Markov-Modulated Hawkes Processes for Sporadic and Bursty Social Interactions

Modeling event dynamics is central to many disciplines. Patterns in observed social interaction events can be commonly modeled using point processes. Such social interaction event data often exhibits self-exciting, heterogeneous and sporadic trends, which is challenging for conventional models. It is reasonable to assume that there exists a hidden state process that drives different event dynamics at different states. In this paper, we propose a Markov Modulated Hawkes Process (MMHP) model for learning such a mixture of social interaction event dynamics and develop corresponding inference algorithms. Numerical experiments using synthetic data demonstrate that MMHP with the proposed estimation algorithms consistently recover the true hidden state process in simulations, while email data from a large university and data from an animal behavior study show that the procedure captures distinct event dynamics that reveal interesting social structures in the real data.

1. Introduction. Understanding social interaction dynamics, such as event occurrences with a temporally heterogeneous intensity, has become an important topic in many research disciplines. For example, user behaviors and interactions on social networking platforms and online service providers are of great importance for resource allocation and user experience improvement. Many real-world social interaction event dynamics are sporadic in nature, rife with irregular event-intense intervals, and heterogeneous in event densities. Study shows that group-living animals dynamically shift their interaction behavior in order to stabilize the social system (Williamson et al., 2017), while the timing of many human actions demonstrates a bursty and heavy-tailed pattern (Barabasi, 2005). It is easy to explain the bursty nature of many event dynamics through the example of an email network: during *active* hours, one individual sends an initial email to *engage* another individual, which acts as a *trigger* that leads to a stream of interactions that follows. These active hours will eventually give way to an *inactive* intersession, during which email arrivals have no trigger effects. Such sporadic event dynamics with stochastic *inactive-active* transitions can be found in many real-world soical interactions, whose irregularity poses challenges for modeling and understanding of the underlying data generating mechanisms.

One example of dynamics of this form is the times of emails between users in a large university, as first collected and analysed by Kossinets and Watts (2006). Figure 1 illustrates emails from one user to another in this network over one academic semester (122 days). The data shows distinct periods of activity, such as at the start and end of this time period, along with periods with less frequent interactions.

Common practices in modeling social interaction event dynamics are generally grouped into two categories. The first type computes aggregated counts of events that are captured in fixed-length time intervals, and then applies time series models for count data (e.g., Blei and Lafferty 2006). The second type directly models continuous-time event occurrences via conditional intensity functions (e.g., Weiss et al. 2012). The first approach requires unnecessary aggregation of the data, which inevitably leads to information loss. In most cases, the time series models make assumptions about the true data generating process that are hard to validate. Another common challenge with the first approach is choosing the "right" length of time intervals that strikes a balance between count sparsity and information loss, as demonstrated further in Supplement S1.1. This can be especially challenging when arrivals are both sporadic and bursty. The continuous-time approach is a more direct modeling of event dynamics. In recent literature, efforts along
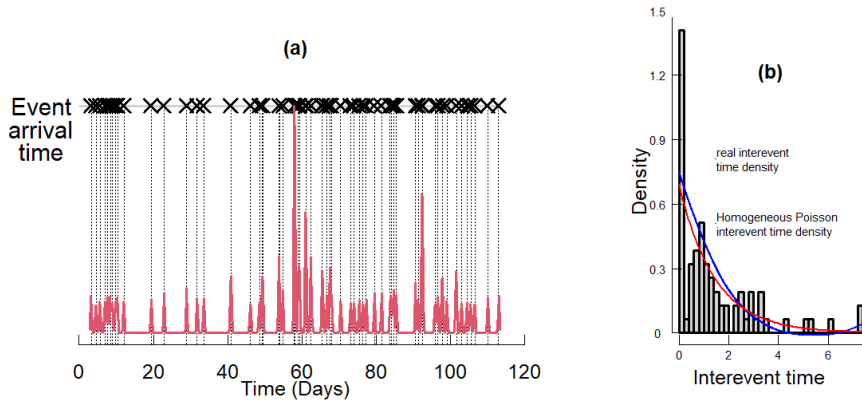
Fig 1: Observed social interactions consisting of emails sent between a pair of users in a large university, as described in Kossinets and Watts (2006). **(a)** describes event arrival times over one semester, with a kernel density estimate for a small bandwidth overlaid, while **(b)** shows the distribution of interevent waiting times. At certain times events arrive at a very high density. The kernel density (red line) is added to help illustrate these events.

3

this line often treat the observed event arrival times as (heterogeneous) point processes and model their intensity functions with incorporated excitatory triggers (Simma and Jordan, 2010; Zhao et al., 2015), event history (Perry and Wolfe, 2013), and/or latent Markov processes (Scott and Smyth, 2003).

Poisson processes are the most widely used models for social interaction event arrival times. These models assume a constant event intensity over time and independent event arrivals. For a Poisson process with rate λ, the waiting time between events follows an exponential distribution with mean $1/\lambda$. However, in practice, the timing of events often follows non-Poisson patterns (Barabasi, 2005). Figure 1-(b) displays the distribution of interevent waiting times for the social interactions between a pair of email users shown in Figure 1-(a). There are departures from an exponential distribution, with the rate being the maximum likelihood estimate. Most notable are departures at the two ends of the distribution that correspond to bursty arrivals and a heavy-tail in waiting times, which have been noted in the literature (Barabasi, 2005).

The Hawkes process (Hawkes, 1971), a self-exciting process, has been proposed as an alternative to address non-Poisson bursts in event dynamics. In a Hawkes process, at the arrival of an event, the event occurrence intensity is elevated. This boost in event rate is sustained for a short period of time that follows. Hawkes processes have been shown to capture bursty patterns in human activities reasonably well (Linderman and Adams, 2014; Wang et al., 2016), but are inadequate to address the existence of extended *'silent period'* and isolated events. Under a Hawkes process, once an event has occurred, it will always induce an incentive for future events to occur in a short period of time immediately following the "triggering" event. However, in reality, long intervals of inactivity or low activity rate between bursts of events are ubiquitous. A more flexible model is needed to address the heavy-tailed distribution of interevent waiting times.

One model that addresses such heterogeneity in interevent waiting times is the Markov Modulated Poisson Process (MMPP) (Fischer and Meier-Hellstern, 1993), which is a doubly stochastic Poisson process with its arrival rate modulated by an underlying Markov process. It assumes that the rate of event arrivals depends on a latent state variable. Conditioning on a given latent state, the arrival of events follows a homogeneous Poisson process. We fit MMPP to data on interactions from the same pair of directed emails in Figure 1, see Figure 2-(a). The inferred latent MMPP state (blue line) fails to capture the different event dynamics patterns. In other words, the observed sporadic event dynamics cannot be explained simply by different rates of incidents. Rather, they suggest different levels of *temporal dependence*.
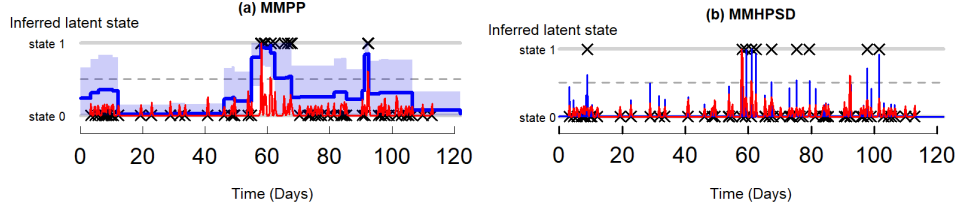
Fig 2: Conventional models applied to the same set of interaction times from a pair of email users as in Figure 1. **(a)** Interactions with inferred latent states (black crosses) by Markov Modulated Poisson Process (MMPP) (Fischer and Meier-Hellstern, 1993) with latent trajectory (blue line) and one-standard-error band (blue shade) bounded to be between 0 and 1.**(b)** The estimated latent state using Markov Modulated Hawkes Process with stepwise decay (MMHPSD) (Wang et al., 2012). Overlaid is a KDE with small bandwidth (red line). Both MMPP and MMHPSD failed to detect segments of different types of social interactions, as further demonstrated in Supplement S1.2.

In this paper, we address the modeling of sporadic interevent waiting times, commonly found in real-world social interaction event dynamics. Combining Markov modulation and the Hawkes process, we propose a Markov Modulated Hawkes Process (MMHP) model and develop corresponding inference algorithms. As shown in Figure 2, the interevent time density from real data includes both short "bursty periods" and extented "silent periods" with isolated events. Our model, on one hand, will address the limitation of the Hawkes process in capturing "silent periods" and isolated events. On the other hand, it will extend the flexibility of the MMPP to allow "bursty periods".

Wang et al. (2012) considered a related strategy, where a Hawkes process with step-wise decay was introduced into the MMPP framework (MMH-PSD) to model variation in seismicity during earthquake sequences. From a modeling perspective, MMHPSD assumes that each event occurrence creates a constant influence on the intensity function that accumulates with that of other events. This influence is reduced to a lower constant by each event that occurs afterwards. This assumption is not flexible enough to describe certain social event arrivals as it ignores time-decaying effects of previous events as time elapses. The estimation of MMHPSD was implemented using

the EM algorithm (Dempster et al., 1977), where the M-step computation highly relies on the piece-wise constant assumption. It is therefore difficult to generalize this procedure to the more widely-used exponential kernel for Hawkes processes. The blue line in Figure 2-(b) shows the fitted latent state using the MMHPSD model, where we can see that the latent process is 'activated' by nearly every event occurrence, classifying several events in the active state. When there is no event, the state immediately drops to state 0. This is the drawback in using a piece-wise constant intensity function that renders the inference dependent only on local patterns (event versus no events). It fails to detect a stable global latent process that represents a mixture of event dynamics.

