Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

# Parameter Estimation of Binned Hawkes Processes

Leigh Shlomovich, Edward A. K. Cohen, Niall Adams & Lekha Patel

**To cite this article:** Leigh Shlomovich, Edward A. K. Cohen, Niall Adams & Lekha Patel (2022): Parameter Estimation of Binned Hawkes Processes, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2022.2050247

**To link to this article:** https://doi.org/10.1080/10618600.2022.2050247

# Parameter Estimation of Binned Hawkes Processes

**Leigh Shlomovich,***

Edward A. K. Cohen,

Niall Adams,

Department of Mathematics, Imperial College London, London, U.K.,

and

Lekha Patel

Department of Mathematics, Imperial College London, London, U.K.

and

Statistical Sciences, Sandia National Laboratories, Albuquerque, USA.

*Leigh Shlomovich is funded by an Industrial Strategy Engineering and Physical Sciences Research Council Scholarship.

leigh.shlomovich14@imperial.ac.uk



***Abstract***

**A key difficulty that arises from real event data is imprecision in the recording of event time-stamps. In many cases, retaining event times with a high precision is expensive due to the sheer volume of activity. Combined with practical limits on the accuracy of measurements, binned data is common. In order to use point processes to model such event data, tools for handling parameter estimation are essential. Here we consider parameter estimation of the Hawkes process, a type of self-exciting point process that has found application in the modeling of financial stock markets, earthquakes and social media cascades. We develop a novel optimization approach to parameter estimation of binned Hawkes processes using a modified Expectation-Maximization algorithm, referred to as Binned Hawkes Expectation Maximization (BH-EM). Through a detailed simulation study, we demonstrate that existing methods are capable of producing severely biased and highly variable parameter estimates and that our novel BH-EM method significantly outperforms them in all studied circumstances. We further illustrate the performance on network flow (NetFlow) data between devices in a real large-scale computer network, to characterize triggering behavior. These results highlight the importance of correct handling of binned data.**

*Keywords:* Hawkes processes, self-exciting processes, aggregated data, binned data, EM algorithm

## 1 Introduction

Point processes on the real line are extensively used to model event data and have found wide applications in many fields including seismology Ogata (1999) and cyber-security Price-Williams and Heard (2020). Let $N(A)$ be a random integer that denotes the number of events in set $A \subset \mathbb{R}$, one representation of a point process is via the counting process $\{N(t), t \in \mathbb{R}\}$, where $N(t) = N((0,t])$ for $t > 0, -N((t,0])$ for $t < 0$ and $N(0) = 0$ Daley and Vere-Jones (2003). Due to limited recording capabilities and storage capacities, retaining event times with a high precision is expensive and often infeasible. Therefore, in much real-world data, it is common to instead observe a times series of the binned process. Here we use 'binned' to mean an aggregation of the latent continuous time process into a series of counts per interval of time. That is,

$$N_t = N\big(\Delta(t+1)\big) - N\big(t\Delta\big),$$

for some interval length, $\Delta > 0$ which we refer to as the *bin width* and $t \in \mathbb{N}$ is now a discrete index. Note that in this paper we will use $N(t)$ to denote a continuous time process and $N_t$ for a discrete process where $t \in \mathbb{N}$, and we use ' binning' synonymously with 'aggregating'. Here we assume observations of the infinite length process are made on a finite window $(0,T]$. This binned process may arise from a predetermined aggregation of the data into counts, or equivalently from the rounding of event times. In the context of network traffic data, for example, the resolution of the recorded times can be anywhere from milliseconds to seconds (as is the case with the Los Alamos National Laboratory (LANL) NetFlow data Turcotte et al. (2018)), or even coarser. In this setting the value of this binned process at each time point is the number of events with that rounded time-stamp. When analysing the data, we cannot retroactively reduce

the aggregation levels chosen, and therefore to apply continuous-time models to count data we require some consideration of the effect of this binning.

Intuitively, when binning data we lose information and essentially ‘blur’ our view of the continuous time point process, making it potentially problematic to apply methods which assume a continuous time framework. Thus, the problem we consider here is to infer upon the underlying continuous process, which often has interpretable parameters, from the observed aggregated data.

The Hawkes process is a type of ‘self-exciting’ process which provides us with a model for contagious event data. Their flexibility and real-world relevancy has resulted in a host of applications. In the case of financial data for example, this allows propagation of stock crashes and surges to be modeled Bacry et al. (2012); Bowsher (2007); Filimonov and Sornette (2012); Fonseca and Zaatour (2014); Embrechts et al. (2011) and insurance claim times estimated Chen and Hall (2016). Propagation of social media events has also been modeled using Hawkes processes, in particular ‘twitter cascades’ are considered in Rizoiu et al. (2017); Kobayashi and Lambiotte (2016). Further applications include the modeling of civilian deaths due to insurgent activity in Iraq Lewis and Mohler (2011), and predicting origin times and magnitudes of earthquakes Ogata (1988); Chen and Stindl (2018). There is also potential for Hawkes processes to be used in the modeling of COVID-19 infections Flaxman et al. (2020).

Formally, the Hawkes process is a class of stochastic process with the property that

$$\begin{aligned} \Pr\{\mathrm{d}N(t) = 1 \mid N(s)\,(s \le t)\} &= \lambda^*(t)\mathrm{d}t + o(\mathrm{d}t), \\ \Pr\{\mathrm{d}N(t) > 1 \mid N(s)\,(s \le t)\} &= o(\mathrm{d}t), \end{aligned} \tag{1}$$

where $\mathrm{d}N(t) = N(t+\mathrm{d}t) - N(t)$. It is characterized via its conditional intensity function (CIF) $\lambda^*(t)$ which, combined with (1), defines the Hawkes process. In particular, the CIF of a Hawkes proces is given by

$$\lambda^*(t) = \nu + \int_{-\infty}^{t} g(t-u)\mathrm{d}N(u), t \in \mathbb{R},$$

where $\nu > 0$ is called the background intensity and $g(\cdot)$ is the nonnegative excitation kernel such that $g(u) = 0$ for $u < 0$. This means the intensity at an arbitrary time-point is dependent on the history of the process, producing self-exciting behavior. Depending on the kernel $g(\cdot)$, the excitation may be quite local, or have longer term effects Hawkes (1971).

We can also consider a Hawkes process as a branching process of time-stamped events. From this viewpoint, formalised in Hawkes and Oakes (1974), events can be seen to arrive either via *immigration* or *birth*. That is, an event can be triggered by the background intensity rate $\nu$, in which case the event is seen as an immigrant. Alternatively, an event which is cause by self-excitation can be considered a descendant, referred to as being generated 'endogenously'. Unlike a homogeneous Poisson point process, where events happen independently and at a constant rate, self-exciting processes are such that each event can have an effect on the likelihood of further events happening in the future, provided the excitation kernel is non-zero Rizoiu et al. (2017). Note that in the case that $g(u) = 0$ everywhere, the Hawkes process reduces to a homogenous Poisson process. The expected number descendants from each event is given by the branching ratio of a Hawkes process, defined as

$$0 < \gamma := \int_0^{\infty} g(u)\mathrm{d}u < 1. \qquad (2)$$

The inequality above ensures that the process is stationary and results from the form of the CIF.

As introduced in Hawkes' original paper, exponential decay is a common choice for the excitation kernel due to the simplifications it provides for the theoretical derivations Hawkes (1971); Laub et al. (2015). In this case we can write the excitation kernel as a sum of $L$ exponential decays,

$$g(u) = \begin{cases} \sum_{l=1}^{L} \alpha_l \exp(-\beta_l u), & u > 0, \alpha_l > 0 \\ 0, & \text{otherwise.} \end{cases} \qquad (3)$$

Here, and as is most common, we let $L = 1$ when considering the exponential kernel. In this case, the branching ratio defined in (2) becomes

$$\gamma = \int_0^\infty \alpha \exp(-\beta u) \mathrm{d}u = \frac{\alpha}{\beta}.$$

Given this model, we wish to estimate the parameter set $\Theta = \{\nu, \alpha, \beta\}$. Other kernels can be used, including a power-law function of form $g(u) = \alpha\beta(1+\beta u)^{-(1+c)} \ 1_{\mathbb{R}_+}(u)$, in which case $\Theta = \{\nu, \alpha, \beta, c\}$. In the continuous time setting, parameter estimation for any of these kernels is straightforward.

#### 1.1 Estimation from Continuous Data

Typically, maximum likelihood estimation (MLE) is used to estimate parameters of a point process from a set of exact event times $\mathcal{T} = \{t_1, \ldots, t_{N(T)}\} \subset (0, T]$. As shown in Proposition 7.2.III of Daley and Vere-Jones (2003), the log-likelihood is given by

$$\log \mathcal{L}(\Theta; \mathcal{T}) = \sum_{j=1}^{N_T} \log \lambda^*(t_j) - \int_0^T \lambda^*(u) \mathrm{d}u. \qquad (4)$$

If specifically considering a Hawkes process with exponential excitation kernel of form $\alpha \exp(-\beta t),$ this log-likelihood can be simplified and expressed recursively as shown in Laub et al. (2015). In this paper we consider methods that are required when we instead observe a binned sequence of event counts. An

alternative but equivalent representation of this is a discretization or rounding of the latent time-stamps, but here we will consider the observed data as the aggregation of $\mathcal{T}$ to bins.

