Stat Inference Stoch Process (2013) 16:63–80
DOI 10.1007/s11203-013-9074-1

Exact and approximate EM estimation of mutually exciting hawkes processes

Jamie F. Olson · Kathleen M. Carley

Received: 22 October 2010 / Accepted: 1 February 2013 / Published online: 23 February 2013

Abstract Motivated by the availability of continuous event sequences that trace the social behavior in a population e.g. email, we believe that mutually exciting Hawkes processes provide a realistic and informative model for these sequences. For complex mutually exciting processes, the numerical optimization used for univariate self exciting processes may not provide stable estimates. Furthermore, convergence can be exceedingly slow, making estimation computationally expensive and multiple random restarts doubly so. We derive an expectation maximization algorithm for maximum likelihood estimation mutually exciting processes that is faster, more robust, and less biased than estimation based on numerical optimization. For an exponentially decaying excitement function, each EM step can be computed in a single $O(N)$ pass through the data, for N observations, without requiring the entire dataset to be in memory. More generally, exact inference is $\Theta(N^2)$, but we identify some simple $\Theta(N)$ approximation strategies that seem to provide good estimates while reducing the computational cost.

Keywords Self-exciting point processes · Estimation · Expectation-maximization

1 Introduction

The formation and evolution of human social relations has long been a topic of research. Previously, such research has been based on a sequence of discrete observations of the entire social system, either through ethnographic observation or self-report surveys. Such observations are difficult, however, and generally such studies would

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Army Research Office, the U.S. Army Engineering Research and Development Center's (ERDC), Topographic Engineering Center, the National Science Foundation or the U.S. government.

J. F. Olson · K. M. Carley (✉)
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: kathleen.carley@cs.cmu.edu

contain only a few, widely spaced time periods. The proliferation of so-called "social-networking" websites and other means of electronic social interaction and communication opens up a new range of possibilities for modeling social dynamics. In particular, it becomes feasible to consider modeling the micro-dynamics of social behavior. Rather than modeling the general change in relationships, we can seek to model how specific social events are influenced by the actions of other participants in the social system. Social activity data also pose several problems, chief among them being the scalability of any useful analysis techniques. By considering not only the existence of a social relation, but the event-based evidence of that social relation, we shift the magnitude of the dataset from the total number of social relations to the total amount of social activity. Any practical model must scale well in order to be useful for the large and dynamic online social systems and even smaller populations of individuals may produce a large volume of social activity (e.g. email within a organization or company).

Much effort has been expended in computational and statistical models of human social behavior. However, many of these models (Snijders and Nowicki 1997; Airoldi et al. 2005; Hoff et al. 2002; Hoff 2008; Krivitsky and Handcock 2008) pertain to complete observations of a social network and are therefore entirely appropriate for event-level modeling of streams of social behavior. For example, specific events may have direct causal relations with other specific events, e.g. email replies, wiki edits, relationships that are not captured through models of the macro-level latent structure. Furthermore,temporal extensions of such models (Sarkar and Moore 2005; Guo et al. 2007; Kolar et al. 2009; Snijders 1996) often require a Markov assumption to yield a computationally feasible model. Although this may often be a useful simplifying assumption, it may not be sufficient to fully capture the event-level dynamics of social behavior. Social activity may not always occur in response to the events or the system state immediately prior. The information system through which activity occurs, cognitive processing required to act and wandering attentional may all conspire to yield non-trivial delays between related social events (e.g. an email and it's caused response). Therefore, we believe that there is a need for more flexible models capable of representing arbitrary distributions of response-times between causally linked events.

Self-exciting point processes provide a statistical model for these kinds of dynamics that can reproduce various qualitative and quantitative features of human social events. Various formulations of Hawkes' processes have been used to model crime (Egedsdal et al. 2010), stock markets (Large 2007), and Youtube movie (Crane and Sornette 2008) and news website (Johansen and Sornette 2000) views and book sales (Deschatres and Sornette 2005) for their ability to model burstiness and endogenous versus exogenous effects. We believe that more widespread adoption of these Hawkes' process models is limited by current estimation techniques, which generally rely on numerical optimization. This can lead to problems both in the robustness of the estimation as well as the computational cost and convergence properties. We derive an efficient expectation-maximization algorithm for mutually exciting Hawkes process for use with any decay function. In Sects. 2 and 3, we will review the relevant prior work on social dynamics and self-exciting Hawkes point processes. Sections 4 and 5 will describe, respectively, exact and approximate inference for the Hawkes using expectation-maximization and Sect. 6 will describe some results on the robustness and accuracy of the EM estimation. Following a brief discussion, we conclude in Sect. 8 by discussing limitations and implications of our findings.

2 Background

3 Self-exciting point processes

The self-exciting Hawkes point process was introduced by Hawkes (1971a,b), who further derived a useful interpretation that we use here (Hawkes and Oakes 1974). Intuitively, a Hawkes process is composed of two stochastic mechanisms. First, some background process continuously and uniformly generates events. Second, any generated event has the potential to lead directly to some future event. Equivalently, there is a population that is altered by both the steady immigration of new individuals and newly born (asexual) descendants of a single individual in the existing population, and the appearance of new individuals corresponds to events in the process. Hawkes (1971b) defines the process in terms of its intensity function:

$$\Lambda(t|H_t) = \mu + \int_{-\infty}^{t} g(t-u)dN(u) \tag{1}$$

The background immigration intensity is represented by u and each individual in the population contributes some excitation $g(t-u) \geq 0$ where u is arrival time of the exciting individual and t is the (potential) arrival time of the excited individual. This process depends only on historical data H_t before time t, not on future data. As with other point processes, the Hawkes process is defined in terms of its counting process and the intensity function.

$$\lim_{\delta \to 0} \frac{1}{\delta} P(N(t, t+\delta) = 1|H_t) = \Lambda(t|H_t) \tag{2}$$

We will refer to the counting process using both the number of events before time t, denoted $N(t)$, as well as the number of events between times s and t, denoted $N(s,t)$, as in Eq. (2). The Hawkes process defined in this way is conditionally orderly (at most one event at any time) and stationary (has constant, finite expectation), only if Eq. (3) holds (Hawkes 1971b). We can interpret Eq. (3) as indicating that the expected number of first generation descendants from any particular individual is strictly less than 1. Stationarity also requires the expected number of all descendants (first and subsequent generations) be finite.

$$\int_{-\infty}^{\infty} g(v)dv < 1 \tag{3}$$

Note that stationarity is defined in terms of the unconditional expectation $E[\Lambda(t)]$ whereas we have defined $\Lambda(t|H_t)$ conditioned on the historical data, H_t in Eq. (1).