Our proposed model considers the Hawkes process with exponential decay, the original more general definition of the Hawkes process. This allows us to efficiently and explicitly model the extent of influence that past events have on the arrival intensity of future events. We derive a novel inference algorithm to solve the computational challenge, providing a close mean-field variational approximation (Blei et al., 2017) of the original likelihood. This novel approximation allows the model inference to be carried out using techniques from the forward-backward algorithm (Rabiner, 1989) and the Viterbi algorithm (Rabiner, 1989), which can be easily generalized to estimate other related models, e.g., MMPP and MMHP with stepwise decay or other kernel functions. This inference procedure can be incorporated into posterior inference under a Bayesian framework that allows us to quantify uncertainty in the model estimates, especially for the latent state process. We evaluate the performance of MMHP using experiments on synthetic data, real email data from Kossinets and Watts (2006) and mice interaction data from Williamson et al. (2016). MMHP is shown to have excellent model estimation and reliable recovery of the latent states (*active* versus *inactive*) for synthetic event dynamics with ground truth. MMHP is also better able to capture the dynamics underlying email interactions than previous models. When applied to interaction dynamics among cohorts of male mice, MMHP identifies two types of fighting activity states with different social structures.

2. Markov Modulated Hawkes Processes. We first introduce the proposed MMHP model for sporadic event dynamics given observed event arrival times. We start with the notation that is necessary for our discussion. Then, we lay down the background on point processes in general with a focus on the Hawkes process. Finally, we introduce the proposed MMHP, a latent variable model with Hawkes process modulated by a Markov process.

Notation for event arrival time data. We consider event arrival time data that consists of all event history up to a *final-observation* time T: $\mathcal{H}(T) = \{t_m\}_{m=0}^{M}$, where $t_0 = 0$, $t_M = T$, and M is the total number of events. The sequence of interevent waiting times is denoted by $\{\Delta t_m := t_m - t_{m-1}\}_{m=1}^{M}$, which is equivalent to the event time $\mathcal{H}(T)$.

Background on point process models. An equivalent representation of a point process $\mathcal{H}(T) = \{t_m\}_{m=0}^{M}$ is via a *counting process*. Let $N(t)$ be a right-continuous point process that records the number of events observed during the interval $[0, t]$. The conditional intensity function given the history up to time t, $\mathcal{H}(t)$, is

$$\lambda(t|\mathcal{H}(t)) = \lim_{\Delta t \to 0} \frac{Pr(N(t + \Delta t) - N(t^-) = 1|\mathcal{H}(t))}{\Delta t}.$$

The likelihood function for a sequence of events up to time T, $\mathcal{H}(T) = \{t_1 < ... < t_M\}$, is then

$$(2.1) \qquad \prod_{m=1}^{M} \lambda(t_m|\mathcal{H}(t_m)) \exp\Big\{ - \int_0^T \lambda(u|\mathcal{H}(u))du\Big\}.$$

The *Hawkes process* (Hawkes, 1971) is a self-exciting process that can explain bursty patterns in event dynamics. For a univariate model with the widely used exponential kernel [1], the intensity function is defined as

$$(2.2) \qquad \lambda(t) = \lambda_1 + \int_0^t \alpha e^{-\beta(t-s)} dN_s = \lambda_1 + \sum_{t_m < t} \alpha e^{-\beta(t-t_m)},$$

where $\lambda_1 > 0$ specifies the baseline intensity, $\alpha > 0$ calibrates the instantaneous boost to the event intensity at each arrival of an event, and $\beta > 0$ controls the decay of past events' influence over time. Hawkes processes have successfully been used to model dynamics in fields such as finance (Hawkes, 2018) and Neuroscience (Linderman and Adams, 2014).

Modulation by a latent Markov process. In Fischer and Meier-Hellstern (1993), the original Poisson process was extended to be *modulated* by a latent continuous-time Markov chain (CTMC), primarily to address the commonly non-Poisson pattern where the event dynamics alternate between long waiting times and intervals of more intensive events. More specifically, this

---

[1]Other kernel functions such as a powerlaw kernel have also been used in applications in seismology and finance.

Markov Modulated Poisson Process (MMPP) model is a doubly stochastic Poisson process whose arrival rate is given by $\lambda_{Z(t)}$. $Z(t)$ is an irreducible Markov process with R-states that is independent of the arrival process. When the Markov process $Z(t)$ is in state r ($r \in \{1, ..., R\}$), arrivals occur according to a Poisson process of rate λ_r. In this paper, we will consider a two-state $Z(t)$ that takes values 0 or 1. All the following discussion can be generalized to an R-state Z(t).

MMPP assumes a constant event intensity conditioning on the latent state $Z(t)$. For sporadic event dynamics with bursts, isolated incidents and long waiting times, we propose to use Hawkes process to model a self-exciting $\lambda_1(t)$ instead of using a constant rate λ_1, when the underlying Markov process is in the active state ($Z(t) = 1$). This Hawkes process $\lambda_1(t)$ considers the whole event history up to current time t, i.e., $\mathcal{H}(t)$. When $Z(t) = 0$, the point process follows homogeneous Poisson process with rate λ_0. Modulating the intensity function using a latent two-state Markov process $Z(t)$ allows us to extract segments with heterogeneous event dynamics. $Z(t)$ is described by an initial probability vector $\delta := (1 - \delta_0, \delta_0)$, and an infinitesimal generator matrix Q, where

$$Q = \begin{bmatrix} q_{1,1} & q_{1,0} \\ q_{0,1} & q_{0,0} \end{bmatrix} = \begin{bmatrix} -q_1 & q_1 \\ q_0 & -q_0 \end{bmatrix}.$$

For each $t \geq 0$, there is a probability transition matrix, denoted as $\mathbf{P}(t) := [\mathbf{P}_{ij}(t)]_{i,j \in \{0,1\}}$. Each entry $\mathbf{P}_{ij}(t)$ is defined as the probability that the chain will be in state j at time $u + t$ ($t > 0$) given the chain is in state i at time u, i.e. for each $u \geq 0$,

$$\mathbf{P}_{ij}(t) = P(Z(u + t) = j | Z(u) = i).$$

The likelihood of the full trajectory of the hidden state $\{Z(t), t \leq T\} := \mathcal{Z}(T)$ can be written in terms of the following sufficient statistics: initial state $Z(0)$, the number of jumps K and the successive transition time points $\{u_1, ..., u_K\}$, given parameters δ, q_1 and q_0. Further, denote s_k as the state of $Z(t)$ during $[u_{k-1}, u_k)$, and $\Delta u_k = u_k - u_{k-1}$, as shown in Figure 3. Then the likelihood function is written as

$$P(\mathcal{Z}(T)|\delta, Q) = \delta_0^{I\{Z(0)=0\}} (1-\delta_0)^{I\{Z(0)=1\}} \Big[\prod_{k=1}^{K} q_{s_k} e^{-q_{s_k} \Delta u_k} \frac{q_{s_k, s_{k+1}}}{q_{s_k}} \Big] e^{-q_{s_{K+1}} \Delta u_{K+1}}.$$

Let Θ denote the entire set of parameters, i.e., $\{\lambda_0, \lambda_1, \alpha, \beta, \delta, q_1, q_0\}$. The complete-data likelihood for Θ, under MMHP, is then

(2.3) $\qquad P(\mathcal{H}(T), \mathcal{Z}(T)|\Theta) = P(\mathcal{Z}(T)|\Theta) \times P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T)),$
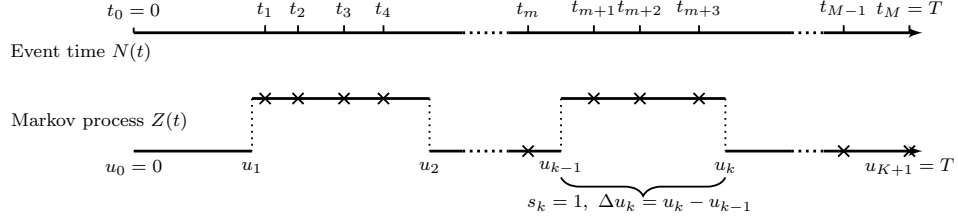
Fig 3: An illustration of the MMHP model.

where $P(\mathcal{Z}(T)|\Theta)$ is for the latent Markov process, which is not dependent on the observation $\mathcal{H}(T)$. $P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T))$ is for observed event data conditioning on the latent process. More specifically,

(2.4)

$$P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T)) = \prod_{m=1}^{M} \lambda_{Z(t_m)}(t_m|\mathcal{H}(t_m)) \exp\{-\int_0^T \lambda_{Z(u)}(u|\mathcal{H}(u))du\}.$$

3. Bayesian inference of MMHP.

3.1. *Inference of model parameters.* In this paper, we adopt a Bayesian framework for the inference of MMHP. We are interested in the posterior distribution of parameter set Θ, given a proposed prior distribution $\pi(\Theta)$ and observed $\mathcal{H}(T)$. It can be written that,

(3.1) $$P(\Theta|\mathcal{H}(T)) \propto \pi(\Theta)P(\mathcal{H}(T)|\Theta) = \pi(\Theta) \sum_{\mathcal{Z}(T)} P(\mathcal{H}(T), \mathcal{Z}(T)|\Theta).$$

For better computational stability and convergence, in (3.1), we integrate out the full latent state trajectory. However, the exact marginalization is computationally infeasible over the entire set of possible full trajectories of continuous-time Markov chains (CTMC), $\mathcal{S} := \{\mathcal{Z}(T)\}$. In practice, we are more interested in the latent state at the time of each event (and also the initial state), i.e. $\tilde{\mathcal{Z}}(T) := \{Z_0, Z_1, Z_2, ..., Z_M\}$. In order to find an efficient approximation of (3.1), we divide the computation into subsets of $\mathcal{S}$ with

$$\mathcal{S} = \bigcup_{\tilde{\mathcal{Z}}(T)} \mathcal{S}_{\tilde{\mathcal{Z}}(T)}, \ \mathcal{S}_{\tilde{\mathcal{Z}}(T)} := \{\mathcal{Z}(T)|Z(t_m) = Z_m, Z_m \in \tilde{\mathcal{Z}}(T), m = 1, ..., M\}.$$

Here $\mathcal{S}_{\tilde{\mathcal{Z}}(T)}$ is the set of CTMC trajectories that result in $\tilde{\mathcal{Z}}(T)$. Given each possible combination of observed event states, $\tilde{\mathcal{Z}}(T)$, the entire set $\mathcal{S}$ can be decomposed into the union of 2^{M+1} such $\mathcal{S}_{\tilde{\mathcal{Z}}(T)}$ realizations.