#### 1.2 Estimation from Binned Data

In the literature, the issue of binned data is handled in many ways, from uniformly redistributing events across the bin Bowsher (2007), to only retaining unique time-stamps and discarding the rest Lorenzen (2012). Here we propose a novel method referred to as *Binned Hawkes Expectation Maximization* (BH-EM), an approach which is related to the MC-EM algorithm. We additionally compare it to two existing approaches, evaluating the performance of parameter estimation for each. The methods compared are:

1. Approximating a binned Hawkes process as an integer-valued auto-regressive (INAR) process, a method developed in Kirchner (2016, 2017) and described in Section 1.3.
2. Formulating a binned log-likelihood which assumes a piecewise constant CIF within each interval as described in Section 1.4.
3. A novel BH-EM approach detailed in Section 2.

There are other methods which have been covered in the literature, but are not considered here due to lack of applicability to this problem. As an example, a significant amount of the literature which aims to work with aggregated data considers binning the time-points such that the process contains at most one event per bin as in Obral (2016). This is inappropriate here as we do not have access to the latent event times and so cannot select an appropriate discretization level, Δ. Likewise, in Brillinger (1988) the binned behavior was modeled as a Bernoulli process. Again, this is invalid here as it fails to account for the number of events in a bin and thus will heavily bias results. There exist methods that handle missing data when we observe continuous time-points with gaps in the recording windows Le (2018). That is, when the data considered

contains precise but intermittent recordings. This is a closely related issue, however differs in the fact that when handling binned data, we do not have any precise times to work with.

We now outline the two methods against which we will compare our novel BH-EM approach.

#### 1.3 Hawkes INAR(*p*) Approximation

It is shown in Kirchner (2016, 2017) that the distribution of the bin-count sequence of Hawkes processes can be approximated by an integer-valued autoregressive model, known as the INAR(*p*) model, further details of which can be found in Kirchner (2016). By representing the binned Hawkes counting process as an INAR(*p*) process, a non-parametric estimator for kernel $g(u)$ is then formulated in terms of conditional least squares (CLS).

Let $\Delta > 0$ be the bin width, the univariate Hawkes process bin-count sequence is denoted $\mathbf{N} = [N_1, \ldots, N_K]$, for $K = \lfloor T/\Delta \rfloor$ and $N_i$ denotes the counts in the $i\text{th}$ bin. Then, defining some support $\Delta < s < T$, the CLS-operator is used on the bin counts $\mathbf{N}$, with maximal lag $p = \lceil s/\Delta \rceil$. Thus,

$$\hat{\mathbf{g}}^{(\Delta,s)} := \frac{1}{\Delta}\mathbf{Y}\mathbf{Z}^{\top}\left(\mathbf{Z}\mathbf{Z}^{\top}\right)^{-1},$$

where the design matrix $\mathbf{Z}$ is given as

$$\mathbf{Z} = \begin{bmatrix} N_p & N_{p+1} & \ldots & N_{k-1} \\ N_{p-1} & N_p & \ldots & N_{k-2} \\ \ldots & \ldots & \ldots & \ldots \\ N_1 & N_2 & \ldots & N_{k-p} \\ 1 & 1 & \ldots & 1 \end{bmatrix},$$

and $\mathbf{Y}$ is the lagged bin-count sequence, being $[N_{p+1}, \ldots, N_K]$. Then the entries of $\hat{\mathbf{g}}^{(\Delta,s)}$,

$$\hat{\mathbf{g}}^{(\Delta,s)} := \left(\hat{g}_1^{(\Delta,s)}, \ldots, \hat{g}_p^{(\Delta,s)}, \hat{\nu}^{(\Delta,s)}\right),$$

are estimates for the excitation kernel at the corresponding time-points. Kernel parameters $\alpha$ and $\beta$ are then estimated by fitting an exponential function to these points.

Simulation studies examining the effect of bin width Δ and parameter $s$ are presented in Kirchner (2016); Kirchner and Bercher (2018), where they determine Δ to have the greatest bearing on the quality of the estimates. There are however several points to note with this method. First, CLS requires the inversion of $\mathbf{Z}\mathbf{Z}^\top$. In this case, this matrix contains the event counts per bin, and so it is possible to have cases where this matrix is singular, specifically when the counting process contains $K - p$ consecutive zeros. Second, as it is currently presented, this method does not constrain the parameter estimates to be those of a stationary Hawkes process, that is such that the estimates satisfy Equation (2). Therefore it is possible to yield infeasible estimates. Infeasible estimates can suggest a poor choice of model for the data, however depending on the use case it can be preferable to constrain the parameter estimates and separately consider applicability of the model.

It is also important to note that the methodology outlined in Kirchner (2016) is intended for choosing parametric models given continuous-time point process data, and not specifically as a method for binned Hawkes process parameter estimation. The method is still of interest as a comparator, being one of the only known methods available that can model binned data.

#### 1.4 Binned Likelihood

An alternative method, briefly mentioned in Mark et al. (2019) and developed here, considers sampling $\lambda^*$ at each discrete time-point $j\Delta$ ($j = 1, \ldots, K$ with $K = \lfloor T/\Delta \rfloor$), thus representing $\lambda^*$ as a piecewise constant function within each bin. Letting $N_j$ be the number of events occurring in the sampling interval

$((j-1)\Delta, j\Delta]$ and using (4) we have that the log-likelihood of the data under this model is

$$\log \mathcal{L}(\Theta; \mathbf{N}) = \sum_{j=1}^{K} N_j \log\left(\Delta\lambda^*(j\Delta)\right) - \Delta\lambda^*(j\Delta), \qquad (5)$$

where $\lambda^*(j\Delta) \equiv \lambda^*(j\Delta \mid \mathcal{H}_{j\Delta})$ and $\mathcal{H}_{j\Delta}$ denotes the history of the process until time $j\Delta$.

The assumption of a piecewise constant CIF is equivalent to assuming $N_j \sim$ Poisson$\left(\Delta\lambda^*\left((j-1)\Delta\right)\right)$. However, it is important to note that this assumption does not comply with the Hawkes process statistics as it ignores the excitation within each bin and therefore will be biased, especially in cases where the intensity is high relative to the bin width, Δ. Nevertheless it provides us with a simple approximation. To estimate the process parameters we maximize (5) with constraints ensuring stationarity, as expressed by (2). In the case of an exponential excitation kernel, explicitly this implies $\nu, \alpha > 0$ and $\alpha/\beta < 1$.

We will now propose an alternative method of parameter estimation which iteratively uses 'consistent' sets of continuous candidate time-points and therefore does not assume a piecewise constant CIF.

**2 BH-EM Algorithm for Binned Data**

The EM algorithm Dempster et al. (1977) is an iterative method for the computation of the maximizer of a likelihood. The idea of this algorithm is to augment the observed data by a latent quantity Wei and Tanner (1990). In the case considered here, the observed data are the event counts per unit time. We denote this by $\mathbf{N} = [N_1, \ldots, N_K]$, where $N_j$ denotes the counts in the $j$th bin ( $j = 1, \ldots, K$ ). The latent data, denoted $\mathcal{T}$ are the unobserved, true event times which are rounded on recording and the set of parameters to be estimated is denoted $\Theta = \{\nu, \alpha, \beta\}$. The algorithm proceeds as follows:

1. In the E (Expectation) step, we evaluate

$$Q_{i+1}(\Theta,\Theta^i) = \int_{\mathbb{T}^*} \log(p(\Theta \mid \mathbf{N},\mathcal{T})\, p(\mathcal{T} \mid \mathbf{N},\Theta^i)\mathrm{d}\mathcal{T}, \qquad (6)$$

   where $\mathbb{T}^*$ denotes the sample space for $\mathcal{T}$. That is, we compute the expectation $Q_{i+1}(\Theta,\Theta^i)$ of the log-posterior $\log(p(\Theta \mid \mathbf{N},\mathcal{T}))$ with respect to the conditional predictive distribution $p(\mathcal{T} \mid \mathbf{N},\Theta^i)$, where $\Theta^i$ is the current, $i$th approximation. Note that in the context presented here, $p(\mathcal{T} \mid \mathbf{N},\Theta^i)$ is the probability density function (PDF) of latent, continuous Hawkes times, conditional on the binned count process and current estimated parameters of the process.

2. In the M (Maximization) step, we update the value of the parameter vector with $\Theta^{i+1}$, being the value which maximises the conditional expectation.

