Hawkes Graphs

Paul Embrechts and Matthias Kirchner

RiskLab, Department of Mathematics, ETH Zurich, Rämistrasse 101, 8092 Zurich, Switzerland

This paper introduces the Hawkes skeleton and the Hawkes graph. These notions summarize the branching structure of a multivariate Hawkes point process in a compact and fertile way. In particular, we explain how the graph view is useful for the specification and estimation of Hawkes models from large, multitype event streams. Based on earlier work, we give a nonparametric statistical procedure to estimate the Hawkes skeleton and the Hawkes graph from data. We show how the graph estimation may then be used for choosing and fitting parametric Hawkes models. Our method avoids the a priori assumptions on the model from a straighforward MLE-approach and it is numerically more flexible than the latter. A simulation study confirms that the presented procedure works as desired. We give special attention to computational issues in the implementation. This makes our results applicable to high-dimensional event-stream data, such as dozens of event streams and thousands of events per component.

1. Introduction

This paper discusses the specification and estimation of Hawkes point process models from large, multitype event-stream datasets such as neural spike-trains, internet search-queries, or limit-order-book data in high-frequency finance. Our approach uses the notion of a Hawkes skeleton and a Hawkes graph. These concepts turn out to be fertile beyond statistical estimation.

The Hawkes process was introduced in Hawkes (1971a,b) as a stationary point process on $\mathbb{R}$ whose points are assigned to a finite number of types. The (stochastic) intensity of a Hawkes process depends on the past of the process itself: given the occurrence of an event, the intensities—the expected mean number of events per time unit and event type—typically jump upwards and then decay. This structure can alternatively be represented as a multitype branching-process with immigration; see Hawkes (1974). The crucial parameters of a Hawkes model are the *excitement functions* or, emphasizing the branching interpretation, the *reproduction intensities* that govern these self- and crosseffects. For a textbook reference that covers many aspects of the Hawkes process; see Daley and Vere-Jones (2003). Maximum likelihood estimation of Hawkes processes has been treated in Ogata (1988) covering calibration issues and introducing a numerically beneficial recursive method for the exponential decay case. Liniger (2009) deals especially with the construction of the multivariate and marked case, whereas Bacry *et al.* (2012) introduces a nonparametric estimation method that is based on the covariance density estimate of the data.

In the present paper, we formally introduce the *Hawkes graph*. The Hawkes graph summarizes the branching structure of a multitype Hawkes point process as a directed graph with weighted vertices and edges. The vertices represent the possible event-types of the corresponding Hawkes process; an edge (i, j) denotes non-zero excitement from event-type i to event-type j. The vertex

This work was supported by RiskLab Zurich and the Swiss Finance Institute.

weights are the corresponding immigration intensities; the weight of an edge (i, j) is the expected number of type-j children that an type-i event generates. The *Hawkes skeleton* is the Hawkes graph disregarding the weights. The graphical description is a natural summary of the Hawkes process. The network view on Hawkes processes has been considered in Song *et al.* (2013), Hall and Willett (2014), Delattre *et al.* (2015), and Bacry *et al.* (2015). Many properties of multivariate Hawkes processes such as feedback, cascades, connectivity, or criticality are properties that are most naturally described via graph terminology. In this paper, we mainly use the graph structure as a tool for estimating high-dimensional Hawkes processes. Concerning Hawkes process estimation, we see three main problems with the standard parametric likelihood approach. First of all, it uses many unjustified assumptions on the shape of the excitement functions. Secondly, the distribution of the MLE-estimator is (in general) not known. In particular, the likelihood approach does not provide tests to decide whether excitement from one event type to another exists *at all*. Finally, there are numerical issues that make it difficult to apply MLE in a straightforward way with large, high-dimensional event-stream datasets.

Our approach leaves the choice of the excitement functions open to the very last. We apply an estimation procedure developed in Kirchner (2015a). This procedure is based on a limit-representation of the Hawkes process that we studied in Kirchner (2015b): we discretize the original process and interpret it as an autoregressive model of bin-counts. The latter is statistically estimated using conditional least-squares. Here, the asymptotic distribution of the resulting estimators can be obtained. Our procedure is numerically more robust than the standard MLE approach. However, for high-dimensional data our procedure also cannot be applied in a straightforward manner. In combination with the concept of a Hawkes skeleton and graph, we tackle the numerical difficulties by the following three-step algorithm:

(i) Given a large multitype event-stream dataseat, we first apply a specific testing scheme to decide whether there is *any* effect from a specific event type to any other event type. The test result yields the *Hawkes-skeleton estimate*. In this first step, we use a parameter allowing us to tune for a *very coarse discretization*; this keeps the computational complexity under control. Despite the resulting discretization error, this approach typically yields a *superset* of the true skeleton. Under the paradigm that the graph of the true underlying multivariate Hawkes model is typically sparse, this estimated superset is still sparse.

(ii) In a second step, we estimate the *Hawkes graph given the skeleton estimate*. The Hawkes graph *quantifies* the remaining excitement effects. The sparseness of the estimated Hawkes-skeleton from (i) reduces the complexity of the estimation problem considerably: there are only few excitements left to estimate and there are fewer "explanatory types" per event type, namely the estimated parent sets. Consequently, we may now choose a much finer discretization parameter and thus retrieve more precise edge and vertex weight estimates—including confidence intervals for all estimated values.

(iii) As a by-product, the calculations in (ii) yield estimates for the values of the non-zero excitement-functions on a finite equidistant grid. We exploit these estimation results graphically to choose appropriate parametric function-families. Finally, we fit the chosen parametric functions to the corresponding estimates by a non-linear least-squares method. This yields parameter estimates for parametric Hawkes models.

The multistep-procedure described above also works in a high-dimensional setting (such as dozens of event streams and thousands of events per component); the approach can be implemented in a straightforward way.