The log likelihood for a point processes is defined in terms of this conditional intensity.

$$L(\theta) = \sum_{t_i} \log \Lambda(t_i|H_t) - \int_{0}^{T} \Lambda(t|H_t)dt \tag{4}$$

$$= \sum_{t_i} \log \Lambda(t_i|H_t) - \left(\mu T + \sum_{t_i} \int_{t_i}^{T} g(v-t_i)dv \right) \tag{5}$$

Hawkes (1971b) also defines the mutually-exciting point process with K different streams, each potentially influencing the other. Formally, Eq. (1) becomes

$$\Lambda_l(t|H_t) = \mu_l + \sum_j \int_{-\infty}^{t} g_{jl}(t-u)dN_j(u) \tag{6}$$

In addition to defining these processes, Hawkes (1971b) also derive the spectra for the specific case where the excitation function, $g(v)$ is a sum of exponential functions.

$$g(v) = \sum_{k=1}^{K} \alpha_k e^{-\beta_k v} \tag{7}$$

We simplify the aforementioned model slightly, re-parameterizing to to separate the decay function from the total expected excitement. Specifically, we define the intensity function as

$$\Lambda_l(t|H_t) = \mu_l + \sum_{k=1}^{K} \int_{-\infty}^{t} \beta_{kl} f_l(t-u)dN(u) \tag{8}$$

and further require that

$$\int_0^{\infty} f(v)dv = 1$$
$$and \quad \beta_{kl} < 1, \forall l, k \tag{9}$$

This, combined with requiring that $f(v) > 0$ implies that $f(v)$ is a probability distribution over $v > 0$, which is convenient both in that it allows us to refer to $P(v)$ but also in that any continuous distribution over $v > 0$ can be used as the excitement function. We can now interpret the β_{kl} parameters as the expected number of first generation descendant events in stream k caused by an event in stream l.

Here we have defined β_{kl} as dependent on the stream, k, containing the "parent" event and the stream, l, containing the "child" event; whereas, the decay function is dependent only on the stream containing the "child" event. Interpreting this in the context of socially active people, we might say that certain people are more excited to act by specific others but that each person has their own characteristic response function that governs the amount of time they take to respond regardless of who they're responding to. The proposed EM inference is not, however, limited to this specific parameterization, rather, we chose this as a simple model that illustrates the more important characteristics of Hawkes processes without unnecessary complications. The model can also be parameterized; e.g., with multiple excitation "processes", depending on observed properties of the individual events rather than this fixed set of "streams" and the interactions between them. That is, the k stream contains the parent event and the l stream contains the child event.

In the case of exponential decay, this becomes

$$f_{\exp}(v) = \alpha e^{-\alpha v} \tag{10}$$

Exponentially decaying excitation has some convenient properties that make exact inference computationally efficient. Distributions with longer range influences, such as the Pareto and Log-Normal distributions pose problems for exact inference and motivate some of the approximation strategies presented in Sect. 4; i.e.,

$$f_{\text{pareto}}(v) = \frac{\omega c^{\omega}}{v^{1+\omega}} \tag{11}$$

$$f_{\text{lnorm}}(v) = \frac{1}{v\sqrt{2\pi\sigma^2}} \exp^{\frac{ln(v-\nu)}{2\sigma^2}} \tag{12}$$

3.1 Simulating hawkes processes

There are many different techniques for simulating self-exciting point processes (Ogata 1981; Møller and Rasmussen 2005). One of the more common techniques begins with a uniform distribution of events and uses filtering to remove most of them. This strategy has been found to introduce biases into the resulting point processes. Instead, we use the cluster process representation, simulating events exactly according to the initial explanation of the self-exciting processes. Events are sampled from a simple homogeneous Poisson process with a rate equal to the background rate of the self-exciting process (μ). Next, each newly generated event will cause some number of future events in the other streams. The number of events to be created is a Poisson random variable with rate β_{kl} if the original event was in stream k and the future event(s) is in stream l. Next, our formulation of $f(v)$ means that it is a probability distribution over the time delay, $t > 0$, so we sample from that distribution to determine the time(s) for the new event(s). We repeat this step for any new events.

4 Exact inference

Maximum likelihood estimation of Hawkes process parameters is widely performed via numerical optimization (Ozaki 1979; Ogata 1998; Zhuang et al. 2002). Ozaki (1979) initially identified the derivatives and Hessian for the exponential-decay process (see Eq. 7) which were subsequently extended for more complex distributions (Ogata 1998). Unfortunately, numerical optimization can lead to substantial computational costs as well as to poor solutions depending on the specific shape of the likelihood function (Veen and Schoenberg 2006).

Although several different fields have seen recent work exploring Hawkes processes, the seismological community has long been using such models in the form of the Epidemic-type Aftershock Sequence (ETAS) (Ogata 1988, 1998). The ETAS model expands the model proposed by Hawkes (1971b) to incorporate both spatial location and magnitude. As such, closed form solutions for parameter estimates are generally not available and often numerical optimization is used. In attempt to solve the inference problems related to numerical optimization, Veen and Schoenberg (2006) proposed a partial information expectation-maximization strategy for parameter estimation. They introduce hidden parameters based on the stochastic deconstruction of the process (Zhuang et al. 2002). This means determining for each event in the process whether it is an immigrant or a descendant and if it is a descendant, which prior event is its immediate ancestor. We adopt this same strategy in developing efficient inference for our simpler process but take advantage of two important recurrence relations to improve the space and time efficiency.