When (6) is analytically intractable we require Monte Carlo methods for numerical computation. This is known as MC-EM Wei and Tanner (1990). If we are able to sample $\mathcal{T}$ directly from $p(\mathcal{T} \mid \mathbf{N},\Theta^i)$, then we can approximate the integral in (6) with

$$\frac{1}{r}\sum_{k=1}^{r} \log(p(\Theta \mid \mathbf{N},\mathcal{T}^{*(k)})),$$

where $\mathcal{T}^{*(k)}$ is the $k$th Monte Carlo sample of $\mathcal{T}$. However, no such sampling regime is possible in the Hawkes process setting. We therefore use importance sampling to simulate a *consistent* proposal for $\mathcal{T}$ (that is, a set of event times that match the binned counts) from an alternative distribution $q(\mathcal{T} \mid \mathbf{N},\Theta^i)$ (see Section 2.1 for details). Each of these proposals is then weighted depending on the probability it came from the desired distribution. That is, given a set of *r* simulated proposals $\mathcal{T}^{*(1)},\ldots,\mathcal{T}^{*(r)}$, we assign weights

$$w_k = \frac{p(\mathcal{T}^{*(k)} \mid \mathbf{N}, \Theta^i)}{q(\mathcal{T}^{*(k)} \mid \mathbf{N}, \Theta^i)}, \qquad (7)$$

and approximate (6) with

$$Q_{i+1}(\Theta, \Theta^i) = \frac{\sum_{k=1}^{r} w_k \log(p(\Theta \mid \mathbf{N}, \mathcal{T}^{*(k)}))}{\sum_{k=1}^{r} w_k}. \qquad (8)$$

We note that the numerator of (7) can be expressed as

$$p\left(\mathcal{T}^{*(k)} \mid \mathbf{N}, \Theta^i\right) \propto p\left(\mathcal{T}^{*(k)} \mid \Theta^i\right) \prod_{j=1}^{K} \delta_{N_j, N_j^*(k)},$$

where we let $N^{*(k)}$ denote the binned count process derived from the *k*th simulated proposals, $\mathcal{T}^{*(k)}$ and the known bin width Δ. Further, $\delta_{a,b}$ is the Kronecker delta, equal to 1 if $a = b$ and 0 otherwise. In this way the product term is 1 if the proposed time stamps are consistent. Therefore, if only proposing consistent event times, we have that

$$p\left(\mathcal{T}^{*(k)} \mid \mathbf{N}, \Theta^i\right) \propto p\left(\mathcal{T}^{*(k)} \mid \Theta^i\right),$$

where, $\log\left(p(\mathcal{T}^{*(k)} \mid \Theta)\right)$ is given by (4).

However, the question remains of how to best sample the latent times with consistent bin counts. Ideally, the distribution we would like to sample from is that of the missing event times given the bin counts and the model parameters $p(\mathcal{T} \mid \mathbf{N}, \Theta^i)$. For this method to be most efficient and to ensure meaningful weights, the alternative distribution $q(\mathcal{T} \mid \mathbf{N}, \Theta^i)$ should be as close as possible to the true distribution of the time-stamps.

**2.1 Method for Simulating Proposals**

It is possible to uniformly redistribute the events across a bin in order to generate a consistent set of event times and this could be incorporated into an MC-EM framework by sampling points within each bin from a uniform distribution. However, especially for Hawkes processes with high activity, this is not ideal as it leads to weights that are too small to compute (8). This is due to the uniform distribution not being close enough to $p(\mathcal{T} \mid \mathbf{N}, \Theta)$. We will show that we can derive an alternative distribution which is closer, in that it captures the inter-bin excitation, however we cannot sample from it. Therefore we propose an alternative method which uses a modified construction of the MC-EM algorithm, referred to as Binned Hawkes Expectation Maximization (BH-EM). As opposed to the stochastic Monte Carlo approximation, we create a theoretically deterministic approximation of the expectation given in 6 by sequentially finding modes of the alternative distribution and using these as the 'samples.' In particular, these ' samples' are obtained by maximising the joint truncated density over each bin, using the known boundaries. Maximizing the joint density yields consistent times for each bin, and repeating this process sequentially for all observed bins generates a continuous time version of the binned Hawkes process. In theory, if the joint density for all bins has a unique maximum then each of the $r$ realisations generated will be identical. In practice, and particularly when handling cases with more than one event per bin, there is not necessarily a unique maximum and randomness is introduced by initializing the numerical maximizer at uniformly distributed time-points within each bin.

In this way, the BH-EM algorithm allows us to more easily utilise a distribution closer to the true distribution $p(\mathcal{T} \mid \mathbf{N}, \Theta)$, as the constraint of having to directly sample from it is removed. In the case of Hawkes processes this is important as it is particularly difficult to sample consistent proposals whilst capturing the inter-bin excitation effect in a tractable way. Note that to assess 'closeness', we consider the empirical variance of the weights $w_k$ from (7), resulting from the uniform sampling method and the sequential method outlined in this section. Across all of the simulations explored in Section 3 we find that the variance is

consistently lower, often by several orders of magnitude, for the proposed method. Furthermore the value of $r$ can also be lower than would be the case in a traditional MC-EM setting. This aids with computational efficiency.

We now outline the proposed approach. Without loss of generality, consider having already simulated until time-point $t_n, n \in \{1,\ldots,N(T)\}$. Then, suppose we know that there are exactly $m$ time-points in the next non-empty bin, being $t_{n+1},\ldots,t_{n+m}$. The joint density of these events can be expressed using factorization. That is

$$f^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1},\ldots,t_{n+m}) = \prod_{i=1}^{m} f^*_{T_{n+i}}(t_{n+i}),$$
$$= \prod_{i=1}^{m} \lambda^*(t_{n+i}) \exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \lambda^*(u)\mathrm{d}u\right).$$

For brevity, we refer to $f^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1},\ldots,t_{n+m})$ as $f^*_{T_{n+1:n+m}}(t_{n+1:n+m})$. Note that for the simplest case of an exponential kernel, we can express this as

$$\prod_{i=1}^{m}\left(\nu + \sum_{j=1}^{n+i-1} \alpha \exp(-\beta(t_{n+i} - t_j))\right) \exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \nu + \sum_{j=1}^{n+i-1} \alpha \exp(-\beta(u - t_j))\mathrm{d}u\right),$$
$$= \prod_{i=1}^{m}\left(\nu + \alpha A(n+i)\right) \exp(-\nu(t_{n+m} - t_n)) \exp\left(\sum_{i=1}^{m}\sum_{j=1}^{n+i-1} \frac{\alpha}{\beta}\left(e^{-\beta(t_{n+i} - t_j)} - e^{-\beta(t_{n+i-1} - t_j)}\right)\right),$$

where

$$A(n+i) = \sum_{j=1}^{n+i-1} \exp(-\beta(t_{n+i} - t_j)).$$

As we wish to simulate possible realizations of the events given observed counts, we should account for the the fact that each time-point is known to have occurred within a given interval. That is, we account for the observed interval range $[b^-, b^+]$ for events in a given bin by considering the truncated joint density. For this we consider the CDF of the event times. Specifically, we require that

$b^- < t_{n+1} < t_{n+2} < \ldots < t_{n+m} < b^+$ . Therefore, we can express the conditional CDF over this region as

$$\int_{b^-}^{t_{n+2}} \cdots \int_{t_{n+m-1}}^{b^+} f^*_{T_{n+1:n+m}}(t_{n+1:n+m}) \mathrm{d}t_{n+m} \ldots \mathrm{d}t_{n+1}.$$

Even in the simplest case of an exponential decay kernel, this appears intractable due to the form of the conditional intensity function for a Hawkes process. Therefore we truncate the PDF by considering the joint CDF. As with the joint PDF, we can use factorization to express the joint CDF as

$$F^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1},\ldots,t_{n+m}) = \prod_{i=1}^{m} F^*_{T_{n+i}}(t_{n+i}),$$
$$:= F_{T_{n+m}}(t_{n+m} \mid \mathcal{H}_{t_{n+m}}) \ldots F^*_{T_{n+1}}(t_{n+1} \mid \mathcal{H}_{t_{n+1}}),$$

where we can use the known form for the CDF of each successive time-point given the history of the process. That is, the joint CDF of time-points $t_{n+1},\ldots,t_{n+m}$ is given by

$$\prod_{i=1}^{m} F^*_{T_{n+i}}(t_{n+i}) = \prod_{i=1}^{m} \left(1 - \exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \lambda^*(u)\mathrm{d}u\right)\right).$$

In the case of an exponential decay kernel, this is

$$\prod_{i=1}^{m} \left(1 - \exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \nu + \sum_{j=1}^{n+i-1} \alpha \cdot \exp(-\beta(u - t_j))\mathrm{d}u\right)\right).$$

Thus the joint truncated PDF of $m$ time-points given the history, $t_1,\ldots t_n$ can be expressed as

$$\frac{f^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1},\ldots,t_{n+m})}{\kappa}, \qquad (9)$$

where

$$\kappa = F^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1}, t_{n+2}, \ldots, t_{n+m-1}, b^+) - F^*_{T_{n+1},\ldots,T_{n+m}}(b^-, t_{n+2} \ldots, t_{n+m-1}, t_{n+m}).$$

Note that the inverse of the joint truncated density given in Equation (9) is intractable for $m > 1$ and therefore we assign the set of $m$ proposed time-stamps for the given bin $\{\tilde{t}_{n+1}, \ldots, \tilde{t}_{n+m}\}$ to be those that maximize (9). In this way we can sequentially simulate a continuous version of the observed binned Hawkes process by progressively handling each bin such that we jointly maximize this likelihood. The proposed BH-EM full algorithm is given in Appendix C. Here the sequential simulation method is developed for the exponential kernel but it is easily extendible to other kernels (see Appendices A and B for details regarding power-law and rectangular kernels, respectively).