The paper is organized as follows: Section 2 gives definitions. In Section 3, we cite earlier results that allow for non-parametric estimation of Hawkes processes; we apply these methods to estimate the Hawkes skeleton, the Hawkes graph, and the remaining non-zero excitement functions. For illustration of the concepts involved, we present a simulation study in Section 4.

In Section 5, we discuss directions for further research.

2. Definitions

In this section we recall the branching construction of a multivariate Hawkes process and introduce the Hawkes skeleton as well as the Hawkes graph. The graph representation summarizes the branching structure of a Hawkes process in a compact and insightful manner.

2.1 *Multivariate Hawkes processes*

Throughout the paper, let $(\Omega, \mathbb{P}, \mathcal{F})$ be a complete probability space rich enough to carry all random variables involved. We give a constructive definition of the Hawkes process that emphasizes the branching view. For a similar construction; see Hawkes (1974) or Chapter 4 in Liniger (2009). The building blocks are Poisson random-measures on $\mathbb{R}$ endowed with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

Definition 2.1: Let $\lambda : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a locally integrable function. We say that M is a *Poisson random-measure* on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with *intensity function* λ whenever both of the following two conditions hold:

(i) If $A_1, A_2, \ldots, A_n \in \mathcal{B}(\mathbb{R})$ with $A_i \cap A_j = \emptyset$, then $M(A_1), M(A_2), \ldots, M(A_n)$ are mutually independent.
(ii) $M(A) \sim \text{Pois}\left(\int_A \lambda(t)\mathrm{d}t\right), \ A \in \mathcal{B}(\mathbb{R})$.

We write $M \sim \text{PRM}(\lambda \mathrm{d}t)$.

In the definition above we use the convention that $X \sim \text{Pois}\,(0) :\Leftrightarrow X \equiv 0$, a.s. and $X \sim \text{Pois}\,(\infty) :\Leftrightarrow X \equiv \infty$, a.s.

A multivariate or, synonymously, multitype Hawkes process is a model for the occurrence of events on $\mathbb{R}$, where the events are assigned to a finite number of types. The different event-types are represented as (in general dependent) random counting-measures. For each event type, there is an immigration process. Each immigrant event independently generates a family. These families consist of cascades of Poisson random measures. A Hawkes process is the superposition of all such families. We formalize this construction in the definitions below. To emphasize the intuition behind the names of immigrants, generations, and families, we use the somewhat unusual letters I, $\mathbf{G}$; and $\mathbf{F}$ for the corresponding processes.

Definition 2.2:

(i) For $(i,j) \in \{1,2,\ldots,d\}^2$, let $h_{i,j} : \mathbb{R} \to \mathbb{R}_{\geq 0}$, locally integrable, $h_{i,j}(t) \equiv 0$, $t \leq 0$, $\xi_t^{(i,j)}(\cdot) := \xi^{(i,j)}(\cdot - t) \sim \text{PRM}(h_{i,j}\mathrm{d}t)$, mutually independent over $(i,j,t) \in \{1,2,\ldots,d\}^2 \times \mathbb{R}$. We call the functions $h_{i,j}$ *reproduction intensities*.
(ii) For $i_0 = 1, 2, \ldots, d$, define the recursions

$$G_j^{(i_0,0)}(\cdot) := 1_{\{j=i_0,\, 0 \in \cdot\}}, \quad j = 1,2,\ldots,d, \tag{1}$$

$$G_j^{(i_0,g)}(\cdot) := \sum_{i=1}^{d} \int_{\mathbb{R}} \xi_t^{(i,j)}(\cdot) G_i^{(i_0,g-1)}(\mathrm{d}t) \quad j = 1,2,\ldots,d, \quad g \in \mathbb{N}\,. \tag{2}$$

For $g \in \mathbb{N}_0$, $\mathbf{G}^{(i_0,g)}(\cdot) := \left(G_1^{(i_0,g)}(\cdot), \ldots, G_d^{(i_0,g)}(\cdot)\right)^\top$ is the g-th *generation process* generated by a type-i_0 event at time zero.

(iii) For $i_0 = 1, 2, \ldots, d$, define

$$\mathbf{F}^{(i_0)}(\cdot) = \sum_{g \geq 0} \mathbf{G}^{(i_0,g)}(\cdot).$$

The random measure $\mathbf{F}^{(i_0)}$ is the *Hawkes i_0-family*, i.e., a family generated by a type-i_0 event at time 0.

The branching structure of the Hawkes family is encoded in recursion (2). Note that the Hawkes family is a particular example for the births in a *general branching process* as defined in Jagers (1975). The following definition clarifies how the Hawkes family process is related to the prototypic branching process, the Galton–Watson process:

Definition 2.3: For $i_0 = 1, 2, \ldots, d$, let $\mathbf{F}^{(i_0)}$ be a Hawkes i_0-family and let $\{\mathbf{G}^{(i_0,g)}\}_{g \in \mathbb{N}}$ be the corresponding generation processes constructed in Definition 2.2 above. For $g \in \mathbb{N}_0$, define

$$\mathbf{Y}_g^{(i_0)} := \left(Y_{g,1}^{(i_0)}, Y_{g,2}^{(i_0)}, \ldots, Y_{g,d}^{(i_0)}\right)^\top, \quad \text{where, for } j = 1, 2, \ldots, d, \quad Y_{g,j}^{(i_0)} := \int_{\mathbb{R}} G_j^{(g)}(t) \mathrm{d}t.$$

We call $\left(\mathbf{Y}_g^{(i_0)}\right)_{g \in \mathbb{N}_0}$ the *embedded generation process* of the Hawkes i_0-family $\mathbf{F}^{(i_0)}$.