4.1 Maximization

Following previous work (Veen and Schoenberg 2006), we introduce hidden variables u_j representing the immediate ancestor (cause) of the event at time t_j as well as indicators u_{ij} which equals 1 if $u_j = i$ and 0 otherwise. Immigrants have no immediate ancestor and so for an immigrant, j, $u_j = 0$ or equivalently $u_{0j} = 1$. We can now write the complete data log likelihood:

$$\begin{aligned} L_{CD}(\theta) &= \sum_{t_j} \left(u_{0j} \log \mu + \sum_{t_i < t_j} u_{ij} \log \beta f(t_j - t_i) \right) \\ &\quad - \int_0^T \Lambda(t|H_t)dt \\ &= \sum_{t_j : u_j = 0} \log \mu + \sum_{t_j : u_j \neq 0} \left(\log \beta + \log f(t_j - t_{u_j}) \right) \\ &\quad - \left(\mu T + \beta \sum_{t_i} F(t_i) \right) \end{aligned} \tag{13}$$

where

$$F(t_i) = \int_{t_i}^{T} f(v - t_i)dv \tag{14}$$

Intuitively, the first group in Eq. (13) indicates how likely the actual sequence of events is, while the other term, $\int_0^T \Lambda(t|H_t)dt$ describes how unlikely it was to have not seen additional events. More precisely, the first group sums over all observed events, the log of the conditional intensity when the event happened, whereas the other, negative term integrates the conditional intensity over the entire time interval. The somewhat simplified integral in Eq. (14) is similarly over (t_i, T) to account for the fact that a specific interval in time was sampled.

The derivatives with respect to μ and β are

$$\frac{\partial L_{CD}(\theta)}{\partial \mu} = \sum_{t_j : u_j = 0} \frac{1}{\mu} - T \tag{15}$$

$$\frac{\partial L_{CD}(\theta)}{\partial \beta} = \sum_{t_j : u_j \neq 0} \frac{1}{\beta} - \sum_{t_i} F(t_i) \tag{16}$$

In order to simplify the notation, we introduce $m_0 = \sum_{t_j : u_j = 0} 1$, the total number of events generated by the background process, and $m_1 = \sum_{t_j : u_j \neq 0} 1$. Conditional maximum likelihood estimates for the μ, β parameters are then

$$\hat{\mu}_{EM} = \frac{m_0}{T} \tag{17}$$

$$\hat{\beta}_{EM} = \frac{m_1}{\sum_{t_i} F(t_i)} \tag{18}$$

Although this maximization of the exact likelihood yields convenient conditional estimates for the μ, β parameters, it is not quite so straightforward to estimate the parameters for the excitation function. Even in the simple case of exponential decay, there is no analytical solution for the rate parameter, α, and instead a root-finding algorithm is necessary

$$
\begin{aligned}
F_{\exp}(t_i) &= \int\limits_{t_i}^{T} f_{\exp}(v)dv \\
&= 1 - e^{-\alpha(T-t_i)} && (19) \\
\frac{\partial L_{CD}(\theta)}{\partial \alpha} &= \sum_{t_j:u_j\neq 0}\left(\frac{1}{\alpha} - (t_j - t_{u_j})\right) - \sum_{t_i}\beta\,(T-t_i)\,e^{-\alpha(T-t_i)} \\
&= \frac{m_1}{\alpha} - \sum_{t_j:u_j\neq 0}\left(t_j - t_{u_j}\right) - \sum_{t_i}\beta\,(T-t_i)\,e^{-\alpha(T-t_i)} && (20)
\end{aligned}
$$

The situation is similar in the case where the excitation function is specified by the Pareto distribution (Eq. (11)). The intensity function and consequently, the log likelihood, increase monotonically with c. Because the Pareto distribution assigns a probability of 0 to any value of v less than c, the maximum likelihood estimate for c is $\min_j(t_j - t_{u_j})$, the minimum elapsed time between a parent and child event, but since this is unknown we instead assign c to be $\min_j(t_j - t_{j-1})$, the minimum elapsed time between any two events. It should be noted that this is a heuristic approximation that could be replaced by an expectation conditional maximization using the expectation information. By fixing c, we only need to consider estimation of the shape parameter, ω

$$
\begin{aligned}
F_{\text{pareto}}(t) &= 1 - \left(\frac{c}{t}\right)^{\omega} \\
\frac{\partial L_{CD}(\theta)}{\partial \omega} &= \sum_{t_j:u_j\neq 0}\left(\frac{1}{\omega} + \log c - \log(t_j - t_{u_j})\right) \\
&\quad -N\beta\sum_{t_i}\frac{c/\,(T-t_i)^{\omega}}{\log\left(c/(T-t_i)\right)} \\
&= m_1\left(\frac{1}{\omega} + \log c\right) - \sum_{t_j:u_j\neq 0}\log\left(t_j - t_{u_j}\right) \\
&\quad -N\beta\sum_{t_i}\frac{c/\,(T-t_i)^{\omega}}{\log\left(c/(T-t_i)\right)} && (21)
\end{aligned}
$$

As with exponential decay, there is no closed-form solution to the exact conditional maximum likelihood estimates for the ω parameter of the Pareto distribution.

Because there is no closed-form solution for the cumulative density function of the Log-Normal distribution, $F_{\text{lnorm}}(v)$, it is helpful to consider the partial derivatives of the two excitation-related components of the complete data likelihood separately. In general, for a parameter η,

$$
\frac{\partial L_{CD}(\theta)}{\partial \eta} = \sum_{t_i}\frac{\partial}{\partial \eta} logf(t_i) - \beta\sum_{t_i}\frac{\partial}{\partial \eta}F(t_i) \qquad (22)
$$

The partial derivatives of the first term, $logf(t)$, are straightforward.

$$
\begin{aligned}
\frac{\partial}{\partial \nu} logf(t) &= \frac{log(t_j - t_{u_j}) - \nu}{\sigma^2} \\
\frac{\partial}{\partial \sigma} logf(t) &= -\frac{1}{\sigma} + \frac{(log(t_j - t_{u_j}) - \nu)^2}{\sigma^3} && (23)
\end{aligned}
$$

The partial derivatives with respect to F(t) require a change of variables in order to apply the fundamental theorem of calculus. Rather than starting from the definition, $F(t_i) = \int_{t_i}^{T} f(\nu - t_i) d\nu$, we use a definition of the cumulative distribution function of the Log-Normal distribution in terms of the complementary error function, $F(t) = erfc\left(\frac{log(t-\nu)}{\sigma\sqrt{2}}\right)$, which offers a clearer derivation.