**3 Simulation Studies**

Given parameters $\nu, \alpha, \beta$ and some maximum simulation time $T$, we can simulate realizations of a Hawkes process. The generated events are those which form the latent space $\mathcal{T}$ , and aggregating these to different Δ allows us to simulate the count data $\mathbf{N}$. We can then apply each of the three methods detailed: the binned log-likelihood, INAR($p$) approximation, and the BH-EM method. We note that for the INAR($p$) method, we have used the AIC minimizing approach detailed in Kirchner (2017) to select the choice of support for each simulation. In cases where the optimal $p$ has been found to be 1, we use the value of $p$ that generates the next smallest value of the AIC, typically found to be $p = 2$ or 3. This is so that estimates of the parameters under an exponential model may be obtained for comparison with the other methods. Figures 1-7 show boxplots for the estimates of each of $\nu, \alpha,$ and $\beta$ for 20 realizations of a Hawkes process with the ground truth parameters specified. The mean value of each parameter estimate is presented on the vertical axis. We clearly see that the INAR($p$) and binned log likelihood methods can yield variable results with the INAR($p$) approximation also yielding negative estimates. In Figures 1, 5 and 7, the boxplots for both the excitation rate, $\alpha$ have been presented on a log-scale in order to show the results on one axis. In these instances the INAR($p$) method has resulted in outliers that

are factors of 10 away from the ground truth. In particular, Figure 7 clearly highlights the effect a larger bin width can have on the INAR($p$) approximation. The remaining figures all likewise show the BH-EM method to perform better than either of the two alternative approaches considered for a range of different parameter sets. The binned log-likelihood method, also appears biased and performs less well than the BH-EM method in all cases. We note that when implementing the INAR($p$) approximation, ideally $\Delta$ is selected such that there is ' about one event on an average per bin' Kirchner (2016). Our application involves handling cases where no such choice is possible, and so this approach expectedly yields highly variable results if $\Delta$ is large relative to the effective support of the excitation kernel.

In Figure 8, we show an example for a power law kernel using the BH-EM method and the INAR($p$) approximation. We see that again the BH-EM method has less variable and less biased results. In Figure 9 we also consider the bias across different levels of aggregation. That is, for each of the realizations of a self-exciting process for a given parameter set, we can bin the data to different levels and compare the bias in the parameter estimates. The right hand plot in Figure 9 presents the bias on a log-scale. It is evident that the INAR($p$) method is more biased for larger bin widths. The binned log-likelihood performs better, however, still not as well as the BH-EM method which most consistently exhibits a low bias.

## 4 Case Study

NetFlow is a protocol operating on routers that assembles records of communications between devices in an enterprise computer network. Originally designed for accounting for network usage, it has potential applications in detecting a variety of malicious network activities Turcotte et al. (2018). Specifically, here it is of interest to detect self-exciting behavior within the communication channel between a pair of network devices. In this section, we apply the proposed methodology to NetFlow data from Los Alamos National Lab

(LANL). These data are collected from core routers in the LANL enterprise network Turcotte et al. (2018) and recorded at a 1 second resolution meaning multiple events frequently occur at the same time point. We have isolated inter-edge communication channels (edges) in the LANL network, which show non-regular behavior in the communications. In particular, we have chosen three cases with different lengths, intensities and behaviors. Case 1 covers a lower intensity process over 20 hour window, Case 2 shows a shorter window with highly variable counts whilst Case 3 is a shorter window but with a lower average count than observed in Case 2. Figure 10 shows their activity. Note that we have calibrated the Hawkes process with respect to three time-windows of different sizes.

We estimate the parameters using each of the three methods for binned Hawkes processes, assuming the exponential kernel given in Equation (3) for $L = 1$. Performance of the methods is explored via goodness of fit. This is an important practical consideration and allows us to check the quality of the estimates given the data. In literature, it is typical to use the random time change theorem given in Daley and Vere-Jones (2003), Laub et al. (2015). This states that given an unbounded, increasing set of time points $\mathcal{T} = \{t_1, t_2, \ldots\} \in (0, \infty)$ and compensator function

$$\Lambda(t_k) = \int_0^{t_k} \lambda^*(u)\mathrm{d}u,$$

the transformed sequence $\mathcal{T}^* = \{t_1^*, t_2^*, \ldots\} = \{\Lambda(t_1), \Lambda(t_2), \ldots\}$ is a realization of a unit rate Poisson process if and only if the original sequence $\mathcal{T}$ is a realization from the point process defined by $\Lambda(\cdot)$. Therefore, with the estimated conditional intensity function, obtained via the parameter estimates $\hat{\Theta}$, it is possible to transform the observed interarrival times and compare them with the theoretical Exp(1) distribution for a unit rate Poisson process via a QQ-plot. In the case of binned data, the observed time-points are typically not unique and therefore the interarrival times of the the transformed points will be zero-inflated and hence not

Exp(1). This raises the question of how to apply the time rescaling theorem when the true location of the event times is unknown. A discrete version of the time rescaling theorem is presented in Haslinger et al. (2010) and is used to counter the biases introduced by naively treating the discrete events as continuous. However this method relies on the count process being binary, making it inapplicable here. To the best of our knowledge, the only approach which has shown to work well in the case of more than one event per bin is given in Gerhard and Gerstner (2010), where it is suggested that observed event times are uniformly distributed within each interval to create a surrogate process. It is then possible to proceed as if working with continuous time-points. Figure 11 shows the QQ-plots corresponding to the count processes shown in Figure 10. The parameter estimates produced by the BH-EM algorithm have a better fit across all three examples and therefore describe a more viable Hawkes process for each data set. In particular, we note the difference in vertical scale in the QQ-plots, particularly for cases 1 and 2, highlighting that both the INAR($p$) and binned log-likelihood methods do not provide viable parameter estimates, whilst the BH-EM method does. In case 3, we omit the qq-plot generated by using the parameter estimates from the INAR($p$) method as it yielded a negative value for $\hat{\alpha}$. This issue of infeasible estimates when using the INAR($p$) method was also found in our simulation study and is mentioned in Section 1.3.

**5 Conclusion**

Here we presented a new technique for handling binned data using a BH-EM algorithm. By simulating times using a distribution close to that of the latent event times given the observed times and current parameter estimate, we have proposed a surrogate consistent set of modal time-points. This allows estimation of parameters using methods for continuous time-points. We further compared this to the INAR($p$) approximation proposed in Kirchner (2016) and a binned log-likelihood method which assumes a piecewise constant CIF within each interval. For the parameter sets considered, the BH-EM method has appeared to out-perform both alternatives. Applying the BH-EM algorithm to NetFlow data has

also demonstrated superior model fitting to existing methods. The BH-EM approach further provides us with additional flexibility in that the level of aggregation does not need to be consistent across the dataset. That is, provided the interval bounds are known, Δ can vary. The issues arising from binned data could also be handled via a Markov Chain Monte Carlo (MCMC) algorithm and exploring this is the subject of future work.

## Appendices

### A Power-Law Kernel

The proposed BH-EM method can still be applied if intending to consider a regularized power-law kernel of the form

$$g(u) = \frac{\alpha\beta}{(1+\beta u)^{1+c}} \; 1_{u \in \mathbb{R}_+}(u),$$

for $\alpha, \beta, c > 0$. In this case, to ensure a stationary Hawkes process, we have that

$$\begin{aligned} \gamma &= \int_0^\infty \frac{\alpha\beta}{(1+\beta u)^{1+c}} \mathrm{d}u < 1, \\ &\Rightarrow \left[ -\frac{\alpha}{c}(1+\beta u)^{-c} \right]_0^\infty < 1, \\ &\Rightarrow \frac{\alpha}{c} < 1. \end{aligned}$$

Therefore the stationarity condition is met for $\alpha < c$ Bacry et al. (2015).