The summation in (3.1) over a CTMC family of infinite dimension is then divided into 2^{M+1} parts,

$$(3.2) \qquad \sum_{\mathcal{Z}(T)} P(\mathcal{H}(T), \mathcal{Z}(T)|\Theta) = \sum_{\tilde{\mathcal{Z}}(T)} \sum_{\mathcal{Z}(T) \in \mathcal{S}_{\tilde{\mathcal{Z}}(T)}} P(\mathcal{H}(T), \mathcal{Z}(T)|\Theta).$$

For each term in the outer summation of (3.2), a variational approximation solution is proposed in section 3.2 (Cohn et al., 2009). Based on these approximated likelihood summands, we construct an inference procedure for the whole marginalization in (3.1), utilizing components of the forward-backward algorithm (Rabiner, 1989). This allows us to implement an MCMC sampling algorithm and derive the posterior distribution of Θ. The latent Markov process trajectory is then estimated by the Viterbi algorithm (Rabiner, 1989) and interpolated by maximizing the likelihood of $\mathcal{Z}(T)$, as outlined in section 3.3. Empirical results (not shown) indicate that this inference procedure, obtaining both the posterior distribution Θ and the latent process trajectory, scales linearly in the number of events observed.

3.2. *Likelihood approximation.* Working with each summand in (3.2),

$$\sum_{\mathcal{Z}(T) \in \mathcal{S}_{\tilde{\mathcal{Z}}(T)}} P(\mathcal{H}(T), \mathcal{Z}(T)|\Theta) = \sum_{\mathcal{Z}(T) \in \mathcal{S}_{\tilde{\mathcal{Z}}(T)}} P(\mathcal{Z}(T)|\Theta) \times P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T))$$

$$(3.3) \qquad\qquad = P(\tilde{\mathcal{Z}}(T)|\Theta) \times E_{\mathcal{Z}|\tilde{\mathcal{Z}}}[P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T))].$$

The calculation of $P(\tilde{\mathcal{Z}}(T)|\Theta)$). This is the likelihood of the Markov process embedding at event times, which can be calculated as

$$(3.4) \qquad P(\tilde{\mathcal{Z}}(T)|\Theta) = \delta_0^{I\{Z_0=0\}} (1-\delta_0)^{I\{Z_0=1\}} \prod_{m=1}^{M} \mathbf{P}_{Z_{m-1}, Z_m}(\Delta t_m).$$

Following Kolmogorov's forward equation

$$\frac{d\mathbf{P}_{ij}(t)}{dt} = \sum_{k \neq j} q_{kj} \mathbf{P}_{ik}(t) - \left(\sum_{k \neq j} q_{jk} \right) \mathbf{P}_{ij}(t),$$

we can readily calculate the matrix $\mathbf{P}(t) = \exp(Qt)$. When there are two states, it has an explicit form as follows,

$$\mathbf{P}(t) = \begin{bmatrix} \mathbf{P}_{11}(t) & \mathbf{P}_{10}(t) \\ \mathbf{P}_{01}(t) & \mathbf{P}_{00}(t) \end{bmatrix} = \frac{1}{q_0 + q_1} \begin{bmatrix} q_0 + q_1 e^{-(q_0+q_1)t} & q_1 - q_1 e^{-(q_0+q_1)t} \\ q_0 - q_0 e^{-(q_0+q_1)t} & q_1 + q_0 e^{-(q_0+q_1)t} \end{bmatrix}.$$

Variational approximation of $E_{\mathcal{Z}|\tilde{\mathcal{Z}}}[P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T))]$. We need to approximate the marginalization of all possible trajectories of $\mathcal{Z}(T)$ given a $\tilde{\mathcal{Z}}(T)$. More specifically, we approximate the expectation in the following equation,

(3.5)

$$E_{\mathcal{Z}|\tilde{\mathcal{Z}}}\left[P(\mathcal{H}(T)|\Theta, \mathcal{Z}(T))\right] = \prod_{m=1}^{M} \lambda_{Z_m}(t_m|\mathcal{H}(t_m)) E_{\mathcal{Z}|\tilde{\mathcal{Z}}}\left[\exp\left\{-\int_0^T \lambda_{Z(u)}(u|\mathcal{H}(u))du\right\}\right],$$

where, for $u \in [0, T]$,

$$\lambda_{Z(u)}(u|\mathcal{H}(u)) = \lambda_1(u|\mathcal{H}(u))Z(u) + \lambda_0(1 - Z(u)) \coloneqq \lambda(Z(u), u).$$

We then derive a variational approximation for

(3.6) $$E_{\mathcal{Z}|\tilde{\mathcal{Z}}}\left[\exp\left(-\int_0^T \lambda(Z(u), u)du\right)\right].$$

Variational approximation is a general numerical tool for approximating any integral and has been widely used for approximating full posteriors (Blei et al., 2017). Cohn et al. (2009) pointed out that inference problems related to CTMC are computationally intractable and it is necessary to construct an approximation. Given the observations of the event times, our problem setting is similar to Cohn et al. (2009), except that our observations are point process. Under our model assumptions, we assume the *evidence* of the states is $\tilde{Z}$, e.g., the latent states at the event times. Thus, for our approximation task, we consider the Markov process density family defined in Cohn et al. (2009):

$$\mathcal{M}_{\tilde{Z}} \coloneqq \{\mu_z(t), \gamma_{z_1, z_2}(t) : 0 \leq t \leq T\}.$$

Here

$$\mu_z(t) = P(Z(t) = z), \ z \in \{0, 1\}$$

$$\gamma_{z_1, z_2}(t) = \lim_{h \to 0} \frac{P(Z(t) = z_1, Z(t + h) = z_2)}{h}, \ z_1, z_2 \in \{0, 1\}, z_1 \neq z_2,$$

which satisfies $\mu_{z_m}(t_m) = 1$, $\mu_z(t_m) = 0, z \neq z_m$, and other positive, normalizing and marginal conditions as stated in Definition 1 of Cohn et al. (2009). This definition is critical for the validity of the variational distribution as proven by Cohn et al. (2009). Although our computation only relies on the definition of $\mu_z(t)$, we include the second moment definition for the completeness of the discussion.

Given $f_{\tilde{Z}} \in \mathcal{M}_{\tilde{Z}}$, consider the integration $E_{f_{\tilde{Z}}}[\exp(-\int_0^T \lambda(Z(u), u)du)]$. It can be readily approximated by applying Jensen's inequality as follows.

$$(3.7) \qquad E_{f_{\tilde{Z}}}[\exp{(-\int_0^T \lambda(Z(u),u)du)]}$$

$$\geq \exp{(-E_{f_{\tilde{Z}}}[\int_0^T \lambda(Z(u),u)du])}$$

$$= \exp{(-\int_\Sigma f_{\tilde{Z}}(\sigma) \int_0^T \lambda(Z(u),u)dud\sigma)}$$

$$= \exp{\left(-\int_0^T \sum_z \int_\Sigma f_{\tilde{Z}}(\sigma)I_{Z(u)=z}\lambda(z,u)d\sigma du\right)}$$

$$= \exp{\left(-\int_0^T \sum_z \lambda(z,u)\mu_z(u)du\right)}.$$

Following Theorem 6 of Cohn et al. (2009), (3.6) can be approximated by the integral (3.7) evaluated at the $f_{\tilde{Z}}$ that minimizes the Kullback-Leibler divergence (Kullback and Leibler, 1951) from the process of interest given $\tilde{\mathcal{Z}}(T)$. $f_{\tilde{Z}}$ satisfies the condition that,

$$\mu_z(u) = \frac{\mathbf{P}_{z_{m-1},z}(u-t_{m-1})\mathbf{P}_{z,z_m}(t_m-u)}{\mathbf{P}_{z_{m-1},z_m}(\Delta t_m)}, \text{ for } u \in [t_{m-1},t_m).$$

Given the formulation of $\mu_z(u)$, we are able to calculate the integration in the (3.7)

$$(3.8) \quad \int_{t_{m-1}}^{t_m} \sum_z \lambda(z,u)\mu_z(u)du$$

$$= \sum_z \int_{t_{m-1}}^{t_m} \lambda(z,u)\frac{\exp{(Q(u-t_{m-1}))}_{z_{m-1},z}\exp{(Q(t_m-u))}_{z,z_m}}{\exp{(Q\Delta t_m)}_{z_{m-1},z_m}}du.$$

Thus, we can calculate the summand in (3.2) by combining (3.3), (3.4), (3.5) (3.7) and (3.8). This variational approximation allows marginalizing the likelihood over $\left\{\mathcal{Z}(T) \in \mathcal{S}_{\tilde{\mathcal{Z}}(T)}\right\}$ to be computationally tractable. Given this intermediate approximated result, which will be denoted as

$$\tilde{P}_\Theta(\mathcal{H}(T),\tilde{\mathcal{Z}}(T)) := \sum_{\mathcal{Z}(T)\in\mathcal{S}_{\tilde{\mathcal{Z}}(T)}} P(\mathcal{H}(T),\mathcal{Z}(T)|\Theta),$$

we will then marginalize over $\tilde{Z}(t)$. This can be carried out in linear time by using the forward component of the forward-backward algorithm (Rabiner, 1989), which is described in Appendix A.3.