$$
\begin{aligned}
\frac{\partial}{\partial \nu} F(t) &= \frac{\partial}{\partial \nu}\frac{1}{2} erfc\left(\frac{log(t-\nu)}{\sigma\sqrt{2}}\right) \\
&= \frac{\partial}{\partial \nu}\frac{1}{2}\left(1 - erf\left(\frac{log(x-\nu)}{\sigma\sqrt{2}}\right)\right) \\
&= \frac{\partial}{\partial \nu}\frac{1}{2}\left(\frac{2}{\sqrt{\pi}} \int\limits_{0}^{\frac{log(t-\nu)}{\sigma\sqrt{2}}} \exp^{-x^2}\right) dx \\
&= \frac{\partial}{\partial \nu}\frac{1}{\sqrt{\pi}} \int\limits_{\nu}^{log(t)} \exp^{-(\frac{x-\nu}{\sigma\sqrt{2}})^2} \frac{1}{\sqrt{2}\sigma} dx \\
&= \begin{cases} -\int_{log(t)}^{\nu} \frac{1}{\sqrt{2\pi}\sigma} \exp^{-(\frac{x-\nu}{\sigma\sqrt{2}})^2} dx & \log(t) < \nu \\ \int_{\nu}^{log(t)} \frac{1}{\sqrt{2\pi}\sigma} \exp^{-(\frac{x-\nu}{\sigma\sqrt{2}})^2} dx & \log(t) > \nu \end{cases} \\
&= sign(log(t) - \nu) f(t)
\end{aligned} \tag{24}
$$

Similarly, the partial derivatives with respect to σ can be derived

$$
\begin{aligned}
\frac{\partial}{\partial \sigma} F(t) &= \frac{\partial}{\partial \sigma}\frac{1}{2} erfc\left(\frac{log(t-\nu)}{\sigma\sqrt{2}}\right) \\
&= \frac{\partial}{\partial \sigma}\frac{1}{2}\left(\frac{2}{\sqrt{\pi}} \int\limits_{0}^{\frac{log(t-\nu)}{\sigma\sqrt{2}}} \exp^{-x^2}\right) dx \\
&= \frac{\partial}{\partial \sigma}\frac{1}{\sqrt{\pi}} \int\limits_{0}^{\sigma} \exp^{-(\sigma(\frac{log(t-\nu)}{\sigma^2\sqrt{2}}))^2} \frac{log(t-\nu)}{\sigma^2\sqrt{2}} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp^{-(x(\frac{log(t-\nu)}{\sigma^2\sqrt{2}}))^2} \frac{log(t-\nu)}{\sigma} \\
&= t f_{\text{lnorm}}(t) \frac{log(t-\nu)}{\sigma}
\end{aligned} \tag{25}
$$

The two components are then combined to compute the complete data partial derivatives.

4.2 Expectation

Naive computation of the hidden variables introduced above would require $O(N^2)$ space requirements for each of the u_{ij} hidden variables. However, the update equations above can be restated in terms of the expectations, $E[m_{0l}]$, $E[m_{kl}]$ and $\sum E[t_j - t_{u_j}]$, which can all be computed in $O(K)$ space. Because we require the process to be conditionally orderly, we can alternatively state the hidden values in terms of $\Delta_j = t_j - t_{u_j}$ because Δ_j uniquely determines

u_j and vice versa. Although it is more intuitive to describe the hidden values in terms of the causal relationship using the u_j, estimation is more straightforward for $E[\Delta_j] = E[t_j - t_{u_j}]$. We will continue to refer the hidden values u_j even though the values we are estimating are the Δ_j. We also introduce variables, z_i, indicating the stream containing the event at time t_i.

$$\hat{\mu}_l = \frac{E[m_{0l}]}{T} \tag{26}$$

$$\hat{\beta}_{kl} = \frac{E[m_{k1}]}{N_k} \tag{27}$$

$$\hat{\alpha}_l = \sum_k E[m_{kl}] \left(\sum_{t_j : z_j = l, u_j = i \neq 0} E[t_j - t_{u_j}] \right)^{-1} \tag{28}$$

The expectations $E[m_{0l}]$, $E[m_{kl}]$ are simple sums of the indicators, u_{ij}.

$$\begin{aligned} E[u_{ij}] &= P(i \text{ produced } j) \\ &= \begin{cases} \mu/\Lambda(t|H_t) & i = 0 \\ \beta_{z_i z_j} \alpha_{z_j} e^{-\alpha_{z_j}(t_j - t_i)}/\Lambda(t|H_t) & i > 0 \end{cases} \end{aligned} \tag{29}$$

$$\begin{aligned} E[m_{kl}] &= \sum_{z_j = l} \sum_{t_i < t_j, z_i = k} \beta_{kl} \alpha_{z_j} e^{-\alpha_{z_j}(t_j - t_i)}/\Lambda(t|H_t) \\ &= \sum_{z_j = l} \beta_{kl} \sum_{t_i < t_j, z_i = k} \alpha_{z_j} e^{-\alpha_{z_j}(t_j - t_i)}/\Lambda(t|H_t) \end{aligned} \tag{30}$$

Define $A_k(t_j)$ as follows and a simple recurrence allows calculate the updated $A_k(t_{j+1})$

$$A_k(t_j) = \sum_{t_i < t_j, z_i = k} \alpha_k e^{-\alpha_k(t_j - t_i)} \tag{31}$$

$$A_k(t_{j+1}) = A_k(t_j) e^{-\alpha_k(t_{j+1} - t_j)} + \alpha_k e^{-\alpha_k(t_{j+1} - t_j)} \tag{32}$$

Now $E[m_{kl}]$ becomes

$$\forall k > 0 \quad E[m_{kl}] = \sum_{z_j = l} \frac{\beta_{kl} A_k(t_j)}{\mu_l + \sum_\kappa \beta_{\kappa l} A_\kappa(t_j)} \tag{33}$$

$$E[m_{0l}] = \sum_{z_j = l} \frac{\mu_l}{\mu_l + \sum_\kappa \beta_{\kappa l} A_\kappa(t_j)} \tag{34}$$

If we now define $B_k(t_i)$, we can also compute $\sum E[t_j - t_{u_i}]$ in a single pass through the data with $O(K)$ memory

$$B_k(t_j) = \sum_{t_i < t_j, z_i = k} (t_j - t_i) \alpha_k e^{-\alpha_k(t_j - t_i)} \tag{35}$$

$$\begin{aligned} B_k(t_{j+1}) &= B_k(t_j) e^{-\alpha_k(t_{j+1} - t_j)} + A_k(t_j)(t_j - t_i) e^{-\alpha_k(t_{j+1} - t_j)} \\ &\quad + (t_j - t_i) \alpha_k e^{-\alpha_k(t_{j+1} - t_j)} \end{aligned} \tag{36}$$

$$E[t_j - t_{u_i}] = \sum_{z_j = l} \frac{\beta_{kl} B_k(t_j)}{\mu_l + \sum_\kappa \beta_{\kappa l} A_\kappa(t_j)} \tag{37}$$

These expectations can thus be calculated in a single pass, requiring only $O(K)$ memory to keep track of the current set of $\mathbf{A}(t_j)$ and $\mathbf{B}(t_j)$.