We also need to consider the proposed method for simulating candidate times and thus the complete log-likelihood of proposed time-points. Both of these points fundamentally rely on expressing the conditional PDF and CDF. Firstly, for the simulation method introduced in Section 2.1, we now have that

$$f^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1},\ldots,t_{n+m}) = \prod_{i=1}^{m} \lambda^*(t_{n+i}) \exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \lambda^*(u)\mathrm{d}u\right),$$

$$= \prod_{i=1}^{m}\left(\nu + \sum_{j=1}^{n+i-1} \frac{\alpha\beta}{(1+\beta(t_{n+i}-t_j))^{1+c}}\right) \exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \nu + \sum_{j=1}^{n+i-1} \frac{\alpha\beta}{(1+\beta(u-t_j))^{1+c}}\mathrm{d}u\right),$$

$$= \prod_{i=1}^{m}\left(\nu + \sum_{j=1}^{n+i-1} \frac{\alpha\beta}{(1+\beta(t_{n+i}-t_j))^{1+c}}\right) \exp(-\nu(t_{n+m}-t_n)) \exp\Big(\sum_{i=1}^{m}\sum_{j}^{n+i-1} \frac{\alpha}{c}\cdot$$

$$\left\{(1+\beta(t_{n+i}-t_j))^{-c} - (1+\beta(t_{n+i-1}-t_j))^{-c}\right\}\Big).$$

Similarly, the joint conditional CDF is given by

$$F^*_{T_{n+1},\ldots,T_{n+m}}(t_{n+1},\ldots,t_{n+m}) = \prod_{i=1}^{m} F^*_{T_{n+i}}(t_{n+i}),$$

$$= \prod_{i=1}^{m}\left(1-\exp\left(-\int_{t_{n+i-1}}^{t_{n+i}} \lambda^*(u)\mathrm{d}u\right)\right).$$

In the case of regularized power-law kernel, this is

$$\prod_{i=1}^{m}\Big(1-\exp\Big(-\int_{t_{n+i-1}}^{t_{n+i}} \nu + \sum_{j=1}^{n+i-1} \frac{\alpha\beta}{(1+\beta(u-t_j))^{1+c}}\mathrm{d}u\Big)\Big).$$

Then, (9) gives the form for the truncated PDF, as previously. All that remains is to adjust the log-likelihood for the CIF with regularized power-law function when implementing the BH-EM algorithm. Using (4) that is,

$$\log \mathcal{L}(\Theta;\mathcal{T}) = \sum_{i=1}^{N_T} \log\left(\nu + \sum_{j=1}^{n+i-1} \frac{\alpha\beta}{(1+\beta(t_{n+i}-t_j))^{1+c}}\right) - \int_0^T \nu + \sum_{j=1}^{n+i-1} \frac{\alpha\beta}{(1+\beta(u-t_j))^{1+c}}\mathrm{d}u.$$

**B Rectangular Kernel**

We can also consider a rectangular kernel of the form

$$g(u) = \frac{n}{\beta-\alpha} 1_{[\alpha,\beta]}(u),$$

$$\equiv \frac{n}{\beta-\alpha}\ 1_{[0,\beta-\alpha]}(u-\alpha),$$

for $\alpha, \beta > 0$. In this case, stationarity holds if

$$\int_{\alpha}^{\beta} \frac{n}{\beta - \alpha} \mathrm{d}u = n < 1.$$

Note that $\alpha$ here represents a small shift of the excitation effect. Therefore, if $\alpha$ = 0, there is an increase in the process intensity immediately after an arbitrary event. In the case of a rectangular kernel,

$$\lambda^*(t) = \nu + \sum_{t_i < t} \frac{n}{\beta - \alpha} 1_{[\alpha,\beta]}(t - t_i),$$

$$= \nu + \frac{n}{\beta - \alpha} \sum_{t_i < t} 1_{[\alpha,\beta]}(t - t_i).$$

The remaining equations follow as previously by substituting the above CIF.

**C The BH-EM algorithm**

Here we provide detailed algorithms for the BH-EM approach. Code has been developed in MATLAB and is available at https://github.com/lshlomovich/MCEM-Univariate-Hawkes.

---

**Algorithm 1** Simulation of times using alternative distribution $q(\mathcal{T}^{*(j)} \mid \mathbf{N}, \Theta^i)$

---

1:

**function** SAMPLE FROM $Q(\mathbf{N}, \Theta^i)$

2:

$\mathcal{T}^{*(j)} \leftarrow \varnothing$

3:

**for** $l$ = 1 to $\left|\{N_k \in \mathbf{N} : N_k \neq 0\}\right|$ **do**

4:

$\{\tilde{t}_1,\ldots,\tilde{t}_m\} \leftarrow$ set which maximizes (9) for $m = N_l$

5:

$\mathcal{T}^{*(j)} \leftarrow \mathcal{T}^{*(j)} \cup \{\tilde{t}_1,\ldots,\tilde{t}_m\}$

6:

**end for**

7:

**end function**

---

**Algorithm 2** BH-EM

---

1:

**function** MCEM($\mathbf{N}, r, \epsilon$)

2:

$\Theta^1 \leftarrow \text{sort}(\text{Unif}(1,3))$

3:

$i \leftarrow 1$

4:

**while** tolerance $> \epsilon$ **do**

5:

**for** $j$ = 1 to $r$ **do**

6:

$\mathcal{T}^{*(j)} \sim q(\mathcal{T} \mid \mathbf{N}, \Theta^i)$ , provided in Algorithm 1

7:

$w_j \leftarrow p(\mathcal{T}^{*(j)} \mid \Theta^i) / q(\mathcal{T}^{*(j)} \mid \mathbf{N}, \Theta^i)$

8:

**end for**

9:

$$Q_{i+1}(\Theta, \Theta^i) \leftarrow \sum_{k=1}^{r} w_k \log(p(\Theta \mid \mathbf{N}, \mathcal{T}^{*(k)})) / \sum_{k=1}^{r} w_k$$

10:

$$\Theta^{i+1} \leftarrow \text{argmax}_{\Theta, \gamma \leq 1} Q_{i+1}(\Theta, \Theta^i)$$

11:

tolerance $\leftarrow \text{norm}(\Theta^{i+1} - \Theta^i)$

12:

$i \leftarrow i+1$

13:

**end while**

14:

**return** $\{\Theta^i\}$ ▷ Set of parameter estimates

15:

**end function**

---

**D Algorithmic Details**

The simulated datasets are generated until a fixed end time $T$. The stationary intensity of a Hawkes process is given in 151971Hawkes by

$$\lambda = \frac{\nu}{1-\gamma}.$$

Thus the expected number events is given by $\lambda T$ and here $T$ is selected to yield between 500 and 1000 events. For the simulations considered here, the run time of the BH-EM algorithm is a factor of $10^3$ times longer than that of the INAR($p$) and binned log-likelihood methods, and this is predominantly due to the simulation method used, which optimises the position of time stamps within each bin via the truncated conditional PDF. In contrast, the INAR($p$) method does not involve any optimisation steps. The simulation study presented has been timed on a 2020 MacBookPro 32GB RAM, 2.3 GHz Quad-Core Intel Core i7, Intel Iris Plus Graphics 1536 MB. The BH-EM algorithm takes an average of 464 seconds per realisation, INAR($p$) takes 0.0707 seconds per realisation, where this includes the optimal selection of $p$ and exponential fit. The binned log-likelihood 0.0282 seconds per realisation. While we have endeavored to code efficiently, we are not claiming this is an optimal implementation of the BH-EM algorithm. Due to the nature of the BH-EM algorithm, volume of events is less of consideration than the average intensity. Having many events in each bin will slow the algorithm more. In the simulations presented here, $r$, is 20. The $L2$ norm

of the difference between successive parameter estimates is used in the BH-EM algorithm for assessing convergence up to a user selected tolerance level. The mean number of iterations is approximately 25, with standard deviation of approximately 15. Further, the memory requirement is $O(n)$, where $n$ is the total number of events.

**References**


E. Bacry, K. Dayri, and J. F. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85 (5):157, 2012.

E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes Processes in Finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015.

C. G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

D. R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological cybernetics*, 59(3):189–200, 1988.

F. Chen and P. Hall. Nonparametric Estimation for Self-Exciting Point Processes–A Parsimonious Approach. *Journal of Computational and Graphical Statistics*, 25(1):209–224, 2016.

F. Chen and T. Stindl. Direct Likelihood Evaluation for the Renewal Hawkes Process. *Journal of Computational and Graphical Statistics*, 27(1):119–131, 2018.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, New York, 2nd edition, 2003.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–22, 1977.

P. Embrechts, T. Liniger, and L. Lin. Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 48(A):367–378, 2011.

V. Filimonov and D. Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, 2012.

S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. C. Vollmer, N. M. Ferguson, L. C. Okell, and S. Bhatt. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, 2020.

