use five datasets in our experiments for evaluation. The first four datasets were used in previous work [82, 40, 16].

1. **Club** [82, 40]: It gathers the tweets on Barcelona getting the first place in La-liga, from May 8 to May 16, 2016.
2. **Election** [82, 40]: It gathers the tweets on British election, from May 7 to May 15, 2015.
3. **Series** [82]: It gathers the tweets on promotion on the TV show "Games of Thrones", from May 4 to May 12, 2015.
4. **Verdict** [40]: It gathers the tweets on the verdict for the corruption-case against Jayalalitha, an Indian politician, from May 6 to May 17, 2015.
5. **BookOrder** [16]: It gathers the limited book order data from NYSE of the high frequency transactions for a stock in one day.

Table 4 lists some of the statistics of theses datasets. For an event $e = (y, t)$, the arrival times are available in each dataset. For the BookOrder dataset, mark y is available in the data. However for the Twitter datasets, we extract y as follows: $y = +1$ if $s \in (0, 1]$, $y = 0$ if $s = 0$ and $y = -1$ if $s \in [-1, 0)$, where s is the sentiment of the message.

F.2 Additional implementation details for base MTPP models

(i) RMTPP [16] [2]: Here we select batch size 64, initial learning rate 0.1 and L_2 penalty as $1e-3$. (ii) THP [86] [3]: Here we select batch size 16, initial learning rate $1e-4$ and L_2 penalty as 0 as the implementation contains no L_2 penalty. In addition, we used dropout rate 0.3, model depth 64, hidden layer size 128, head number 4 and layer number 4. (iii) SAHP [84] [4]. Here we select batch size 16, initial learning rate $1e-4$ and L_2 penalty as $3e-4$. In addition, we used dropout rate 0.1, model depth 16, hidden layer size 16, head number 8 and layer number 4. (iv) Linear model: For experiments with synthetic data, we used linear models[5][72] to build event simulator and parameter estimator.

F.3 Compute resource

We used a single AMD Ryzen 7 3700X CPU with 32GB RAM machine, no graphical card is needed given all the models and data size in the experiments are moderate.

[2] https://github.com/musically-ut/tf_rmtpp

[3] https://github.com/SimiaoZuo/Transformer-Hawkes-Process

[4] https://github.com/QiangAIResearcher/sahp_repo

[5] https://github.com/sandeepsoni/MHP

G Additional experiments with real data

	Time prediction error in terms of MAE $\mathbb{E}(\lvert t - \widehat{t}\rvert)$		**Mark prediction error** $\Pr(y \neq \widehat{y})$	
	TPP-SELECT	Base MTPP	TPP-SELECT	Base MTPP
Club	**3.78±0.00**	3.80±0.00	**0.49±0.01**	0.79±0.02
Election	**13.66±0.00**	13.71±0.01	**0.62±0.01**	0.79±0.03
Series	**1.98±0.01**	2.13±0.02	**0.62±0.00**	0.67±0.01
Verdict	**3.50±0.00**	3.58±0.05	**0.65±0.00**	0.69±0.02
BookOrder	**0.29±0.01**	0.55±0.05	**0.38±0.01**	0.49±0.01

Table 5: Performance in terms of $\mathbb{E}(\lvert t - \widehat{t}\rvert)$, *i.e.*, the mean absolute error (MAE) for time prediction (left half) and the misclassification error $\Pr(y \neq \widehat{y})$ for the mark prediction (right half) for TPP-SELECT and the corresponding base MTPP model which is Self Attentive Hawkes Process (SAHP) in this case. Here, this base MTPP model is learned over the entire sequence of observed events. In all experiments, we considered 60% training, 20% validation and 20% test set.