Temporal decay functions other than the simple exponential decay will not have such efficient exact expectation calculations. The quantities, $A_k(t_j)$, $B_k(t_j)$ can be more generally stated as

$$A_k(t_j) = \sum_{t_i < t_j, z_i = k} f(t_j - t_i) \tag{38}$$

$$B_k(t_j) = \sum_{t_i < t_j, z_i = k} (t_j - t_i) f(t_j - t_i) \tag{39}$$

Using the Pareto distribution for the temporal decay function, these do not seem to have any useful recurrence relation that would allow us to simplify these quantities.

$$A_k(t_j) = \sum_{t_i < t_j, z_i = k} \frac{\omega c^{\omega}}{(t_j - t_i)^{1+\omega}} \tag{40}$$

$$B_k(t_j) = \sum_{t_i < t_j, z_i = k} (t_j - t_i) \frac{\omega c^{\omega}}{(t_j - t_i)^{1+\omega}} \tag{41}$$

Direct calculation from these formulas is possible, but for each observation, i, this requires iterating over all $i - 1$ previous observations, leading to a computational cost $O(N^2)$ for N observations, for each step of the EM algorithm. This also requires the ability to hold the entire sequence in memory as the final observation will require iterating over all $N - 1$ previous observations to compute the quantities $A_k(t_N)$, $B_k(t_N)$. While this is reasonable for small or moderate sized datasets, it is impractical for very large collections of many different streams.

5 Approximate inference

In the case of an exponentially decaying excitation function, we derived an efficient E-step that requires only a single pass through the data and can be computed in $O(N)$ time for N events. For the Pareto distribution and in general, the exact E-step requires $O(N^2)$ time to calculate. Furthermore, the exact conditional maximum likelihood estimates in the M-step require potentially many passes through the data, depending on the convergence of the root-finding algorithm. We seek now to improve upon these results by using several approximation strategies.

5.1 Maximization

Recall that even in the case of exponentially decaying excitation, there is no closed form M-step for estimating the parameters of the excitation function:

$$F_{\exp}(t_i) = \int_{t_i}^{T} f_{\exp}(v) dv \tag{42}$$

$$= 1 - e^{-\alpha(T - t_i)} \tag{43}$$

$$\frac{dL_{CD}(\theta)}{d\alpha} = \sum_{t_j : u_j \neq 0} \left(\frac{1}{\alpha} - (t_j - t_{u_j}) \right) - \sum_{t_i} \beta (T - t_i) e^{-\alpha(T - t_i)} \tag{44}$$

$$= \frac{m_1}{\alpha} - \sum_{t_j:u_j \neq 0} \left(t_j - t_{u_j}\right) - \sum_{t_i} \beta \left(T - t_i\right) e^{-\alpha(T-t_i)} \tag{45}$$

This function is well-behaved and in many situations this may not be problematic. For very large sequences of observations it may become quite computationally expensive to compute $\frac{dL_{CD}(\theta)}{d\alpha}$ because it requires access to each element t_i in the stream. If the integral in $F(t_i)$ is taken from (t_i, ∞), ignoring any effect of the time interval sampled, this quantity is by Eq. (3) equal to 1. If we assume a rapidly decaying excitation function, e.g. exponential decay, then we are already assuming that long-range influences are few. As such, the only observations for which this censoring might impact are the final few and far large sequences, these censoring factors, $F(t_i)$, are almost always close to 1. For decay functions $f(v)$ that assign higher likelihood to long-range influences, this assumption may need to be reconsidered, but should be reasonable for large datasets. Veen and Schoenberg (2006) used this simplification with a power law decay in the excitation function without issue.

$$L_{CD}(\theta) \approx \sum_{t_j:u_j=0} \log \mu + \sum_{t_j:u_j \neq 0} \left(\log \beta + \log \alpha - \alpha(t_j - t_{u_j})\right) \\ -\mu T - \beta \sum_{t_i} 1 \tag{46}$$

Using this simplification, the derivative and estimate for β are now

$$\frac{dL_{CD}(\theta)}{d\beta} = \sum_{t_j:u_j \neq 0} \frac{1}{\beta} - \sum_{t_j} 1 \tag{47}$$

$$\hat{\beta}_{EM} = \frac{m_1}{N} \tag{48}$$

There are now a closed-form estimates for both exponentially decaying excitation

$$\frac{dL^{exp}_{CD}(\theta)}{d\alpha} = \sum_{t_j:u_j \neq 0} \left(\frac{1}{\alpha} - (t_j - t_{u_j})\right) \tag{49}$$

$$\hat{\alpha}_{EM} = m_1 \left(\sum_{t_j:u_j \neq 0} t_j - t_{u_j} \right)^{-1} \tag{50}$$

As well as Pareto excitation

$$\frac{dL^{\text{pareto}}_{CD}(\theta)}{d\omega} = \sum_{t_j:u_j \neq 0} \left(\frac{1}{\omega} + \log c - \log(t_j - t_{u_j})\right) \tag{51}$$

$$\hat{\omega}_{EM} = \frac{m_1}{\log(t_j - t_{u_j}) - \log c} \tag{52}$$

Both the exact and approximate estimates change only slightly for mutually-exciting point processes. With z_i indicating the process containing the event t_i and an intensity function defined as

$$\Lambda_l(t|H_t) = \mu_l + \sum_{t_j<t} \beta_{z_j l} \int_0^t \alpha_k e^{-\alpha_k(t-t_j)} \tag{53}$$

The update equations become

$$m_{0l} = \sum_{t_j:z_j=l,u_j=0} 1 \tag{54}$$

$$m_{kl} = \sum_{t_j:z_j=l,u_j=i\neq 0,z_i=k} 1 \tag{55}$$

$$\hat{\mu}_l = \frac{m_{0l}}{T} \tag{56}$$

$$\hat{\beta}_{kl} = \frac{m_{kl}}{N_k} \tag{57}$$

$$\hat{\alpha}_l = \sum_k m_{k1} \left(\sum_{t_j:z_j=l,u_j=i\neq 0} t_j - t_{u_j} \right)^{-1} \tag{58}$$