J. D. Fonseca and R. Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

F. Gerhard and W. Gerstner. Rescaling, thinning or complementing? On goodness-of-fit procedures for point process models and Generalized Linear Models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23. Curran Associates, Inc., 2010.

R. Haslinger, G. Pipa, and E. Brown. Discrete Time Rescaling Theorem: Determining Goodness of Fit for Discrete Time Statistical Models of Neural Spiking. *Neural Computation*, 22(10): 2477–2506, 2010.

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

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

M. Kirchner. Hawkes and INAR($\infty$) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, 2016.

M. Kirchner. An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4): 571–595, 2017.

M. Kirchner and A. Bercher. A nonparametric estimation procedure for the Hawkes process: comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation*, 88(6):1106–1116, 2018.

R. Kobayashi and R. Lambiotte. TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. *Proceedings of the International AAAI Conference on Web and Social Media*, 10 (1):191–200, 2016.

P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes Processes. *arXiv:1507.02822 [math, q-fin, stat]*, 2015.

T. M. Le. A Multivariate Hawkes Process With Gaps in Observations. *IEEE Transactions on Information Theory*, 64(3):1800–1811, 2018.

E. Lewis and G. Mohler. A Nonparametric EM algorithm for Multiscale Hawkes Processes. In *Proceedings of the 2011 Joint Statistical Meetings*, pages 1–16, 2011.

F. Lorenzen. *Analysis of Order Clustering Using High Frequency Data: A Point Process Approach*. PhD thesis, Tilburg School of Economics and Management, Aug. 2012.

B. Mark, G. Raskutti, and R. Willett. Network Estimation From Point Process Data. *IEEE Transactions on Information Theory*, 65(5):2953–2975, 2019.

K. Obral. Simulation, Estimation and Applications of Hawkes Processes. Master's thesis, University of Minnesota, June 2016.

Y. Ogata. Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

Y. Ogata. Seismicity Analysis through Point-process Modeling: A Review. In *Seismicity Patterns, their Statistical Significance and Physical Meaning*, Pageoph Topical Volumes, pages 471–507. Birkhäuser, 1999.

M. Price-Williams and N. A. Heard. Nonparametric self-exciting models for computer network traffic. *Statistics and Computing*, 30:209–220, 2020.

M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie. A Tutorial on Hawkes Processes for Events in Social Media. In *Frontiers of Multimedia Research*, pages 191–218. Association for Computing Machinery and Morgan & Claypool, 2017.

M. J. M. Turcotte, A. D. Kent, and C. Hash. Unified Host and Network Data Set. In *Data Science for Cyber-Security*, Security Science and Technology, pages 1–22. World Scientific, 2018.

G. C. G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85 (411):699–704, 1990.

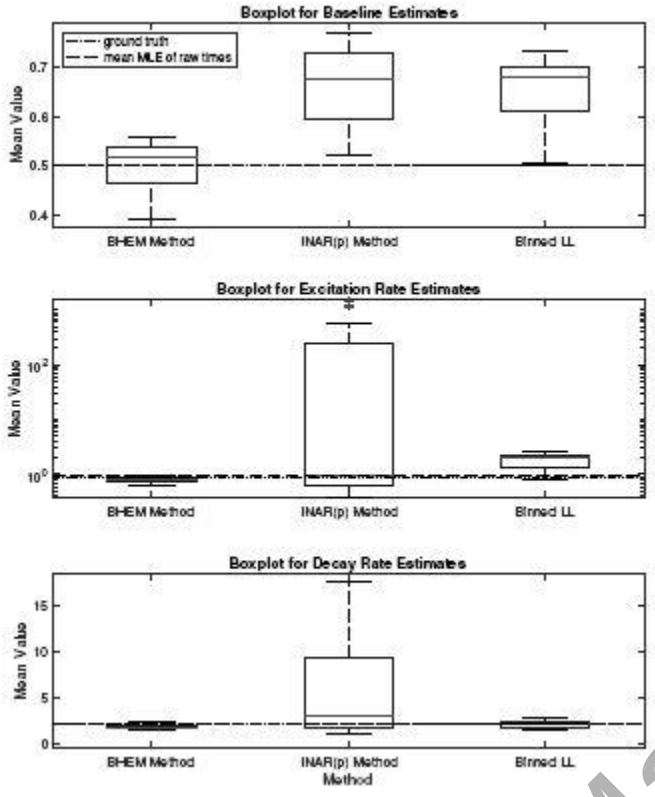


**Fig. 1** Parameter set: $[\nu, \alpha, \beta] = [0.5, 0.9, 2.0]$, $\Delta$ = 1. The mean optimal $p$ was 2.05 with variance 3.73 and 60% equal to 1.

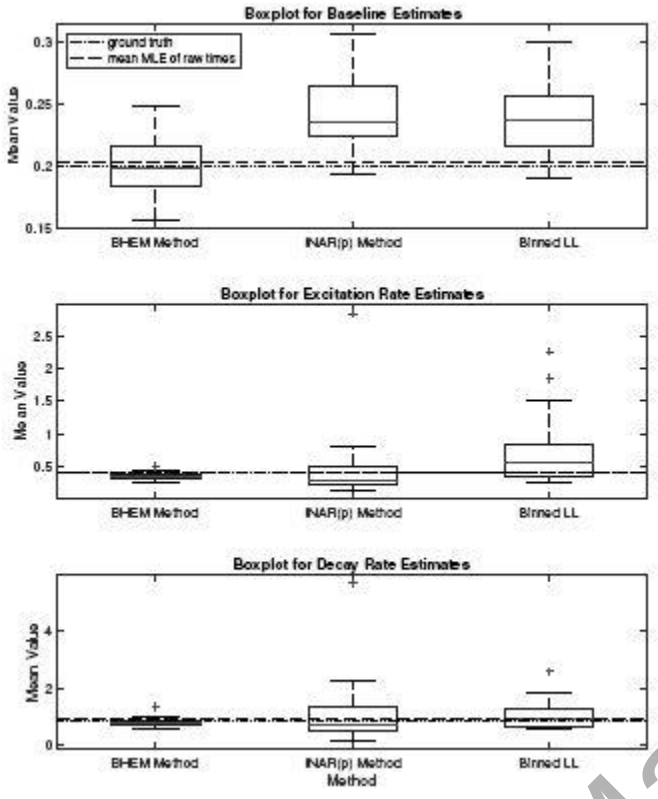


**Fig. 2** $[\nu, \alpha, \beta] = [0.2, 0.4, 0.9]$, Δ = 1. The mean optimal $p$ was 3.59 with variance 6.29 and 10% are found as 1.

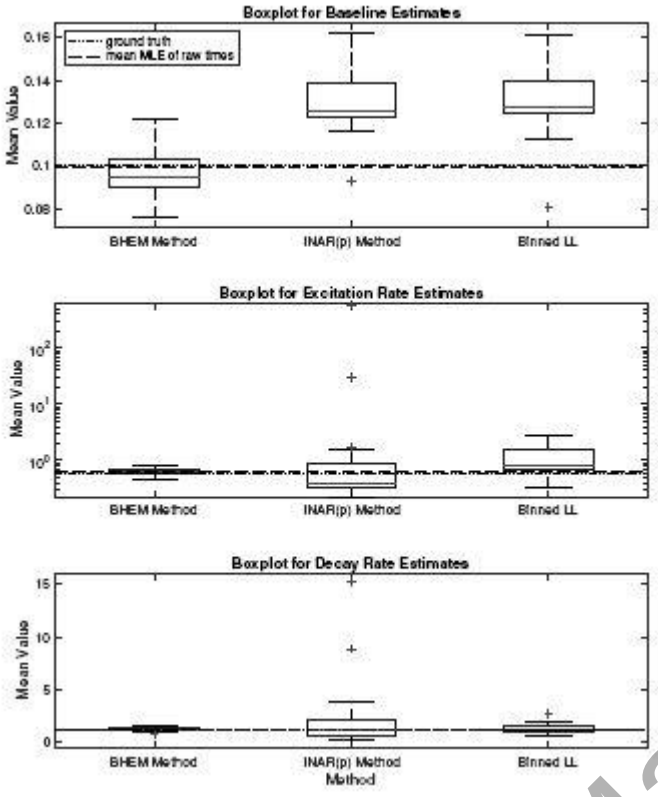


**Fig. 3** $[\nu, \alpha, \beta] = [0.1, 0.6, 1.2]$, Δ = 1. The mean optimal $p$ was 3.65 with variance 5.61 and 20% equal to 1.