Bayesian inference of MMHP using MCMC. We impose weakly informative priors for the model parameters: $\delta_0 \sim U(0,1)$, $\alpha \sim \mathrm{N}(0,5)$, $\beta \sim \log \mathrm{N}(0, 0.5)$, $\lambda_1 \sim \log \mathrm{N}(0,1)$, $\lambda_0 < \lambda_1$, $q_0, q_1 \sim \log \mathrm{N}(-1, 1)$. λ_1 is baseline intensity for the Hawkes process, which is greater than the rate in state 0, λ_0, in order to address model identifiability issue. As part of the MCMC sampler, we incorporate the above likelihood approximation algorithm and forward algorithm as in Appendix A.3 to obtain posterior draws of the parameters. See Algorithm 1 in Appendix A.1 for more details. Computation was carried out in Stan (Guo et al., 2014).

3.3. *Inference of the latent process.* Given a posterior draw $\hat{\Theta}$, we may infer the most likely sequence of hidden states, $\hat{z}_{1:M}$, corresponding to the observed events, $\hat{z}_{1:M}$. We apply the Viterbi algorithm (Forney, 1973), which maximizes the conditional probability: $P(Z_{1:M} = \hat{z}_{1:M} | \Theta, \Delta t_{1:M})$, as shown in Algorithm 2 in Appendix A.2. Then, the full latent trajectory $\mathcal{Z}(T)$ given $\hat{\Theta}$ and $\hat{z}_{1:M}$ is interpolated by maximizing the likelihood of there being no event between two observed events, given the estimated states at these two events' times and the parameter estimate, i.e., $P(Z(t) | Z_m = \hat{z}_m, Z_{m+1} = \hat{z}_{m+1}, \hat{\Theta}, \Delta t_{1:m+1}), t \in [t_m, t_{m+1})$. Based on the posterior latent trajectories corresponding to a sample of the posterior draws of $\hat{\Theta}$, $\mathcal{Z}(T)$ can be estimated by their majority vote at each t.

4. Experiments.

4.1. *Experiments on synthetic data.* To evaluate the validity of our proposed algorithms for estimating MMHP, we simulate event arrival times from a generative MMHP model and explore parameter recovery. The simulation of point processes is based on the *thinning algorithm* (Ogata, 1981), a common approach for simulating inhomogeneous Poisson processes.

Estimation of model parameters. Given one fixed set of parameters Θ, we simulate $S = 100$ independent sets of synthetic MMHP processes, each with an independent latent Markov process, and let them run to $M = 50$ events.

Model estimation was carried out using Algorithm 1 with MCMC sampling. We run four parallel chains with random initial values and 1000 iterations per chain, using the first half of each chain for burn-in. For each parameter we obtain $\hat{R}_{\max} < 1.1$ (Gelman et al., 2013). Since the posterior distributions are highly skewed, we calculated simulation-efficient Shortest Probability Intervals (SPIn) (Liu et al., 2015) for the parameter estimates. For our 95% posterior probability interval using SPIn, the coverage rates of true values are all above 95%.

We further examined the sampling distribution of the posterior means from the $S = 100$ independent simulated data sets. Figure 4 displays 100 estimated posterior distributions for each model parameters in grey lines. The purple points and horizontal bars are the average of these posterior mean estimates and their shortest 95% probability interval. The blue lines represent prior distributions and red vertical lines indicate the ground truth. It shows that the sampling distribution of our estimates is centered at the true value with reasonable precision.
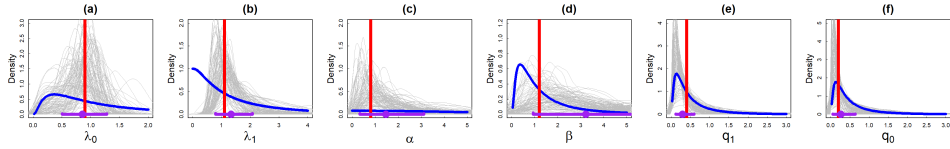

Fig 4: Estimation of the parameters for simulated MMHPs. Red lines: ground truth; grey lines: posterior distributions; purple segments: 95% Shortest Probability Intervals (SPIn) (Liu et al., 2015) based on the sampling distribution; blue lines: prior distributions.

Estimation of the latent Markov process. Using synthetic data, we evaluate how well our proposed algorithm recovers the latent state process. We first simulate one fixed instance of the latent Markov process $\mathcal{Z}(T)$, and then generate $S = 100$ sets of event arrival times given this fixed $\mathcal{Z}(T)$. Numerical experiments are conducted under different lengths of the latent Markov process, where the average numbers of events are $M = 50, 100, 200$ and 500.

Following the inference procedure described in section 3.3, we show the estimates of the latent Markov process in Figure 5-(a), where the thick black line is the ground truth of the latent process, the grey lines are the majority vote estimates for each synthetic processes among $S = 100$ processes, and the thick blue line and the blue shades are the average latent trajectory of $S = 100$ estimates and its one-standard-deviation confidence bands. The estimated trajectories reflect the hidden true state process $\mathcal{Z}(T)$, even at $M = 50$. As the number of events increases, our estimation of the latent process becomes more accurate.

Figure 5-(b) provides a comparison between the true intensity function for one simulated 50-event process and the estimated intensity function that given the posterior draws of the model parameters and the latent process. The accuracy of the estimated intensity serves as another validation of our estimation of the model parameters and the latent state process.

For comparison, we include in Figure 5-(c) the inferred trajectories on

the same simulated data sets using MMPP, which suffers from biases and high variability. The inference of the latent process deviates more from the ground truth when events are more bursty in the state 1. This is because MMPP assumes a constant intensity in each state and is not flexible enough to capture the highly heterogeneous event time data that is generated by the MMHP model. As a result, MMPP systematically underestimates the intensity in the state 1 and overestimates the intensity in the state 0. This causes the estimated mean latent process to regress towards 0.5, with the most sizable deviation taking place during the transition period.
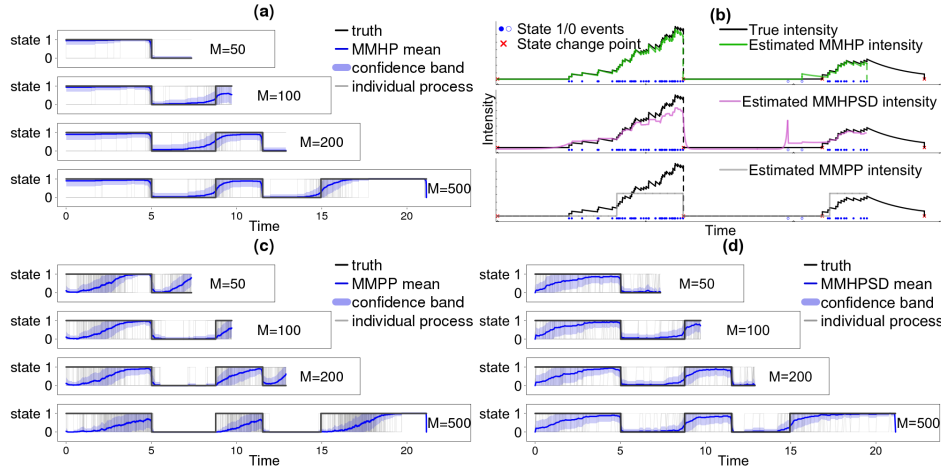

Fig 5: Estimation of the latent Markov process **(a)** Estimated latent trajectory using the proposed algorithm. **(b)** Event intensity for one synthetic process. **(c)** Estimated latent trajectory using MMPP. **(d)** Estimated latent trajectory using MMHPSD.

We also evaluated the performance of MMHPSD (Wang et al., 2012), with results shown in Figure 5-(d). MMHPSD suggests transitions between states that are much more frequent than the ground truth. To quantify this difference between our MMHP model and MMHPSD, we calculate the *integrated absolute error* of the inferred latent process, i.e., $\int_0^T |Z(t) - \hat{Z}(t)|dt$, as shown in Figure 6-(a). Figure 6-(b) shows the comparison between MMHP and MMHPSD, from which we can see that the overly sensitive state-transition of MMHPSD leads to a larger integrated absolute error.

4.2. *Application to email interactions.* We can now fit our proposed model to the directed email pair presented in Figure 1, consisting of emails from one user to another in a large university over one academic semester.
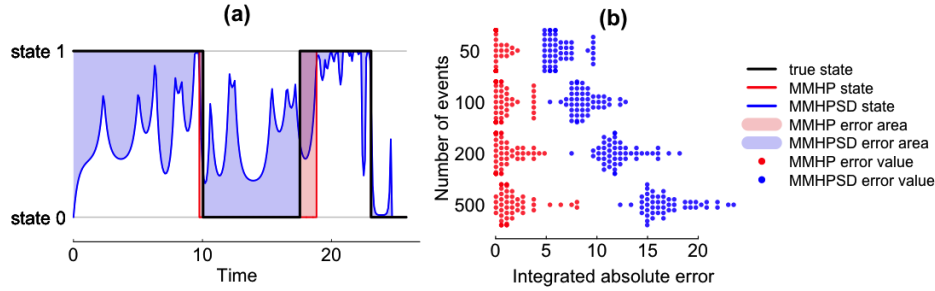
Fig 6: Comparison of integrated absolute error between MMHP and MMH-PSD. **(a)** An illustration of the integrated absolute error of a latent process, i.e., $\int_0^T |Z(t) - \hat{Z}(t)| dt$. Here, the black line indicates the true trajectory, the red and blue lines represents the estimated latent process by MMHP and MMHPSD correspondingly, and the red and blue shaded areas correspond to the absolute error of the inferred states using these two methods. **(b)** Comparison of the integrated absolute error between MMHP and MMH-PSD under different simulation scenarios.