The embedded generation processes are multitype Galton–Watson processes where each type-i individual has Pois($\int h_{i,j} \mathrm{d}t$) offspring of type j; see Section 2.3 in Haccou *et al.* (2005). This is why $a_{i,j} := \int_0^\infty h_{i,j}(t) \mathrm{d}t$, $1 \leq i, j \leq d$, are called *branching coefficients* and why the matrix $A := (a_{i,j}) \in \mathbb{R}_{\geq 0}$ is called *branching matrix*. If the spectral radius of the branching matrix is strictly smaller than 1, it then follows that, with probability 1, the total number of events in all Hawkes families will almost surely remain finite. In other words that, for $i_0 = 1, 2, \ldots, d$, the embedded generation processes die out in finite time.

Definition 2.4: For $i_0 = 1, 2, \ldots, d$, let $\eta_{i_0} \geq 0$ be *immigration intensities* and let $I_{i_0} \sim \mathrm{PRM}(\eta_{i_0} \mathrm{d}t)$, independent over i_0, be *Hawkes immigration processes*. Furthermore, let $\mathbf{F}_t^{(i_0)}(\cdot) := \mathbf{F}^{(i_0,t)}(\cdot - t)$, where $\mathbf{F}^{(i_0,t)}$, $t \in \mathbb{R}$, $i_0 = 1, 2, \ldots, d$, are independent copies of the generic Hawkes family processes $\mathbf{F}^{(i_0)}$ from Definition 2.2 above—also independent from the immigration processes. Set

$$\mathbf{N}(\cdot) := \big(N_1(\cdot), \ldots, N_d(\cdot)\big)^\top := \sum_{i_0=1}^{d} \int_{\mathbb{R}} \mathbf{F}_t^{(i_0)}(\cdot)\, I^{(i_0)}(\mathrm{d}t).$$

The random counting-measure $\mathbf{N}$ is a *d-variate or d-type Hawkes process*. If $N_i(\{T\}) = 1$, for some $i \in \{1, 2, \ldots, d\}$, we say that T is a *type-i event* or, synonymously, an *event in component i*. The Hawkes process $\mathbf{N}$ is *subcritical* if the corresponding embedded generation processes are subcritical, i.e., if the spectral radius of its branching matrix is strictly smaller than 1.

From Hawkes (1974) we have that in the subcritical case, a Hawkes process $\mathbf{N}$, constructed as in Definition 2.4 above, is a stationary solution to the system of implicit equations

$$\begin{aligned}\Lambda_j(t) &:= \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{E}\left[\mathbf{N}_j\big((t, t+\delta]\big) \Big| \sigma\Big(\mathbf{N}\left((a, b]\right),\, a < b \leq t\Big)\right] \\ &= \eta_j + \sum_{i=1}^{d} \int_{-\infty}^{t} h_{i,j}(t-s) N^{(i)}(\mathrm{d}s), \quad j = 1, 2, \ldots, d. \end{aligned} \tag{3}$$

We call $\mathbf{\Lambda}(t) := (\Lambda_j(\cdot), \ldots, \Lambda_j(\cdot))^\top$ the *conditional intensity* of $\mathbf{N}$. In terms of intensities, the value of a reproduction function at time t, $h_{i,j}(t)$, denotes the effect of an event $T^{(i)}$ in component i on the intensity of component j at time $T^{(i)} + t$.

Remark 1: In most work on Hawkes processes, including the original introduction (Hawkes 1974) and also including our earlier work (Kirchner 2015a,b), the function $h_{i,j}$ models the excitement *from component j on component i*. This somewhat counter-intuitive notation stems from the linear algebra used when writing (3) with matrix multiplication. In the present graph-driven work, "$h_{i,j}$", "$a_{i,j}$" and "$(i,j) \in \mathcal{E}$" *all refer to the effect from component i on component j.*

2.2 *Hawkes skeleton and Hawkes graph*

We interpret the branching structure of the Hawkes process in terms of "causality". The overall goal of causality research is to describe dependencies in a directed manner—rather than applying commutative concepts such as correlation or dependence; see Pearl (2009) for a recent overview. The notion of causality is subtle. For Hawkes processes, however, the use of the term seems justified; in the context of event streams, things cannot become much more "causal" than in the recurrent parent/children relation of a branching process: if we delete an event in the branching construction from the definitions in Section 2.1 above, its offspring vanishes. So—without discussing causality formally—we postulate that given an event in component i, it directly *causes* $\text{Pois}(a_{i,j})$ new events in component j. This makes the branching coefficient $a_{i,j}$ an obvious measure for the strength of the causal effect from component i to component j. Such causal effects are often represented as directed graphs. In the literature on causality, a graphical approach for modeling the interdependence of event streams for instance can be found in Meek (2014) or Gunawardana *et al.* (2014)—without any mentioning of "Hawkes". This shows how natural the definition of a Hawkes graph is. First, we introduce some general graph terminology:

Definition 2.5: A *graph* $\mathcal{G}$ is a 2-tuple $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \ldots, d\}$ is a set of *vertices* and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ a set of edges. Vertex i is a *parent* of vertex j if $(i,j) \in \mathcal{E}$; we write $\text{PA}(j) = \{i : (i,j) \in \mathcal{E}\}$. Vertex j is a *source vertex* if $\text{PA}(j) = \emptyset$. For $n \geq 1$, $(k_0, k_1, \ldots, k_n) \in \mathcal{V}^{n+1}$ is a *walk* of length n from vertex i to vertex j if $k_0 = i, k_n = j$ and $(k_{l-1}, k_l) \in \mathcal{E}$, $l = 1, 2, \ldots, n$; $(k_0, k_1, \ldots, k_n) \in \mathcal{V}^{n+1}$ is a *closed walk* starting in i if it is a walk from i to i. Vertex i is an *ancestor* of j if there exists a walk from i to j. We denote the sets of ancestors of a vertex i by $\text{AN}(i)$. The vertices i and j are *connected* if there exists a set $\{(k_{l-1}, k_l), l = 1, \ldots, n : \ k_0 = i, k_n = j, (k_l, k_{l-1}) \in \mathcal{E}$ **or** $(k_{l-1}, k_l) \in \mathcal{E}\}$ for some $n \in \{1, \ldots, d\}$. A graph is *strongly connected* if all pairs of its vertices are connected. A graph is *fully connected* if $(i,j) \in \mathcal{E}$, $(i,j) \in \{1, 2, \ldots, d\}^2$.