5.2 Expectation

Although the exact E-step for the exponentially decaying Hawkes process can be computed efficiently, in general the exact E-step has a computational cost of $O(N)$. This exact E-step considers the potential effect that each event might have on any and all future events. For contexts where each event has only a small number of likely ancestor events, it may be sufficient to allow for "forgetting" of the more distant past. For example, if a sequence covers more than a year of observations, it may be reasonable to consider only the most recent 6 months in computing the quantities $A_k(t_N)$, $B_k(t_N)$. This is similar to truncating the temporal decay distribution at some chosen point[1]. Because we are making strong parametric assumptions, we can choose a probabilistic threshold based on the most recent estimates for the parameters. The Pareto distribution has cumulative distribution function

$$P(x > t) = 1 - \left(\frac{c}{t}\right)^{\omega} \tag{59}$$

so for a cutoff probability of p we can forget, at time t, any historical information, at time t_i with a probability less than p of causing an event at time t. To be clear, we are only considering the probability that a parent at time t_i produces a child at time t and not the probability that the event at time t is a child of time t_i given the event t and the rest of the history.

$$P(x > t - t_i) \leq p \tag{60}$$

$$\Rightarrow t - t_i \geq c/p^{1/\omega} \tag{61}$$

This means that we can forget any event that occured more than $c/p^{1/\omega}$ units of time in the past. Intuitively, this means disregarding causal effects that are predicted by the model to have extremely low probabilities of occurring. This will likely improve performance but it somewhat defeats the purpose in choosing a temporal decay function that allows long-range influences if we choose to ignore some of the possible long-range influences. The extent to which this will improve performance will depend largely on the characteristics of the data being modeled. For very large numbers of streams (K) with extremely bursty behavior, it may still be necessary to consider a large number of observations as potential causal ancestors.

[1] This implies a temporal decay proportional to the truncated distribution, but as it no longer integrates to 1 it is not a valid probability distribution. This could be remedied by rescaling the delay function and updating the corresponding conditional maximum likelihood estimates.

Table 1 True values for the parameters of the excitation function

Parameter	Value
μ	0.01
α_{exp}	0.10
ω_{pow}	10/(10–0.1)
c_{pow}	0.10
ψ_{lnorm}	log 10
σ_{lnorm}	1

This will also bias the estimation procedure to some degree, although it may be possible to derive a modification to the update equations to correct for any bias introduced.

6 Results

We used simulated data to examine the bias and robustness of the proposed EM estimation, using three prototypical scenarios to assess the inference in between them. The specific scenarios of interest are symmetric excitation, asymmetric excitation, and no cross-process excitation. We selected excitation parameters in an attempt to produce *parent* → *child* inter-arrival times distinctly different from *immigrant* → *immigrant*, such that the expected time between a parent and child is one tenth of the expected time between immigrants(see Table 1). Next, we assume the self-excitement parameter, $\beta_{ii} = 0.5$, such that each event in a process can be always expected to cause 0.5 future events in that process, and where cross-process excitation exists, it is half as strong as self-excitation, $\beta_{i[j \neq i]} = 0.25$. In the symmetric scenario, both processes excite each other($\beta_{12} = \beta_{21} = 0.25$), whereas in the asymmetric case only one process excites the other($\beta_{12} = 0.25, \beta_{21} = 0.0$) and in the no excitement case neither excites the other($\beta_{12} = \beta_{21} = 0.0$), although both excite themselves($\beta_{11} = \beta_{22} = 0.5$). For each scenario, 100 instances of the two processes are sampled such that when combined they contain 100 events. Each inference strategy is applied on each sampled instance, running either to convergence[2] or 500 iterations.

The results for these simulations are shown in terms of relative error in Figs. 1, 2 and 3 displaying results for exponentially decaying excitement, Pareto excitement and Log-Normal excitement, respectively. Only graphics for the exact M-step are shown as in each case the approximate M-step produced estimates nearly identical to the exact M-step. We show relative error ([absolute error] / [true value]) here for ease of comparison across parameters, as several parameters have quite different absolute values. Although numerical optimization fails to converge in almost every scenario (see Table 2), it performs fairly well in certain cases, e.g. exponential symmetric and asymmetric excitation. However, it has has a severe lack of robustness in several of the scenarios tested. In particular, its estimates when no cross-stream excitation exists are extremely inconsistent to the point that they would seem to be of little practical utility. Although it may be possible to improve the performance in these specific scenarios using an $l1$-regularization, this is unlikely to be sufficient in general, as the numerical optimization performed nearly as poorly in the Log-Normal symmetric excitation and Pareto asymmetric excitation scenarios.

[2] Convergence was defined for the EM estimation as no change in the first three significant digits of any of the parameters. In the case of numerical optimization it was defined as an increase in the likelihood function of less than $1e - 6$.