In total, several million emails between more than 40k members of the university were collected. Here we consider the single directed pair shown in Figure 1, which consists of 78 emails over one academic semester of 122 days.

We apply the proposed MMHP model to this email interaction data, using Gamma$(1,1)$ priors for both α and β, with the other priors being the same as those used in Sec 4.3. Figure 7-(a) plots the inferred latent states with observed event times and the estimated state trajectory with the one-standard-error band. MMHP separates bursts of many emails over a period of only a few days from less frequent periods, consisting of at most one or two emails per day. This seems well suited to describing behaviour between individuals in this setting, with several periods of frequent contact throughout the semester interspersed between regular but less frequent communication.

A common approach to test the goodness-of-fit of point process models is the time rescaling theorem (Brown et al., 2002). As it states, the *compensators* $\{\Lambda_m := \int_{t_m}^{t_{m+1}} \lambda(u)du\}_{m=0}^{M-1}$, are expected to be independently distributed following an exponential distribution with rate 1. The MMHP compensators closely follow the exponential distribution with rate 1, as shown in the QQ plot in Figure 7-(b) and Kolmogorov-Smirnov plot in Figure 7-(c). Compared to a homogeneous Poisson process, MMPP, MMHPSD and Hawkes process models, the MMHP model gives the best fit in terms of
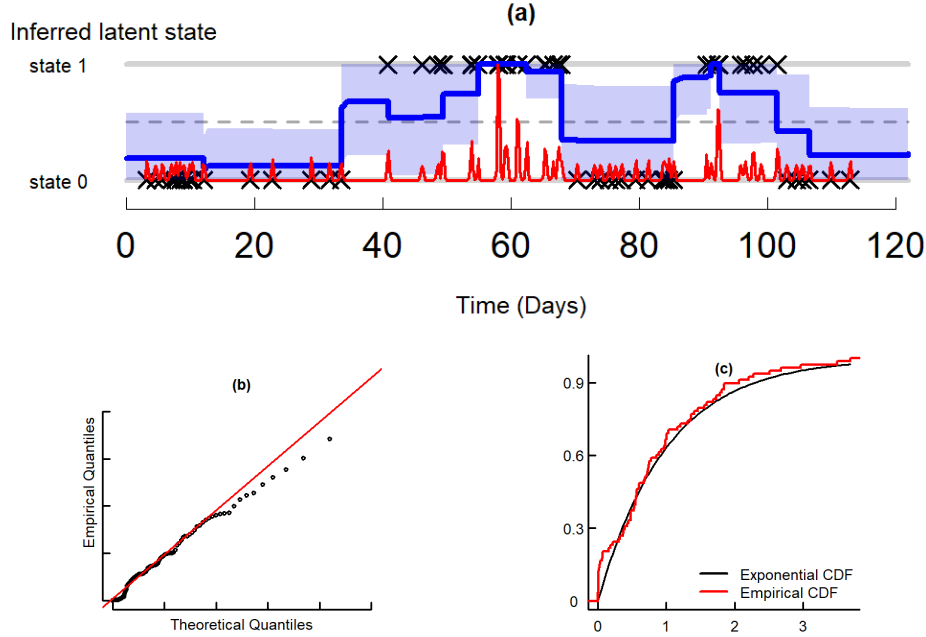
Fig 7: Modeling email interaction data. (a) Estimated latent process for one directed pair of emails from Figure 1, where the crosses are estimated states at observed event times; blue line is inferred $\mathcal{Z}(T)$ based on the majority vote given posterior draws of latent trajectory; the shaded blue area is the one-standard-error band, constrained to be between 0 and 1. KDE estimate overlaid with red line. (b) and (c) show the QQ plot and KS plot for point process compensators of this directed pair.

KS statistic. Further comparison with existing models is given in Supplement S1.2.

4.3. *Experiments with social interaction data among mice.* In Williamson et al. (2016), twelve male mice were placed in a large vivarium at the age of nine weeks. These mice fight each other to establish a social hierarchy. For twenty-one consecutive days, observations were taken during the dark phase of the light cycle when mice are most active. During each observation interval, trained observers recorded the location, time stamp and action of pair-wise mice interactions. For each interaction, the actor (i.e., winner) and recipient (i.e., loser) were recorded. The goal of this study was to understand

how mice collectively establish and navigate their social hierarchy over time and to identify inconsistent deviations from a linear order.

Based on our study of the dataset, we conjecture that the social interaction dynamics of mice exhibit two states: *active* and *inactive*, which can be detected using the proposed MMHP model. After we separate all the interactions into two states, we expect the active state interactions to follow a *linear* hierarchy more closely than with all interactions combined, which suggests an explanation on how social dominance is established among a group of mice. Using multiple measurements of *linearity* for an animal social hierarchy, we show that it is indeed the case. In addition, the inactive state interactions offer insights on social structures among the mice that deviate from the dominance hierarchy. A cluster analysis of the inactive state interactions shows that, as time progresses, the extent of between-cluster interactions decreases, which suggests the social structure may be stabilizing.

From this dataset, we considered the relational event dynamics on a fixed set of actors $V = \{1, 2, ..., N = 12\}$. The data consist of all historic events up to a termination time T: $\mathcal{H}(T) = \{(i_m, j_m, t_m)\}_{m=1}^{M}$. For each directed pair of actors $(i \to j)$, the sequence of interaction event times is denoted by $\mathcal{H}^{(i \to j)}(T) = \{t_m^{(i \to j)}\}_{m=0}^{M^{(i \to j)}}$, where $t_0^{(i \to j)} = 0$, and $M^{(i \to j)}$ is the number of interaction events initiated by i and received by j. For the purpose of this paper, we treat each observation window separately and independently. This allows us to disregard non-observation time.

Assume that, for each directed pair $(i \to j)$, the dyadic events $\mathcal{H}^{(i \to j)}(T)$ follow the MMHP model. The parameters of each process share the same prior distribution and vary across pairs. To improve the estimation of the state trajectories, we assume that

$$\lambda_0^{(i \to j)} \sim \mathrm{N}^+ \left(\frac{1}{\Delta t_{\max}^{(i \to j)}}, 0.1 \right),$$

where $\Delta t_{\max}^{(i \to j)}$ is the maximum over all interevent times for $(i \to j)$. We assume that the latent state transitions should be less frequent than the fight frequencies, and set $q_0 = w_0 \lambda_0, q_1 = w_1 \lambda_1$, where $w_0 \sim \mathrm{Beta}(0.5, 0.5), w_1 \sim \mathrm{Beta}(0.5, 0.5)$. The other parameters share the following priors:

$$\alpha \sim \log \mathrm{N}(\mu_\alpha, \sigma_\alpha), \beta \sim \log \mathrm{N}(\mu_\beta, \sigma_\beta).$$

These log-Normal priors allow for the longer tails present considering all interactions within a cohort, compared to the single email interaction pair.

The fights between one pair of these mice are shown in Figure 8. The mice are observed each day, with some observation periods longer than others. In several of these observation periods no fights occur. The continuous time
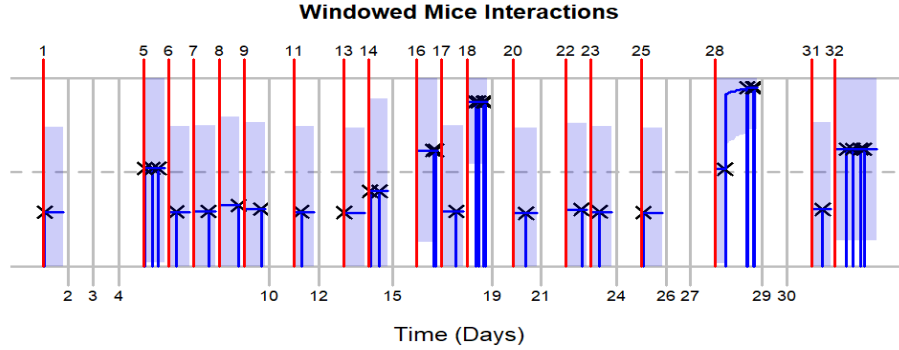
Windowed Mice Interactions

Time (Days)

Fig 8: Interactions between two mice over several weeks. Here the red lines indicate observation windows (and corresponding days) with interactions while grey lines indicate time periods where no interactions were observed. Although we only observe interactions in some of the time windows, we are still able to estimate the latent process in the windows with observed events. This also illustrates the latent state changing within a single time window.

MMHP model is flexible in that it allows the latent state to change during longer observation periods, as seen on day 8 in Figure 8. The latent state is not constrained to remained fixed within a given observation window.