Note that in our definition, a graph allows cycles and, in particular, loops. Also note that a vertex may or may not be an ancestor and, in particular, a parent of itself.

Definition 2.6: Let $\mathbf{N}$ be a d-variate Hawkes process with immigration intensities $\eta_1, \eta_2, \ldots, \eta_d$, reproduction intensities $h_{i,j}$, and branching coefficients $a_{i,j} (= \int h_{i,j}(t) \mathrm{d}t)$; see Definitions 2.2 and 2.4. The *Hawkes graph skeleton* $\mathcal{G}^*_{\mathbf{N}} = (\mathcal{V}^*_{\mathbf{N}}, \mathcal{E}^*_{\mathbf{N}})$ of $\mathbf{N}$ consists of a set of vertices $\mathcal{V}^*_{\mathbf{N}} = \{1, 2, \ldots, d\}$ and a set of edges

$$\mathcal{E}^*_{\mathbf{N}} := \Big\{(i,j) \in \mathcal{V}^*_{\mathbf{N}} \times \mathcal{V}^*_{\mathbf{N}} : \ a_{i,j} > 0\Big\}.$$

For $j = 1, 2, \ldots, d$, we denote the parent and ancestor sets with respect to the Hawkes skeleton by $\text{PA}_{\mathbf{N}}(j)$ and $\text{AN}_{\mathbf{N}}(j)$. For the *Hawkes graph* $\mathcal{G}_{\mathbf{N}} = (\mathcal{V}_{\mathbf{N}}, \mathcal{E}_{\mathbf{N}})$ of $\mathbf{N}$, each vertex and each edge

of the corresponding skeleton is supplied with a weight:

$$\mathcal{V}_{\mathbf{N}} := \Big\{(i;\eta_i): \quad i \in \mathcal{V}_{\mathbf{N}}^* \text{ and } \eta_i \text{ is the } i\text{-th immigration intensity of } \mathbf{N}\Big\},$$
$$\mathcal{E}_{\mathbf{N}} := \Big\{(i,j;a_{i,j}): \quad (i,j) \in \mathcal{E}_{\mathbf{N}}^* \text{ and } a_{i,j} = \int h_{i,j}(t)\mathrm{d}t\Big\}.$$

We call the branching matrix $A = (a_{i,j}) \in \mathbb{R}_{\geq 0}^{d\times d}$ of $\mathbf{N}$ the *adjacency matrix* of $\mathcal{G}_{\mathbf{N}}$.

(i) A Hawkes graph is *strongly, respectively, fully connected* if the corresponding skeleton is strongly, respectively, fully connected; see Definition 2.5.

(ii) Vertex j of a Hawkes graph is a *redundant vertex* if $\eta_j = 0$ and, in addition, $\eta_i = 0$ for all $i \in \mathrm{AN}_{\mathbf{N}}(j)$.

(iii) A Hawkes graph is *subcritical* if

$$\sum_{g=1}^{\infty} \sum_{w_g^{(i)} \in \mathcal{W}_g^{(i)}} |w_g^{(i)}| < \infty, \quad i = 1,2,\ldots,d, \tag{4}$$

where $\mathcal{W}_g^{(i)}$ denotes the set of closed walks starting in i of length g in $\mathcal{G}_{\mathbf{N}}^*$ and, for any walk in a Hawkes graph, $w = (i_0, i_1, \ldots, i_g)$, we define the *walk weights* $|w| := \prod_{l=1}^{g} a_{i_{l-1},i_l}$ with a_{i_{l-1},i_l} the corresponding edge-weights.

As a side remark, note that the term *Hawkes graph* already has been used for the graph representation of a specific finite group that has been introduced in Hawkes (1968). Neither the author of the latter paper, *T.* Hawkes, nor its content has anything to do with our notion of a Hawkes graph.

Obviously, the Hawkes graph does not fully specify the corresponding Hawkes process; it only captures the structure of the embedded generation processes from Definition 2.3 together with the immigration intensities. Despite this simplification, the Hawkes graph gives relevant insight for the underlying Hawkes process. For example, connectivity and redundancy of nodes are two graph-based concepts that become increasingly important the higher the dimension of the considered model is. If a Hawkes graph is disconnected, we may consider the *strongly connected* subgraphs separately and correspondingly use separate, lower-dimensional Hawkes processes. The notion of *redundant vertices* is important because typically we only want to consider "accessible" event types. If the Hawkes graph of some Hawkes process has redundant vertices, the model is typically misspecified. Furthermore, one can show that the criterion for *subcritical* Hawkes graphs in (4) corresponds to the underlying Hawkes process being subcritical, that is, the spectral radius of the branching matrix being strictly less than 1. However, the criterion in (4) for the Hawkes graph gives a more concrete meaning to the somewhat abstract eigenvalue-based criterion for the Hawkes process. It can be used when constructing Hawkes graphs, respectively, models. And—if the graph is sparse and the closed walks are not too numerous—one can check subcriticality without even calculating any eigenvalue; see Section 4.1. Also the walk weights "$|w|$" themselves might be worth calculating in some applications: for any walk in a Hawkes graph $w = (i_0, i_1, \ldots, i_g)$, we have that

$$|w| = \prod_{l=1}^{g} a_{i_{l-1},i_l} = \mathbb{E}\, Y_{g,i_g}^{(i_0)}, \quad i = 1,2,\ldots,d, \tag{5}$$

where $(\mathbf{Y}_g^{(i_0)}) = (Y_{g,1}^{(i_0)}, Y_{g,2}^{(i_0)}, \ldots, Y_{g,d}^{(i_0)})$ is the embedded generation process from Definition 2.3. Last but not least, the graph structure obviously allows for attractive self-explaining illustrations; see Figures 1 and 2. In the following proposition, we collect some specific statistical information

that may be calculated from the adjacency matrix of a Hawkes graph:

Proposition 2.7: *Let $A = (a_{i,j})$ be the adjacency matrix of a subcritical Hawkes graph with $|\mathcal{V}_{\mathbf{N}}| \geq 2$. Then we have that*

(i) $a_{i,j}$ *is the expected number of type-j events stemming (directly) from a type-i event;*
(ii) $a_{i,j} > 0 \quad \Leftrightarrow \quad i \in \mathrm{PA}(j)$*;*
(iii) $a_{i,j} = 0,\ j = 1, 2, \ldots, d \quad \Leftrightarrow \quad$ *vertex i is a sink vertex;*
(iv) $a_{i,j} = 0,\ i = 1, 2, \ldots, d\ a_{i,j} = 0,\ i = 1, 2, \ldots, d \quad \Leftrightarrow \quad$ *vertex j is a source vertex;*
(v) $(A^n)_{i,j} > 0 \quad \Leftrightarrow \quad$ *there is a walk of length n from i to j;*
(vi) $(A^n)_{i,j} > 0$ *for some* $n \in \{1, 2, \ldots, d-1\} \quad \Leftrightarrow \quad i \in \mathrm{AN}(j)$*;*
(vii) $a_{i,j} > 0,\ i, j = 1, 2, \ldots, d \quad \Leftrightarrow \quad$ *the Hawkes graph is fully connected*
(viii) A *is irreducible* $\quad \Leftrightarrow \quad$ *the Hawkes graph is strongly connected;*
(ix) $(1_{d\times d} - A)$ *is invertible, the limit* $\mathbb{R}_{\geq 0}^{d\times d} \ni (e_{i,j}) := \lim_{n\to\infty} \sum_{k=1}^{n} A^k$ *exists and is equal to* $A(1_{d\times d} - A)^{-1}$*;*
(x) the value $e_{i,j}$ is the expected total number of offspring in the j-th component of a Hawkes i-family (without counting the initial event) or, equivalently, with the notation from Definition 2.3, $e_{i,j} = \mathbb{E} \sum_{g=1}^{\infty} Y_{g,j}^{(i)}$*;*
(xi) for $i = 1, 2, \ldots, d$*,* $\tilde{c}_i := \sum_{j=1}^{d} e_{i,j}$ *is the expected total number of events in a Hawkes i-family (again without counting initial event) or, equivalently,* $\tilde{c}_i = \mathbb{E} \sum_{j=1}^{d} \sum_{g=1}^{\infty} Y_{g,j}^{(i)}$*.*

The properties above can easily be checked. They may help to describe the role and the relationships between Hawkes process components, respectively, Hawkes graph vertices. Two specific $\mathbb{R}_{\geq 0}^{d}$-vectors might be particularly meaningful statistical summaries of a Hawkes graph:

Definition 2.8: Let $\mathbf{N}$ be a d-variate Hawkes process. Furthermore, let A be the adjacency matrix of the corresponding Hawkes graph. Consider $\mathbb{R}_{\geq 0}^{d\times d} \ni (e_{i,j}) = A(1_{d\times d} - A)^{-1}$ and $\tilde{c}_i = \sum_{j=1}^{d} e_{i,j}$ from Proposition 2.7. For $i, j = 1, 2, \ldots, d$, define

$$c_i := \tilde{c}_i / \sum_{j=1}^{d} \tilde{c}_j \quad \text{and} \quad f_j := e_{j,j} / \sum_{i=1}^{d} e_{i,j}.$$

We call $(c_i)_{i=1,\ldots,d}$ the *cascade coefficients* and $(f_j)_{j=1,\ldots,d}$ the *feedback coefficients.*

One way of tuning a specific Hawkes graph can be achieved by deleting all outgoing edges of selected vertices. The coefficients defined above summarize the effect of such a deletion. In view of Proposition 2.7, we have the following interpretations: the *cascade coefficients* are important from a *systemic* point of view. They allow to compare the d vertices with respect to their total effect on the system. Which type of events trigger large cascades? Values larger than $1/d$ indicate a relatively large impact, values smaller than $1/d$ a relatively small impact. The *feedback coefficients* are more important from an *individual* point of view. They indicate how much of the total reproduction excitement (again disregarding immigration intensities) that each vertex *experiences* is due to closed walks, i.e., feedback. We illustrate both concepts in Section 4.1.

3. Estimation

In this section we give a summary of earlier work, where we introduced a nonparametric estimation procedure for the multivariate Hawkes process. Based on this approach, we introduce an estimation procedure for the Hawkes skeleton and the Hawkes graph. In particular, we clarify how one can bypass numerical problems in high-dimensional settings. Finally, we explain how one can use the results for completely specifying and estimating a parametric Hawkes model.