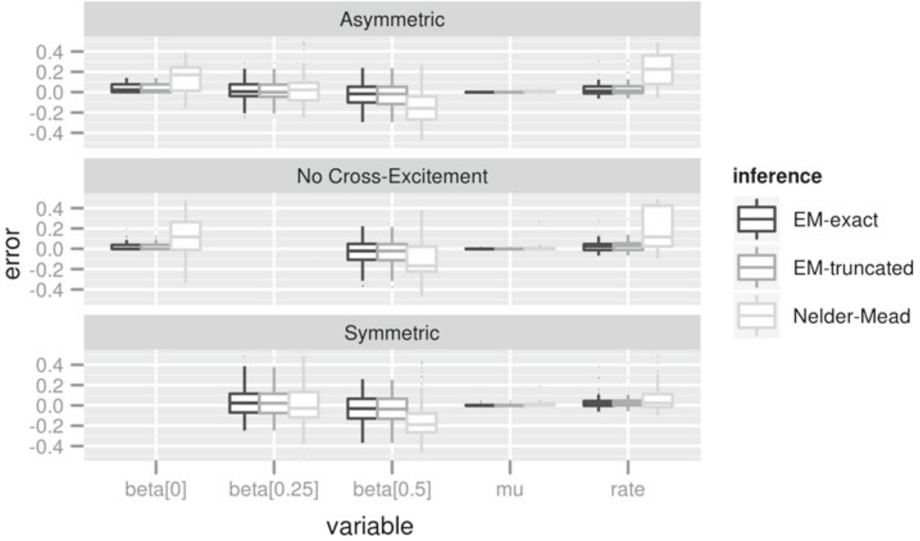

Fig. 1 Comparison of inference strategies for exponentially decaying excitation

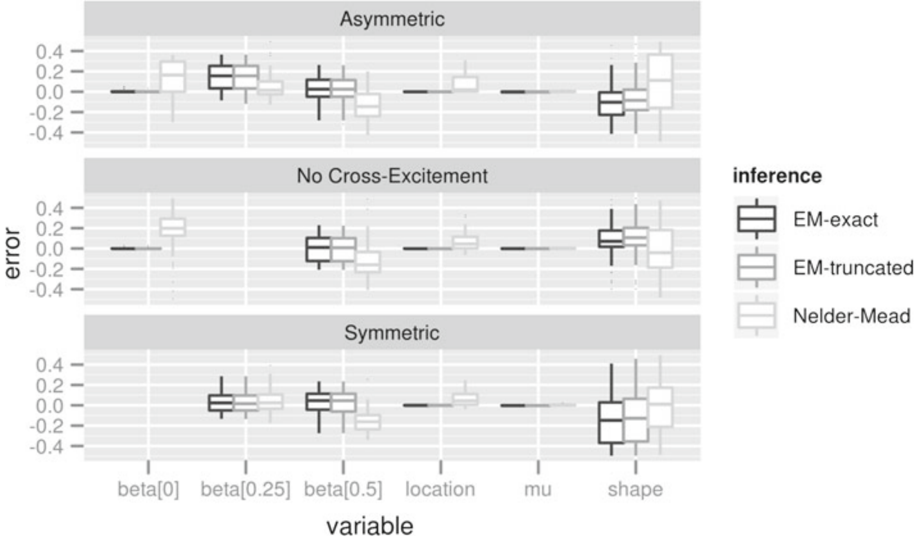

Fig. 2 Comparison of inference strategies for Pareto excitation

We also performed nonparametric hypothesis tests using the Wilcoxon rank sum. Paired two-sample comparisons indicate that across the range of parameters sampled, the EM estimates had significantly smaller error than the numerical optimization estimates($p < 0.001$) for the background rate, μ, and for the expected self/mutual excitation, β. The sole exception to this pattern was for estimation of the β parameter under log-normal distributed excitation. This is consistent with a visual inspection of Fig. 3, where the EM estimation appears to be consistently worse than Nelder–Mead numerical optimization. Although the EM estimates were significantly better than the numerical optimization estimates, they were still

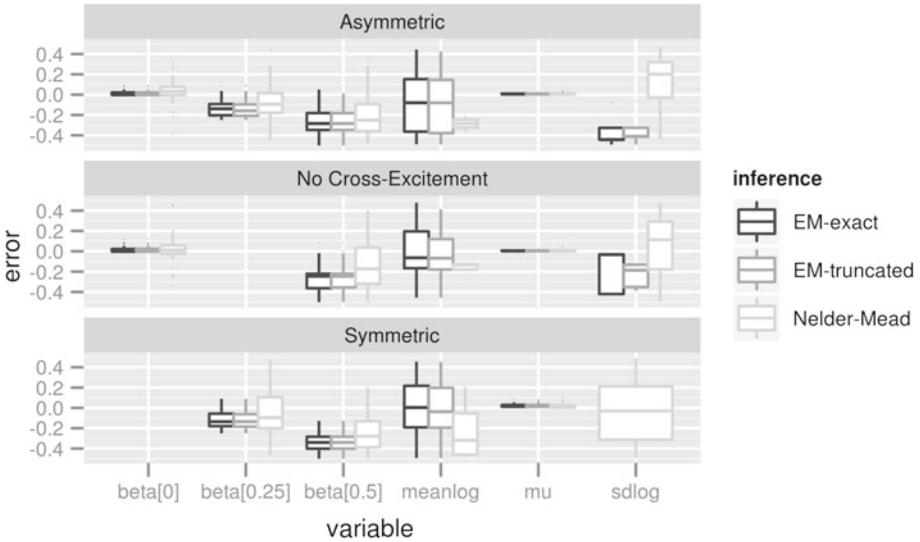

Fig. 3 Comparison of inference strategies for log-normal excitation

significantly biased. The only situations where we failed to reject the null hypothesis that the estimates are unbiased(mean error equals 0) are for the simple exponentially decaying excitation. Among the EM estimation algorithms, there were no significant differences either in the exact versus approximate M-step or the exact versus truncated E-step.

Table 2 shows the average running times and rates of convergence after 500 iterations for the different excitation functions and estimation procedures. The numerical optimization failed to converge in almost every instance it was tested with. Although it may seem that this alone is sufficient to explain the poor performance of the numerical optimization, the convergence rates across the different excitation functions are nearly identical despite wide variations in the accuracy of the estimates after 500 iterations. In addition to failing to converge or to provide a robust parameter estimate, numerical optimization was also extremely computationally demanding. Although both the numerical optimization and the EM estimation are iterative algorithms, the EM estimation terminated 10–100 times faster, likely due to its superior convergence rate.

Although the EM estimation performed uniformly superior to the numerical optimization, there are subtle variations among the various exact and approximate EM estimates. The exact E-step and the truncated E-step provide essentially equally good estimates, however inference using the truncated E-step converged faster in every case, sometimes beating the exact E-step by over a factor of 6. Differences between the exact and approximate M-steps were less substantial with similar accuracy and run-times except in the case of Pareto distributed excitation, where the approximate M-step ran 25–40 % faster.