The *active/inactive* state separation by MMHP can also help us understand the inconsistency between mice interaction behaviors and their hierarchy ranks. Given a set of fights among a group of mice, one can calculate the *win/loss* matrix, which is a frequency sociomatrix of wins and losses (So et al., 2015). The (i, j)-th entry in the *win/loss* matrix represents the number of times i won against j. For one cohort, Figure 9 plots the win/loss matrices for all fights and by the *active/inactive* state. There were 929 fights in total in this cohort, of which 805 were classified as belonging to the active state. The order of rows and columns corresponds to the ranks of mice using the *I&SI* method (Vries, 1998). If the interactions strictly follow the social hierarchy, we expect to see all the interactions in the upper triangle with a few exceptions that are close to the diagonal. This is not the case in the overall panel (Figure 9-(a)). The upper triangular structure in Figure 9-(b) suggests that our identified active state of bursty and intensive fights agrees with hierarchical rank. The inactive state interactions are deviations from the hierarchy. This suggests that these interactions might be motivated by the mice's need to explore the social hierarchy without intensive engagements.
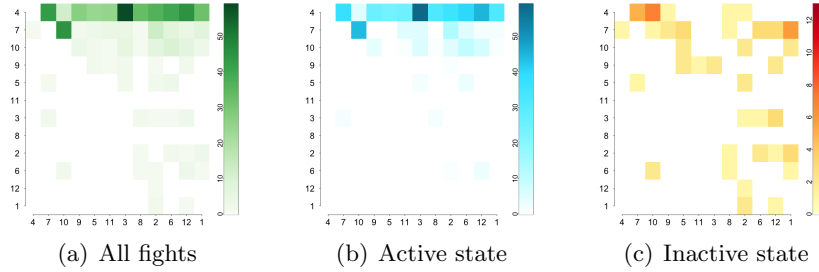
(a) All fights   (b) Active state   (c) Inactive state

Fig 9: Plots of mice win/loss matrices for all fights and by states (sorted by *I&SI* method (Vries, 1998)).

Measurement	All fights	Active state	Inactive state
Directional Consistency	0.96	0.99	0.85
Triangle transitivity	1.00	1.00	0.80
Inconsistencies in ranking	3	1	2

TABLE 1

Measures of social hierarchy linearity in the same cohort

To better quantify how closely a set of interactions follow a linear hierarchy, we calculate the following three measures of social hierarchy linearity for the *win/loss* matrix as in Figure 9: *Directional consistency* (Leiva et al., 2008), *Triangle transitivity* (McDonald and Shizuka, 2012) and *Inconsistency* in the ranking (Vries, 1998). Given a win/loss matrix W, *directional consistency* is defined as $\sum_{i<j} \frac{\max(W_{ij},W_{ij}) - \min(W_{ij},W_{ij})}{W_{ij}+W_{ij}}$, which is the difference in the proportions of fights won by the more dominant individuals and that by the more subordinate individuals. For three individuals (i,j,k), *triangle transitivity* measures the proportion of triad relations satisfying transitivity, i.e. $\mathbb{1}_{\{i>j,j>k,i>k\}}$, where $i>j$ represents i dominates j and $\mathbb{1}_{\{\cdot\}}$ is the indicator function. For $\tilde{W}$, which is W with its rows and columns reordered according to a ranking, *inconsistency* in the ranking equals to $\sum_{i>j} \mathbb{1}_{\{\tilde{W}_{ij}>\tilde{W}_{ji}\}}$. A perfect linear hierarchy ranking would give zero inconsistency. Table 1 shows the results of the above measures corresponding to the win/loss matrix shown in Figure 9.

Figure 10-(a) shows the boxplots of these three measurements for ten cohorts, which are calculated using all, active, and inactive interactions. All three measurements suggest a stronger linear hierarchy among the active-state interactions comparing to all interactions combined.

Williamson et al. (2016) found that day 5 and day 10 were the start and end of social hierarchy stabilization. Figure 10-(b) summarized this
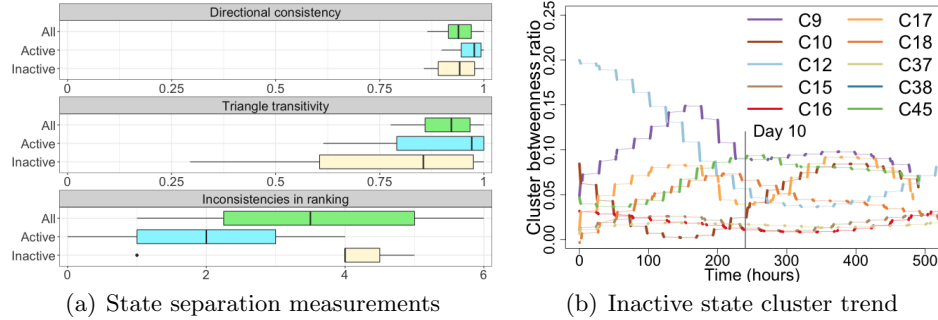
(a) State separation measurements   (b) Inactive state cluster trend

Fig 10: Summary of state separation result in 10 cohorts.

trend for all cohorts: each line in the plot indicates a cohort's evolving ratio of between-cluster interactions and within-cluster interactions. The thicker line segments represent the ratio values calculated for each day by using interactions within a sliding one-hour time window. Due to the fact that the study has different observation times within a day across cohorts, we align all the cohorts by interpolating the calculated segments (thick lines) with a thin line. We see that all cohorts show stable clustering after the 10th day.

5. Discussion and conclusion. Although dynamic social interactions have been widely studied recently, existing models cannot adequately capture the patterns of event dynamics in social interaction data – sporadic with bursts and long wait time. In this paper, we proposed a Markov Modulated Hawkes Process (MMHP) model and its inference algorithms, which segments different dynamic patterns in event arrival times data. Results from numerical experiments provide validating evidence on the advantages of the proposed method over comparable existing methods, in both simulated studies and a real application to both animal behavior data and email data.

In this paper, the MMHP model and its inference framework are built under the assumption that the latent Markov process has two states for better model interpretability. It can readily be extended to a R-state latent Markov process by modifying the computation of the forward-backward algorithm, as we have included in Appendix B. However, it is challenging to make this framework computationally practical for $R > 2$. This also raises identifiability concerns between the underlying latent process and the Hawkes processes. Veen and Schoenberg (2008) pointed out that the log likelihood function of Hawkes processes may be multi-modal and locally flat. This be-

comes more notable when we consider R generally parametrized Hawkes processes, modularized by a Markov process. One potential avenue is to utilise other inference schemes, such as Markov Jump processes Rao and Teh (2013). Initial simulation studies (not shown here) indicate this is a promising direction, however this procedure struggles to infer a stable latent process in our model for $R = 2$. Further work is therefore required to investigate this inference procedure in this setting before considering the general R-state problem.

MMHP uses the classic exponential kernel in Hawkes processes, where α represents the influence of historical events on the intensity and β represents the decay rate of such influence. A potential extension of the proposed model could be introducing covariates into α and β. Subsequent inference can be easily conducted under our framework. Although our current work emphasizes more on the model interpretation due to the motivation from the social interaction data, our model can also make predictions on the expected number of future events with measures of uncertainty.

In this paper, we focused on modeling the sporadic dynamic of one sequence of event history that can be assumed as one non-stationary point process. In practice, some event streams might be driven by multiple dynamic processes. One such example is Du et al. (2015), who fitted a nonparametric Dirichlet mixture of Hawkes processes to a large collection of document streams and clustered these news articles into separate streams, with each stream represented by a Hawkes process. A similar mixture of Hawkes processes is considered by Xu and Zha (2017). This mixture of Hawkes process is able to capture the self excitatory patterns seen in mixtures of document streams, but is not able to capture the long wait times observed in social interaction data that we consider here. Adding an inactive state with sporadic event arrivals to such mixture models via Markov modulation is a future direction worth exploring.

Furthermore, we assumed independence structure when applying the model to a network of animals. However, more research can be carried out in terms of introducing a dependence structure among different animal pairs in a network. Such a network-structured MMHP will lead to improvements in model estimates and interpretability. On the other hand, combined with network models, MMHP provides a way of inferring network structure based on continuous-time event data.

References.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346.

Cohn, I., El-Hay, T., Friedman, N., and Kupferman, R. (2009). Mean field variational approximation for continuous-time bayesian networks. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 91–100. AUAI Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J., and Song, L. (2015). Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 219–228.

Fischer, W. and Meier-Hellstern, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance evaluation*, 18(2):149–171.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Guo, J., Betancourt, M., Brubaker, M., Carpenter, B., Goodrich, B., Hoffman, M., Lee, D., Malecki, M., and Gelman, A. (2014). RStan: The R interface to Stan.

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

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *science*, 311(5757):88–90.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Leiva, D., Solanas, A., and Salafranca, L. (2008). Testing reciprocity in social interactions: A comparison between the directional consistency and skew-symmetry statistics. *Behavior Research Methods*, 40(2):626–634.

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

Liu, Y., Gelman, A., and Zheng, T. (2015). Simulation-efficient shortest probability intervals. *Statistics and Computing*, 25(4):809–819.

McDonald, D. B. and Shizuka, D. (2012). Comparative transitive and temporal orderliness in dominance networks. *Behavioral Ecology*, 24(2):511–520.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rao, V. and Teh, Y. W. (2013). Fast mcmc sampling for markov jump processes and extensions. *J. Mach. Learn. Res.*, 14(1):3295–3320.

Scott, S. L. and Smyth, P. (2003). The markov modulated poisson process and markov poisson cascade with applications to web traffic data. *Bayesian statistics*, 7:671–680.

Simma, A. and Jordan, M. I. (2010). Modeling events with cascades of poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 546–555.

So, N., Franks, B., Lim, S., and Curley, J. P. (2015). A social network approach reveals associations between mouse social dominance and brain gene expression. *PloS one*, 10(7):e0134509.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

Vries, H. d. (1998). Finding a dominance order most consistent with a linear hierarchy: a new procedure and review. *Animal Behaviour*, 55(4):827–843.