3.1 *Earlier results*

In earlier work, we showed that the distributions of the bin-count sequences of a Hawkes process can be approximated by the distribution of so called *integer-valued autoregressive time series* INAR(p); see Kirchner (2015b). This approximation yields an estimation method for the Hawkes process: we fit the approximating model on observed bin-counts of point process data. The resulting estimates can be used as estimates of the Hawkes reproduction intensities on a finite and equidistant grid; see Kirchner (2015a). For illustration, consider a univariate Hawkes process N with reproduction intensity h and immigration intensity η. Given data from N in a time window $(0,T]$, $\Delta > 0$, small, bin counts $X_n^{(\Delta)} := N\big(((n-1)\Delta, n\Delta]\big)$, $k = 1, 2, \ldots, n := N\big((0,T]\big)$, and some $p \in \mathbb{N}$, large, we calculate

$$\left(\hat{\alpha}_0^{(\Delta)}, \hat{\alpha}_1^{(\Delta)}, \ldots, \hat{\alpha}_p^{(\Delta)}\right) := \operatorname{argmin}_{(\alpha_0^{(\Delta)}, \alpha_1^{(\Delta)}, \ldots, \alpha_p^{(\Delta)})} \sum_{k=p+1}^{n} \left(X_k^{(\Delta)} - \alpha_0^{(\Delta)} - \sum_{l=1}^{p} \alpha_l^{(\Delta)} X_{k-l}^{(\Delta)} \right)^2 . \quad (6)$$

Given (6), we estimate the reproduction-intensity values $h(k\Delta)$, $k = 1, 2, \ldots, p$, of N by $\hat{h}_k := \hat{\alpha}_k^{(\Delta)}/\Delta$ and the immigration intensity η by $\hat{\eta} := \hat{\alpha}_0^{(\Delta)}/\Delta$. The multivariate case is conceptually equivalent but somewhat cumbersome notationwise. Furthermore—due to the special distribution of the errors—the covariance matrix of the estimates is nonstandard. This is why we give all formulas in some detail. The following definitions and properties are taken from Kirchner (2015a)—modulo transposition as stated in Remark **1**.

Definition 3.1: Let $\mathbf{N} = (N_1, N_2, \ldots, N_d)^\top$ be a subcritical d-variate Hawkes process with immigration intensity $\eta \in \mathbb{R}_{\geq 0}^d \setminus \{0_d\}$ and reproduction intensities $h_{i,j} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $(i,j) \in \{1, 2, \ldots, d\}^2$. Let $T > 0$ and consider a sample of the process on the time interval $(0,T]$. For some $\Delta > 0$, construct the $\mathbb{N}_0^d$-valued *bin-count sequence* from this sample:

$$\mathbf{X}_k^{(\Delta)} := \left(N_j\Big(\big((k-1)\Delta, k\Delta\big]\Big)\right)_{j=1,\ldots,d}^{\top}, \quad k = 1, 2, \ldots, n := \lfloor T/\Delta \rfloor . \quad (7)$$

Define the *multivariate Hawkes estimator* with respect to some support s, $\Delta < s < T$,

$$\widehat{\mathbf{H}}^{(\Delta,s)} := \frac{1}{\Delta} \left(\mathbf{Z}^\top \mathbf{Z} \right)^{-1} \mathbf{Z}^\top \mathbf{Y} \quad \in \mathbb{R}^{(dp+1)\times d} . \quad (8)$$

Here,

$$\mathbf{Z}\left(\mathbf{X}_1^{(\Delta)}, \ldots, \mathbf{X}_n^{(\Delta)}\right) := \begin{pmatrix} (\mathbf{X}_p^{(\Delta)})^\top & (\mathbf{X}_{p-1}^{(\Delta)})^\top & \ldots & (\mathbf{X}_1^{(\Delta)})^\top & 1 \\ (\mathbf{X}_{p+1}^{(\Delta)})^\top & (\mathbf{X}_p^{(\Delta)})^\top & \ldots & (\mathbf{X}_2^{(\Delta)})^\top & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (\mathbf{X}_{n-1}^{(\Delta)})^\top & (\mathbf{X}_{n-2}^{(\Delta)})^\top & \ldots & (\mathbf{X}_{n-p}^{(\Delta)})^\top & 1 \end{pmatrix} \in \mathbb{R}^{(n-p)\times(dp+1)} \quad (9)$$

is the *design matrix* and $\mathbf{Y}\,(\mathbf{x}_1, \ldots, \mathbf{x}_n) := \left(\mathbf{X}_{p+1}^{(\Delta)}, \mathbf{X}_{p+2}^{(\Delta)}, \ldots, \mathbf{X}_n^{(\Delta)}\right)^\top \in \mathbb{R}^{(n-p)\times d}$.

For the following considerations, we drop the "(Δ, s)" superscript. Note that also the matrices $\mathbf{Z}$ and $\mathbf{Y}$ depend on Δ. Additional notation clarifies what the entries of the matrix $\widehat{\mathbf{H}}$ in (8)

actually estimate:

$$\begin{pmatrix} \widehat{H}_1 \\ \dots \\ \widehat{H}_p \\ \hat{\eta} \end{pmatrix} := \widehat{\mathbf{H}} \in \mathbb{R}^{(dp+1)\times d}, \quad \text{where} \quad \widehat{H}_k := \begin{pmatrix} \hat{h}_{1,1}(k\Delta) & \hat{h}_{1,2}(k\Delta) & \dots & \hat{h}_{1,d}(k\Delta) \\ \hat{h}_{2,1}(k\Delta) & \hat{h}_{2,2}(k\Delta) & \dots & \hat{h}_{2,d}(k\Delta) \\ \dots & \dots & \dots & \dots \\ \hat{h}_{d,1}(k\Delta) & \hat{h}_{d,2}(k\Delta) & \dots & \hat{h}_{d,d}(k\Delta) \end{pmatrix}. \tag{10}$$