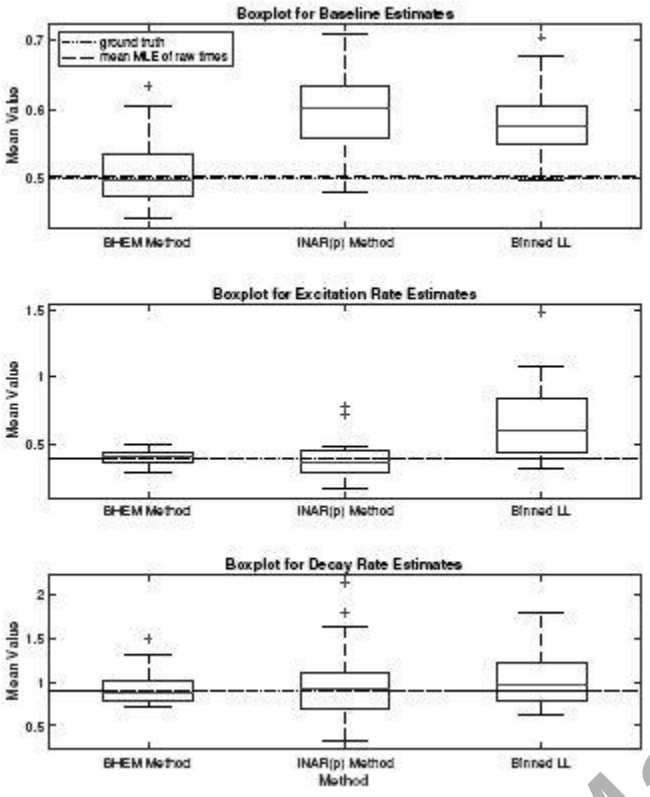


**Fig. 4** $[\nu, \alpha, \beta] = [0.5, 0.4, 0.9]$, Δ = 1. The mean optimal $p$ was 2.85 with variance 4.03 and 10% equal to 1.

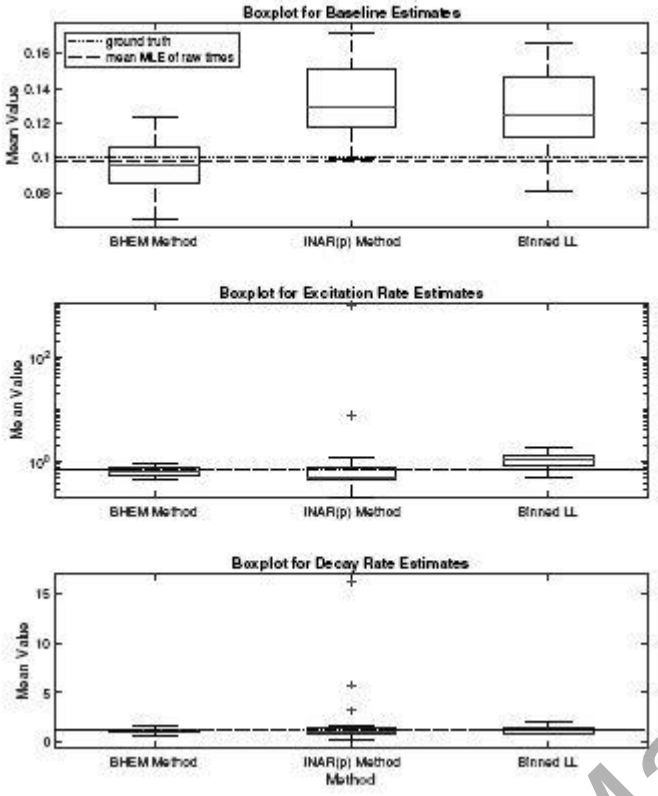


**Fig. 5** $[\nu, \alpha, \beta] = [0.1, 0.7, 1.2]$, Δ = 1. The mean optimal $p$ was 3.25, variance 3.78 and 5% equal to 1.

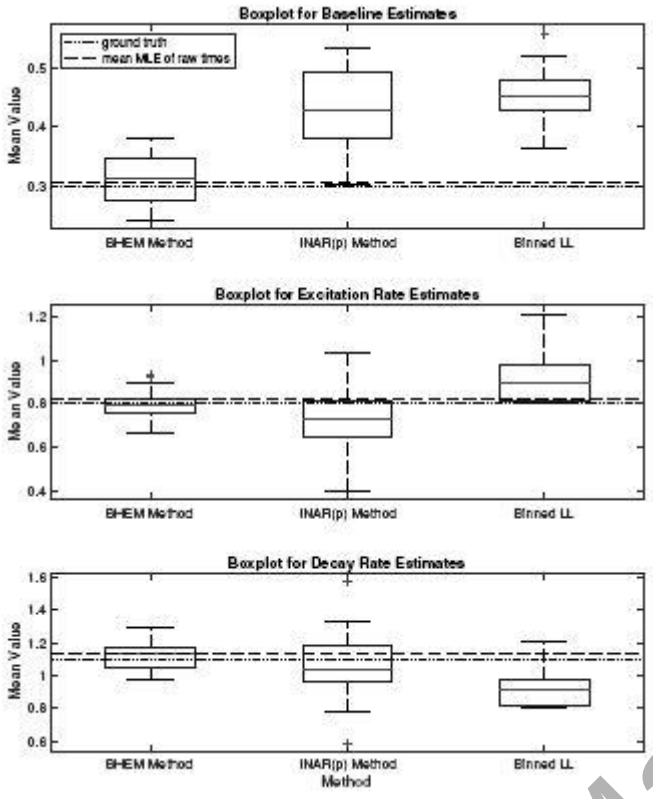


**Fig. 6** $[\nu, \alpha, \beta] = [0.3, 0.8, 1.1]$, Δ = 1. The mean optimal $p$ was 3.00 with variance 1.68 and none equal to 1.

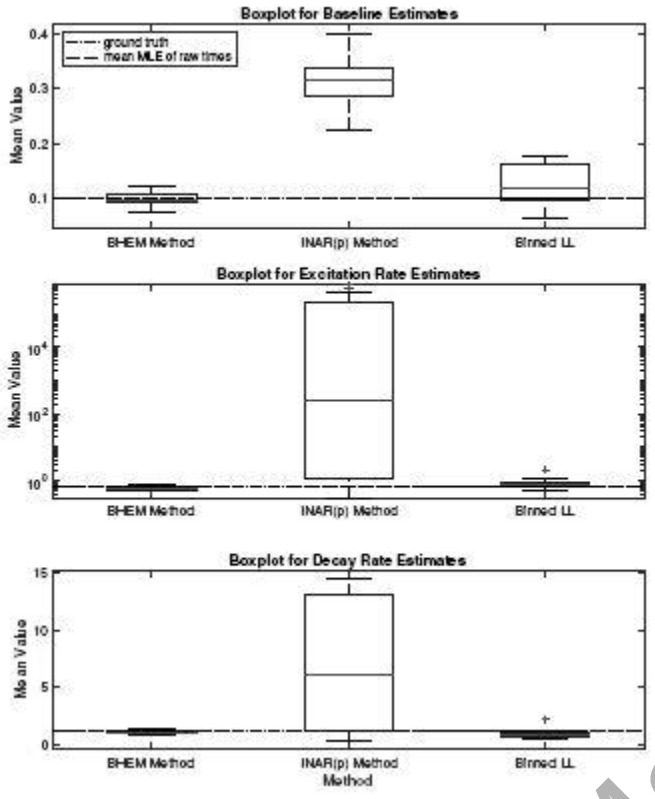


**Fig. 7** Parameter set: $[\nu, \alpha, \beta] = [0.1, 0.7, 1.2]$, $\Delta = 2$. The mean optimal $p$ was 1.55, variance 1.31 and 75% being 1.

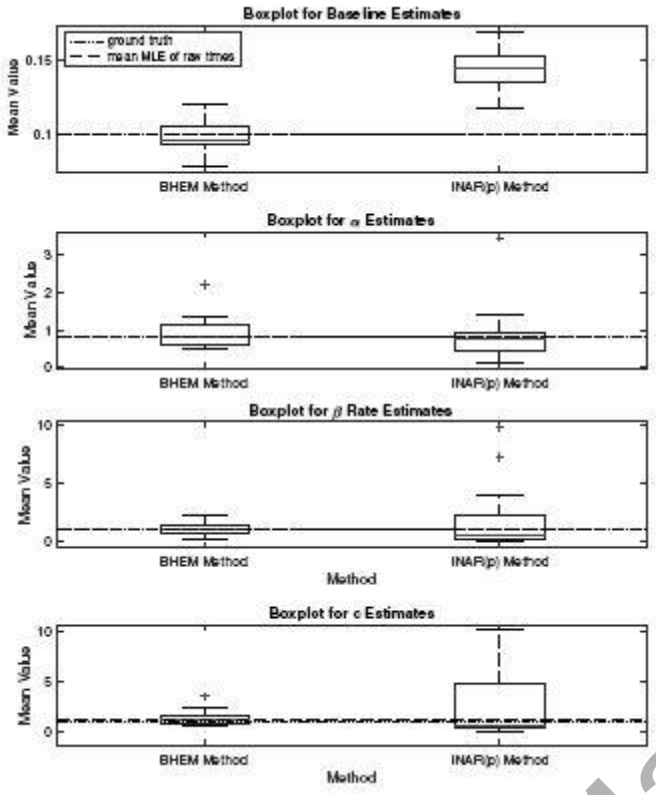


**Fig. 8** Power-law kernel with $[\nu, \alpha, \beta, c] = [0.1, 0.8, 1, 1.2]$, Δ = 1. The form for this kernel is given in Appendix A.