Wang, T., Bebbington, M., and Harte, D. (2012). Markov-modulated Hawkes process with stepwise decay. *Annals of the Institute of Statistical Mathematics*, 64(3):521–544.

Wang, Y., Du, N., Trivedi, R., and Song, L. (2016). Coevolutionary latent feature processes for continuous-time user-item interactions. In *Advances in Neural Information Processing Systems*, pages 4547–4555.

Weiss, J., Natarajan, S., and Page, D. (2012). Multiplicative forests for continuous-time processes. In *Advances in neural information processing systems*, pages 458–466.

Williamson, C. M., Lee, W., and Curley, J. P. (2016). Temporal dynamics of social hierarchy formation and maintenance in male mice. *Animal Behaviour*, 115:259–272.

Williamson, C. M., Romeo, R. D., and Curley, J. P. (2017). Dynamic changes in social dominance and mpoa gnrh expression in male mice following social opportunity. *Hormones and Behavior*, 87:80–88.

Xu, H. and Zha, H. (2017). A dirichlet mixture model of hawkes processes for event sequence clustering. In *Advances in Neural Information Processing Systems*, pages 1354–1363.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. ACM.

APPENDIX A: ALGORITHMS

A.1. Posterior Samples. Here we provide the algorithm use to draw samples from the posterior of the proposed MMHP model.

Algorithm 1 Posterior sampling of parameters for MMHP.

Inputs:

   M(number of events), $\Delta t_{1:M}$(interevent time)

Prior:

   $\delta_0 \sim U(0,1), \alpha \sim \mathrm{N}(0,5), \beta \sim \log \mathrm{N}(0,0.5),$

   $\lambda_1 \sim \log \mathrm{N}(0,1), \lambda_0 < \lambda_1, q_0, q_1 \sim \log \mathrm{N}(-1,1).$

Likelihood:

   Initialize: $m \leftarrow 0; \mathcal{A}_m[1] \leftarrow (1 - \delta_0); \mathcal{A}_m[2] \leftarrow \delta_0.$

   while $m < M$ **do**

      $m \leftarrow m + 1$

      $\mathcal{A}_m[1] \leftarrow \mathcal{A}_{m-1}[1] \times \tilde{P}_\Theta(Z_m = 1, \Delta t_m | Z_{m-1} = 1, \Delta t_{1:m-1}) +$
         $\mathcal{A}_{m-1}[2] \times \tilde{P}_\Theta(Z_m = 1, \Delta t_m | Z_{m-1} = 0, \Delta t_{1:m-1})$

      $\mathcal{A}_m[2] \leftarrow \mathcal{A}_{m-1}[1] \times \tilde{P}_\Theta(Z_m = 0, \Delta t_m | Z_{m-1} = 1, \Delta t_{1:m-1}) +$
         $\mathcal{A}_{m-1}[2] \times \tilde{P}_\Theta(Z_m = 0, \Delta t_m | Z_{m-1} = 0, \Delta t_{1:m-1})$

   $P_\Theta(\Delta t_{1:M}) \leftarrow \mathcal{A}_M[1] + \mathcal{A}_M[2]$

Posterior:

   Use MCMC to sample from posterior distribution.

Outputs:

   Posterior draws of Θ.

A.2. The Viterbi algorithm. Algorithm 2 describes the Viterbi algorithm discussed in the main text to infer the latent states of MMHP.

A.3. The forward backward algorithm. The forward-backward algorithm is a dynamic programming algorithm for computing the marginal likelihood of a sequence of observations from complete-data likelihood, by iteratively marginalizing out hidden state variables. As shown in Section 3.2, we obtained the approximated likelihood after marginalizing over $\mathcal{Z}(T) \in \mathcal{S}_{\tilde{\mathcal{Z}}(T)}$, i.e. $\tilde{P}_\Theta(\mathcal{H}(T), \tilde{\mathcal{Z}}(T))$. This is a function of $\tilde{\mathcal{Z}}(T) = \{Z_0, Z_1, ..., Z_M\}$, which are the states at event times. Utilizing the forward variable, we can then marginalize over $\tilde{\mathcal{Z}}(T)$ computationally efficiently. For convenience, we denote the event history up to m-th event $\mathcal{H}(t_m)$ through their interevent times $\Delta t_{1:m} := \{\Delta t_i := t_i - t_{i-1}, i = 1, ..., m\}$. Hence the forward variable can be defined as $\mathcal{A}_m(z_m, \Delta t_{1:m}) := \tilde{P}_\Theta(Z_m = z_m, \Delta t_{1:m})$. It satisfies the initial condition that $\mathcal{A}_0(z_0) = P(Z_0 = z_0)$. The forward iteration can be

---

Algorithm 2 Viterbi algorithm for Markov Modulated Hawkes Process

---

Inputs:
$\hat{\Theta}$(Estimation of parameters), M(number of events), $\Delta t_{1:M}$(interevent time)

Initialize:
$m \leftarrow 0$; $v_m[1] \leftarrow \log{(1 - \hat{\delta}_0)}$; $v_m[0] \leftarrow \log{(\hat{\delta}_0)}$

while $m < M$ **do**

$\qquad m \leftarrow m + 1$

$\qquad v_m[1] \leftarrow \max_k v_{m-1}[k] + \log \tilde{P}_{\hat{\Theta}}(Z_m = 1, \Delta t_m | Z_{m-1} = k, \Delta t_{1:m-1})$

$\qquad v_m[0] \leftarrow \max_k v_{m-1}[k] + \log \tilde{P}_{\hat{\Theta}}(Z_m = 0, \Delta t_m | Z_{m-1} = k, \Delta t_{1:m-1})$

$\qquad b_m[1] \leftarrow \arg\max_k v_{m-1}[k] + \log \tilde{P}_{\hat{\Theta}}(Z_m = 1, \Delta t_m | Z_{m-1} = k, \Delta t_{1:m-1})$

$\qquad b_m[0] \leftarrow \arg\max_k v_{m-1}[k] + \log \tilde{P}_{\hat{\Theta}}(Z_m = 0, \Delta t_m | Z_{m-1} = k, \Delta t_{1:m-1})$

$z_M^* \leftarrow \arg\max_k v_M[k]$

while $m' \leq M$ **do**

$z_{M-m'}^* \leftarrow b_{M-m'+1}[z_{M-m'+1}^*]$

Outputs:
Global optimal sequence of latent state $(z_0^*, z_1^*, ..., z_M^*)$

---

derived as

$$\mathcal{A}_m(z_m, \Delta t_{1:m})$$
$$= \tilde{P}_\Theta(Z_m = z_m, \Delta t_{1:m})$$
$$= \sum_{z_{m-1}} \tilde{P}_\Theta(Z_{m-1} = z_{m-1}, \Delta t_{1:m-1}) \tilde{P}_\Theta(Z_m = z_m, \Delta t_{1:m} | Z_{m-1} = z_{m-1}, \Delta t_{1:m-1})$$
$$= \sum_{z_{m-1}} \mathcal{A}_{m-1}(z_{m-1}, \Delta t_{1:m-1}) \tilde{P}_\Theta(Z_m = z_m, \Delta t_m | Z_{m-1} = z_{m-1}, \Delta t_{1:m-1})$$

Given the last forward variable, we obtain the whole marginalized likelihood as

$$\sum_{\mathcal{Z}(T)} \tilde{P}_\Theta(\mathcal{H}(T), \tilde{\mathcal{Z}}(T)) = \sum_{z_M} \mathcal{A}_M(z_M, \Delta t_{1:M})$$

APPENDIX B: GENERAL R STATE ALGORITHM

For completeness, we describe our MMHP procedure for general R states here. For this, we define CTMC with R states as $Z(t), t \in [0, T], Z(t) \in \{1, 2, ..., R\}$. The initial probability vector is $\delta := \{\delta_1, ..., \delta_R\}$, where $\sum_{r=1}^R \delta_r = 1$. The infinitesimal generator matrix Q is a $R \times K$ matrix defined as

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,R} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ q_{R,1} & q_{R,2} & \cdots & q_{R,R} \end{bmatrix},$$

where $q_{r,r} = -\sum_{j \neq r} q_{r,j}$.

For each $t \geq 0$, there is a probability transition matrix, denoted as $\mathbf{P}(t) := [\mathbf{P}_{ij}(t)]_{i,j \in \{0,1\}}$. Each entry $\mathbf{P}_{ij}(t)$ is defined as the probability that the chain will be in state j at time $u + t$ $(t > 0)$ given the chain is in state i at time u, i.e. for each $u \geq 0$,

$$\mathbf{P}_{ij}(t) = P(Z(u + t) = j | Z(u) = i).$$

Following Kolmogorov's forward equation, we know that $\mathbf{P}(t) = \exp(Qt)$.

Given $Z(t) = r$, define the Hawkes intensity as $\lambda_r(t) = \lambda_{0,r} + \alpha_r \sum_{t_m < t} e^{-\beta_r(t - t_m)}$.

To derive the forward algorithm:

$$\mathcal{A}_m(z_m, \Delta t_{1:m})$$
$$= \tilde{P}_\Theta(Z_m = z_m, \Delta t_{1:m})$$
$$= \sum_{z_{m-1}} \tilde{P}_\Theta(Z_{m-1} = z_{m-1}, \Delta t_{1:m-1}) \tilde{P}_\Theta(Z_m = z_m, \Delta t_{1:m} | Z_{m-1} = z_{m-1}, \Delta t_{1:m-1})$$
$$= \sum_{z_{m-1}} \mathcal{A}_{m-1}(z_{m-1}, \Delta t_{1:m-1}) \tilde{P}_\Theta(Z_m = z_m, \Delta t_m | Z_{m-1} = z_{m-1}, \Delta t_{1:m-1})$$

Here, to be precise,

$$\tilde{P}_\Theta(Z_m = z_m, \Delta t_m | Z_{m-1} = z_{m-1}, \Delta t_{1:m-1})$$
$$= \mathbf{P}_{z_{m-1}, z_m}(\Delta t_m) \lambda_{z_m}(t_m | \mathcal{H}(t_m)) \exp\{-\int_{t_{m-1}}^{t_m} \lambda(u | \mathcal{H}(u)) du\}.$$

We give the full details of the algorithms required for the general R-state model in Algorithm 3 and Algorithm 4.