In Kirchner (2015a), we find that, for large T, small Δ and large p, the entries of $\widehat{\mathbf{H}}$ are approximately jointly normally distributed around the true values. Furthermore, the covariance matrix of $\mathrm{vec}\left(\widehat{\mathbf{H}}^\top\right) \in \mathbb{R}^{d(dp+1)}$ can be consistently estimated by

$$\widehat{S^2} := \frac{1}{\Delta^2}\left(\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \otimes 1_{d\times d}\right) \mathbf{W} \left(\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \otimes 1_{d\times d}\right) \in \mathbb{R}^{d(dp+1)\times d(dp+1)}. \tag{11}$$

Here, $\mathbf{Z}$ is the design matrix from (9) and $\mathbf{W} := \sum_{k=p+1}^{n} \mathbf{w}_k \, \mathbf{w}_k^\top \in \mathbb{R}^{d(dp+1)\times d(dp+1)}$, where, for $k = p+1, p+2, \dots, n$,

$$\begin{aligned} \mathbf{w}_k := &\left(\left(\left(\mathbf{X}_{k-1}^{(\Delta)}\right)^\top, \left(\mathbf{X}_{k-2}^{(\Delta)}\right)^\top, \dots, \left(\mathbf{X}_{k-p}^{(\Delta)}\right)^\top, 1\right)^\top \otimes 1_{d\times d}\right) \\ &\cdot \left(\mathbf{X}_k^{(\Delta)} - \Delta\hat{\eta} - \sum_{l=1}^{p} \Delta \widehat{H}_l^\top \, \mathbf{X}_{k-l}^{(\Delta)}\right) \in \mathbb{R}^{d(dp+1)\times 1}. \end{aligned} \tag{12}$$

Above, we consider $\mathrm{vec}(\mathbf{H}^\top)$ instead of $\mathrm{vec}(\mathbf{H})$ in order to apply the results from Kirchner (2015b) more directly; see Remark **1**. We will discuss below how one retrieves specific values from the covariance matrix estimation in (11). The estimator from Definition 3.1 above depends on a support s, $0 < s << T$ and a bin size Δ, $0 < \Delta \leq s$. Automatic methods for the choice of these estimation parameters are discussed in Kirchner (2015b). In the context of the present paper, we assume s fixed. Often an upper bound for the support of the reproduction functions can be guessed from the context of the data. The choice of Δ however will be crucial in the high-dimensional context. We will use it as a tuning parameter for controlling numerical complexity.

3.2 *Estimation of the Hawkes skeleton*

Our first goal is to identify the edges of the Hawkes skeleton from data; see Definition 2.6. The idea is simple: for $(i,j) \in \{1,2,\dots,d\}^2$, we estimate the edge weight $a_{i,j} = \int h_{i,j}(t)\mathrm{d}t$ by $\hat{a}_{i,j} := \Delta \sum_{k=1}^{p} \hat{h}_{i,j}(k\Delta)$; see (10) for the notation. Calculating the covariance estimate (11), we can check whether $\hat{a}_{i,j}$ is significantly larger than zero. If this is the case, we set $(i,j) \in \widehat{\mathcal{E}}^*$. In order to ease implementation, we explicitly give the necessary transformations for the estimates from Definition 3.1 and discuss numerical issues.

Definition 3.2: Given d-variate event stream data on $(0,T]$, calculate the Hawkes estimator $\mathbf{H}^{(\Delta_{\mathrm{skel}},s)}$ from Definition 3.1 with respect to some s, $0 < s < T$ and some Δ_{skel}, $0 < \Delta_{\mathrm{skel}} \leq s$. For $j = 1,2,\dots,d$, let $b_j \in \{0,1\}^{(dp+1)\times 1}$ be column vectors with all entries zero but 1s at entries $(k-1)d + j$, $k = 1,2,\dots,p = \lfloor s/\Delta_{\mathrm{skel}} \rfloor$. Let $B := (b_1, b_2, \dots, b_d)^\top$, and calculate

$$(\hat{a}_{i,j})_{1\leq i,j\leq d} = \Delta_{\mathrm{skel}} B \, \mathbf{H}^{(\Delta_{\mathrm{skel}},s)}. \tag{13}$$

Fix $\alpha_{\text{skel}} \in (0,1)$ and define the *Hawkes-skeleton estimator* as a graph $\widehat{\mathcal{G}}^* := (\{1,2,\dots,d\}, \widehat{\mathcal{E}}^*)$, with

$$\widehat{\mathcal{E}}^* := \Big\{(i,j) \in \{1,2,\dots,d\}^2 : \ \hat{a}_{i,j} > \hat{\sigma}_{i,j} z^{-1}_{1-\alpha_{\text{skel}}}\Big\}. \tag{14}$$

Here, for $\beta \in (0,1)$, z_β^{-1} denotes the β-quantile of a standard normal distribution. Efficient calculation of $(\hat{\sigma}_{i,j})_{1\le i,j\le d}$ will be given in Algorithm **1** below.

The main point of this first estimation step is that we hope that $|\mathcal{E}^*|$ and, consequently $|\widehat{\mathcal{E}}^*|$are typically much smaller than d^2, respectively, that $\text{PA}_{\mathcal{G}^*_{\mathbf{N}}}(j)$ and, consequently, $\widehat{\text{PA}}_{\mathcal{G}^*_{\mathbf{N}}}(j)$ are typically much smaller than d. If this is the case, the knowledge of the skeleton simplifies the estimation of the Hawkes graph considerably:

The role of $\boldsymbol{\Delta_{\text{skel}}}$ On one hand, the smaller we choose the bin size, the better the discrete approximation described in Section 3.1 works. On the other hand, the matrices involved in the calculation of the Hawkes estimator from Definition 3.1 become increasingly large when Δ decreases. More specifically, (8) involves the construction and multiplication of matrices with about ds/Δ rows and about T/Δ columns. Furthermore, we have to invert matrices of size $\lceil ds/\Delta\rceil \times \lceil ds/\Delta\rceil$. The crucial observation is that in the Hawkes-skeleton estimation, we may choose Δ_{skel} quite large for two reasons:

(i) The probability of *missing* a true edge in the skeleton estimation is very small—even if we choose Δ_{skel} as large as $\Delta_{\text{skel}} = s$. In other words, for $j = 1,2,\dots,d$, we have that $\text{PA}_{\mathbf{N}}(j) \subset \widehat{\text{PA}}_{\mathbf{N}}(j)$ with high probability. (Given reasonable values for α_{skel}, say $\alpha_{\text{skel}} \ge 0.01$).