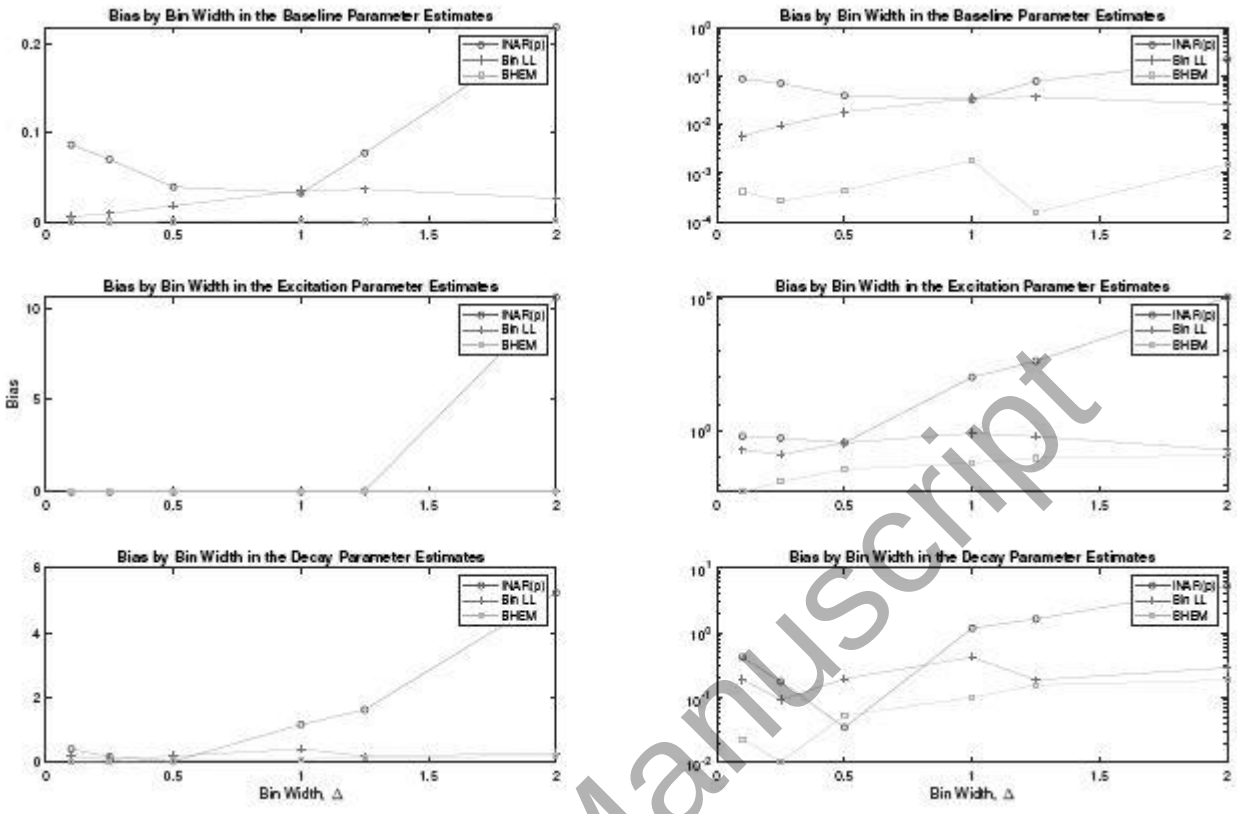


**Fig. 9** Parameter set: $[\nu, \alpha, \beta] = [0.1, 0.7, 1.2]$, $\Delta = [0.1, 0.25, 0.5, 1, 1.25, 2]$ . Left figure shows the results on a linear scale, whilst the right shows a log scale.

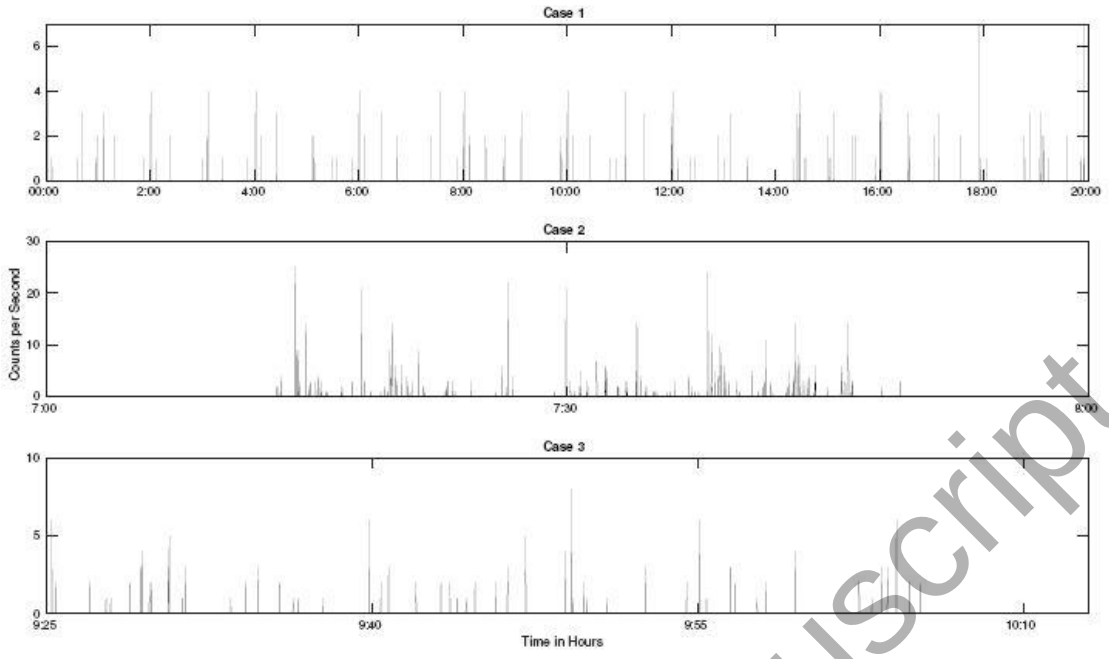


**Fig. 10** Time-stamps of NetFlow data on an edge in the LANL network.

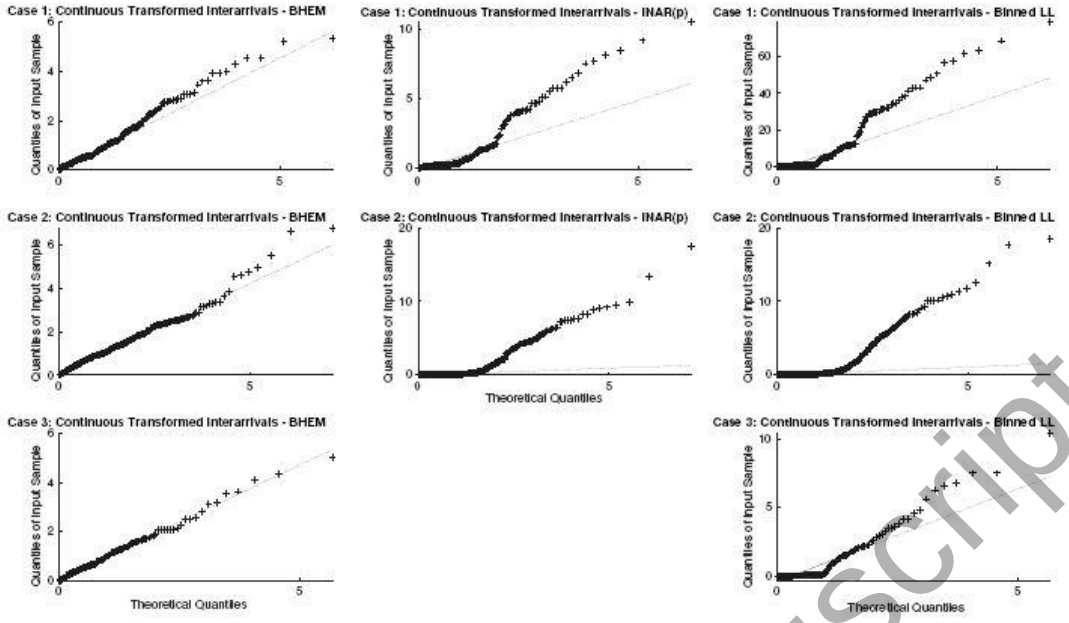


**Fig. 11** QQ-plots of transformed time-points using parameters estimated from each of the three methods considered for an edge in the LANL network. The missing figure in case 3 is due to infeasible parameter estimates generated by the INAR(*p*) method.