Algorithm 3 Posterior sampling of parameters for R-state MMHP.

Inputs:
    M(number of events), $\Delta t_{1:M}$(interevent time)
Prior:
    $\delta_0 \sim U(0,1), \alpha \sim \mathrm{N}(0,5), \beta \sim \log \mathrm{N}(0,0.5),$
    $\lambda_r \sim \log \mathrm{N}(0,1), Q \sim \log \mathrm{N}(-1,1).$
Likelihood:
  Initialize: $m \leftarrow 0;\ \mathcal{A}_m[r] \leftarrow \delta_r$, for $r \in \{1,...,R\}$.
  while $m < M$ **do**
      $m \leftarrow m + 1$
      for r=1,...,R **do**
        $\mathcal{A}_m[r] \leftarrow \sum_{j=1}^{R} \mathcal{A}_{m-1}[j] \times \tilde{P}_\Theta(Z_m = r, \Delta t_m | Z_{m-1} = j, \Delta t_{1:m-1})$
  $P_\Theta(\Delta t_{1:M}) \leftarrow \sum_{r=1}^{R} \mathcal{A}_M[r]$
Posterior:
    Use MCMC to sample from posterior distribution.
Outputs:
    Posterior draws of Θ.

Algorithm 4 Viterbi algorithm for Markov Modulated Hawkes Process

---

Inputs:
    $\hat{\Theta}$(Estimation of parameters), M(number of events), $\Delta t_{1:M}$(interevent time)
Initialize:
    $m \leftarrow 0;\ v_m[r] \leftarrow \log(\hat{\delta}_r), r = 1,...,R.$
while $m < M$ **do**
      $m \leftarrow m + 1$
      for r=1,...,R **do**
        $v_m[r] \leftarrow \max_j v_{m-1}[j] + \log \tilde{P}_{\hat{\Theta}}(Z_m = r, \Delta t_m | Z_{m-1} = j, \Delta t_{1:m-1})$
        $b_m[r] \leftarrow \arg\max_j v_{m-1}[j] + \log \tilde{P}_{\hat{\Theta}}(Z_m = r, \Delta t_m | Z_{m-1} = j, \Delta t_{1:m-1})$
$z_M^* \leftarrow \arg\max_r v_M[k]$
while $m' \leq M$ **do**
  $z_{M-m'}^* \leftarrow b_{M-m'+1}[z_{M-m'+1}^*]$
Outputs:
    Global optimal sequence of latent state $(z_0^*, z_1^*, ..., z_M^*)$

SUPPLEMENTARY MATERIAL

S1.1. Discussion of data binning. To further illustrate the difficulties of correctly binning continuous time data, we illustrate kernel density estimates with a range of bandwidths. As can be seen in Figure S1, it is only when the bandwidth of the kernel density is sufficiently small that it begins to capture the sporadic and bursty nature of interaction times. This indicates that a continuous time model is needed to fully capture the dynamics in this data.

S1.2. Further model comparison. Here we illustrate further how existing models fail to capture the latent dynamics present in the data we analysed.

We fit each of the other models from the literature to the email interaction data and illustrate lack of model fit using the methods described in Sec 4.2. Figures S2,S3 and S4 so model evaluation tools for each of the alternative models considered in the paper. As shown by the compensator plots, each of these models are not able to capture the nature of this arrival data. Similarly, performing a Kolmogorov-Smirnov test on the rescaled inter-event times under each model indicates departures from independent identically distributed exponential random variables.
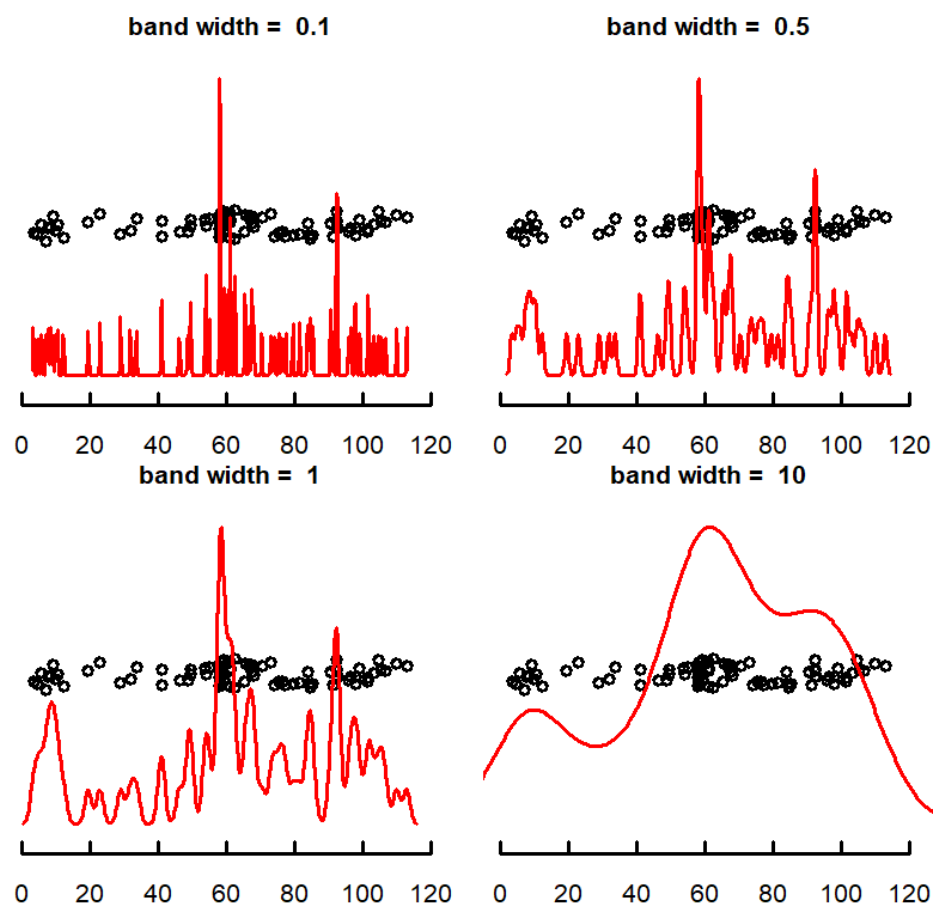
Fig S1: Plots of the jittered event times for the email interaction data along with corresponding kernel density estimates for a range of decreasing bandwidths.
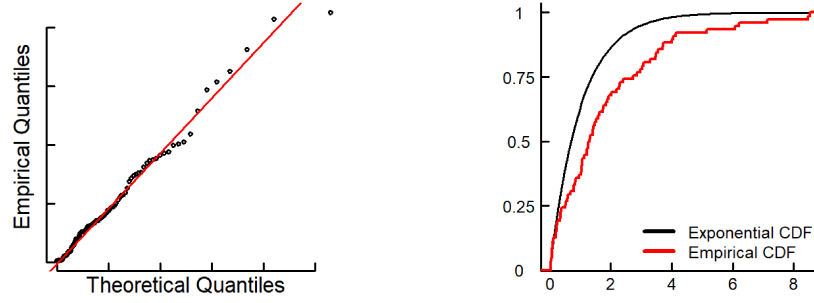
Fig S2: Compensator plots for Hawkes. These illustrate how this model is not able to capture the latent dynamics present in the data.
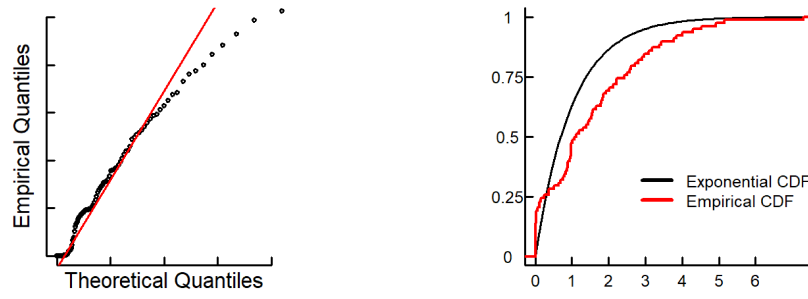

Fig S3: Compensator plots for MMPP. These illustrate how this model is not able to capture the latent dynamics present in the data.
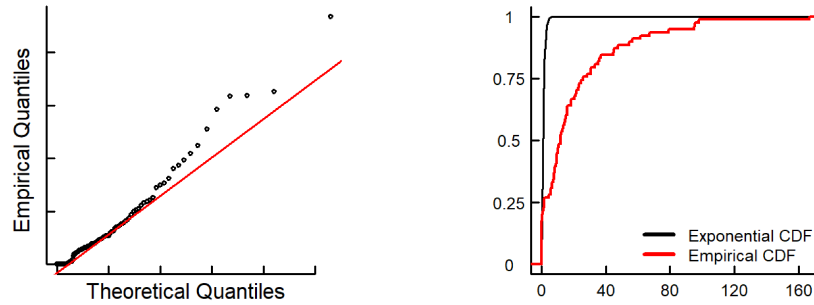

Fig S4: Compensator plots for MMHPSD. These illustrate how this model is not able to capture the latent dynamics present in the data.