7 Discussion

We compared the performance of the Expectation-Maximization(EM) strategies proposed to numerical optimization using the Nelder–Mead algorithm and found that our EM algorithms consistently provided significantly better estimates than numerical optimization. However,

Table 2 Average timing and convergence performance of the various inference strategies

	Time (s)	Convergence rate (%)
Exponential		
EM-exact	16.80	1.00
EM-truncated	3.16	1.00
Nelder–Mead	276.24	0.04
Exponential-approx		
EM-exact	17.69	1.00
EM-truncated	3.25	1.00
Nelder–Mead	277.44	0.04
Pareto		
EM-exact	65.10	0.40
EM-truncated	50.28	0.32
Nelder–Mead	1113.37	0.00
Pareto-approx		
EM-exact	49.57	1.00
EM-truncated	29.86	1.00
Nelder–Mead	1114.83	0.00
Log-normal		
EM-exact	30.13	1.00
EM-truncated	4.80	0.95
Nelder–Mead	274.57	0.00
Log-normal-approx		
EM-exact	30.82	1.00
EM-truncated	4.53	0.94
Nelder–Mead	274.25	0.00

our selection of 500 iterations as the cutoff for both the Nelder–Mead and EM algorithms may have unduly handicapped the Nelder–Mead numerical optimization. Increasing this cutoff may not be practical, though, as the Nelder–Mead algorithm still required as much as 20 min to complete its 500 iterations. The computational requirements of the Nelder–Mead algorithm are unlikely to decrease in moving from the single stream of 100 events to larger real-world datasets.

Although our results generally seem to recommend the use of the EM estimation strategies, they also highlighted an important weakness. Both the statistical bias and robustness, as well as the computational demands of the EM estimation vary quite dramatically with the specific excitation function. Although the EM estimates for both exponentially decaying and Pareto distributed excitation were quite good, estimates of the log-normal excitation systematically underestimated both self and cross-stream excitation (the β parameters). Similarly, while the truncated Expectation step provided large computational benefits in estimating either exponentially decaying or log-normal excitation, it provided almost no benefit under Pareto distributed excitation. This, at least, makes intuitive sense. Although the log-normal distribution decays more slowly than the exponential distribution, it does not permit nearly as many large time delays between *parent* and *child* events. These long-range influences are directly related to the amount of truncation and so it would seem that the more frequent long-range influences under the Pareto excitation limit the ability to rapidly truncate the recent history and “forget” the previous events.

The accuracy and efficiency of the EM estimation depend substantially on the specific excitation function, suggesting that before applying the proposed strategies, they should first be tested with the specific excitation function chosen for the application. However, the computational efficiency of EM estimation makes it the only real option for modeling large-scale phenomena with mutually exciting Hawkes processes.

8 Conclusions

Self and mutually-exciting Hawkes processes hold great potential for modeling a wide variety of phenomena, especially in human social systems. We generalize previous work on the ETAS model and propose a variety of exact and approximate estimation algorithms based on Expectation-Maximization. The exact inference is straightforward but may be computationally demanding for large datasets. However, simple approximations can be made both in the Expectation-step and the Maximization-step to reduce the computational burden. We use simulated data with known parameters to assess the statistical accuracy and computational efficiency of the EM estimates and find both to be superior to maximum likelihood estimates through numerical optimization. In particular, the truncated approximate E-step provides minimally biased estimates in addition to offering a tremendous reduction in computational cost even for the small datasets we simulated. The proposed EM algorithm is not limited to the specific model we use here but can be applied to a wide variety of excitation functions and parameterizations. Our maximum-likelihood EM estimation is accurate, robust and efficient, providing a practical way for using mutually-exciting Hawkes processes for continuous-time modeling of interacting streams of activity.

Acknowledgments This work was supported in part by the Office of Naval Research (N00014-06-1-0104) for adversarial assessment and (N00014-08-11186) for rapid ethnographic assessment, the Army Research Office and ERDC-TEC (W911NF0710317) and the National Science Foundation (#0943168). Additional support was provided by CASOS—the center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University.

References

Airoldi EM, Blei DM, Xing E, Fienberg SE (2005) A latent mixed membership model for relational data. In: http://portal.acm.org/citation.cfm?id=1134283 (ed) Proceedings of the international workshop on link discovery, ACM, pp 1–8

Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. Proc Natl Acad Sci 105(41):15649–15653

Deschatres F, Sornette D (2005) Dynamics of book sales: endogenous versus exogenous shocks in complex networks. Phys Rev E 72(1):16112

Egedsdal M, Fathauer C, Louie K, Neuman J (2010) Statistical and stochastic modeling of gang rivalries in Los Angeles. SIAM Undergrad Res Online 3(3):72–94

Guo F, Hanneke S, Xing EP (2007) Recovering temporally rewiring networks : a model-based approach. In: Proceedings of the 24th international conference on machine learning. Corvallis, OR

Hawkes AG (1971a) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58(1):83

Hawkes AG (1971b) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58(1):83

Hawkes AG, Oakes D (1974) A cluster process representation of a self-exciting process. J Appl Probab 11(3):493–503

Hoff PD (2008) Multiplicative latent factor models for description and prediction of social networks. Comput Math Organ Theory 15(4):261–272

Hoff PD, Raftery AE (2002) Latent space approaches to social network analysis. J Am Stat Assoc 97(460):1090–1098

Johansen A, Sornette D (2000) Download relaxation dynamics on the WWW following newspaper publication of URL. Phys A: Stat Mech Appl 276(1–2):338–345

Kolar M, Song L, Ahmed A, Xing EP (2009) Estimating time-varying networks. Ann Appl Stat 4(1):1–33

Krivitsky PN, Handcock MS (2008) Fitting latent position cluster models for social networks with latentnet. J Stat Softw 24(5):1–23

Large J (2007) Measuring the resiliency of an electronic limit order book. J Financial Mark 10(1):1–25

Møller J, Rasmussen J (2005) Perfect simulation of Hawkes processes. Adv Appl Probab 37(3):629–646

Ogata Y (1981) On Lewis' simulation method for point processes. IEEE Trans Inf Theory 27(1):23–31

Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. J Am Stat Assoc 83(401):9–27

Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Stat Math 50(2): 379–402

Ozaki T (1979) Maximum likelihood estimation of Hawkes' self-exciting point processes. Ann Inst Stat Math 31(1):145–155

Sarkar P, Moore A (2005) Dynamic social network analysis using latent space models. ACM SIGKDD Explor Newsl 7(2):40

Snijders TAB (1996) Stochastic actor-oriented models for network change. J Math Sociol 21:149–172

Snijders TAB, Nowicki K (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J Classif 14(1):75–100

Veen A, Schoenberg FP (2006) Estimation of space-time branching process models in seismology using an EM-Type algorithm. J Am Stat Assoc

Zhuang J, Ogata Y, Vere-Jones D (2002) Declustering of space-time earthquake occurrences. J Am Stat Assoc 97(458):369–380