(ii) The actual *quantitative* estimation of the interactions between different event types will be performed in a second step when we consider the Hawkes *graph*. In this second step, due to the sparseness of the Hawkes skeleton, we are typically able to choose a much finer bin-size Δ_{graph}. So we may ignore the bias stemming from too rough discretization in the first step.

By choosing $\Delta_{\text{skel}} = s/k$ for some small $k \in \mathbb{N}$ in the calculations of Definition 3.2 above, even Hawkes-skeleton estimates of very high-dimensional models (such as $d > 20$) become computationally tractable.

The role of $\boldsymbol{\alpha_{\text{skel}}}$ A value of $\alpha_{\text{skel}} = 1$ will yield a fully connected estimated graph as Hawkes skeleton. When α_{skel} decreases, the skeleton estimate becomes sparser and sparser. However, for $\alpha_{\text{skel}} \ge 0.01$, we typically still *overestimate* the true edge set. Under $H_0 : a_{i,j} \equiv 0$, we have that $\mathbb{P}_{H_0}[\hat{a}_{i,j} > \sigma^2_{i,j} z^{-1}_{1-\alpha_{\text{skel}}}] \approx \alpha_{\text{skel}}$. Because of the multiple testing setup over (i,j), the dependency between the tests, and the typically very rough discretization, the parameter $\alpha_{\text{skel}} \in (0,1)$ should not so much be thought of as an actual confidence level. It is more a flexible tuning parameter that allows for controlling the degree of sparseness in the estimated graph. Despite this warning, note that in the simulation study from Section 4.2 the corresponding empirical rates are very close to the different choices of α_{skel}.

Variance estimate calculation The most elaborate step from a computational point of view in Definition 3.1 is the calculation of the covariance estimator in (11). Here, we deal with matrices of size $\lceil d^2 s/\Delta\rceil \times \lceil d^2 s/\Delta\rceil$. Furthermore, we have to calculate approximately T/Δ vectors of size $d^2 s/\Delta$ and calculate and sum their crossproducts $\mathbf{w}_k\, \mathbf{w}_k^\top$. This is the numerical bottleneck of the procedure—in particular for high-dimensional setups. For the Hawkes skeleton estimator from Definition 3.2, we simplify the calculation. First of all, we note that in the matrix $\widehat{S}^2$ from (11),

we estimate many more covariance values than we actually need for the (marginal) distribution of the edge-weight estimates. After some linear algebra, we find that one can avoid the tedious computation of the $\mathbf{W}$ matrix from (11) by the following matrix manipulations.

Algorithm 1: Let $\mathbf{E} \in \{0,1\}^{d^2 \times (d^2 p + d)}$ be a matrix with all entries zero but, for $i, j = 1, 2, \ldots, d$, in row $(i-1)d + j$ we have 1s at entries $(k-1)d^2 + (i-1)d + j$, $k = 1, 2, \ldots, d$. Let $\mathbf{E}_{l,\cdot}$ denote the l-th row of $\mathbf{E}$. With $\widehat{S}^2$ from (11) and for $(i,j) \in \{1, 2, \ldots, d\}^2$, we have that $\hat{\sigma}^2_{i,j} := \Delta^2 \, \mathbf{E}^\top_{(i-1)d+j,\cdot} \, \widehat{S}^2 \, \mathbf{E}_{(i-1)d+j,\cdot}$ are the variance estimates for the $\hat{a}_{i,j}$ *from* (13). These estimates can be computed in the following way:

(i) Compute $E\left((\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top\right) \otimes 1_{d \times d}) \in \mathbb{R}^{d^2 \times d(n-p)}$ and stack the rows of the result in a vector. Row-wise fill this vector in a $d^2(n-p) \times d$ matrix $\mathbf{C}$.

(ii) Set $\mathbf{U} = (\mathbf{Y} - \Delta \, \mathbf{Z} \, \widehat{\mathbf{H}}) \in \mathbb{R}^{(n-p) \times d}$. Denoting $(U_{p+1}, U_{p+2}, \ldots, U_n)^\top := \mathbf{U}$, we have that

$$U_k = \left(\mathbf{X}^{(\Delta)}_k - \Delta \hat{\eta} - \sum_{l=1}^{p} \Delta \widehat{H}_l^\top \, \mathbf{X}^{(\Delta)}_{k-l} \right), \quad k = p+1, p+2, \ldots, n.$$

Furthermore, let $\mathbf{U}^{(\mathrm{rep})} \in \mathbb{R}^{d^2(n-p) \times d}$ be a matrix consisting of d^2 repetitions of the $\mathbf{U}$ matrix stacked on top of each other.

(iii) Point-wise multiply $\mathbf{C}$ from (i) and $\mathbf{U}^{(\mathrm{rep})}$ from (ii) and square the row sums of the resulting matrix. Row-wise fill the resulting vector into a $d^2 \times (n-p)$ matrix and compute the row sums of this matrix.

(iv) Row-wise fill the result from (iii) into a $d \times d$ matrix. This yields $\left(\hat{\sigma}^2_{i,j}\right)_{1 \le i,j \le d}$.

3.3 *Estimation of the Hawkes graph*

Given an estimate $\widehat{\mathcal{G}}^*_{\mathbf{N}}$ of the Hawkes skeleton $\mathcal{G}^*_{\mathbf{N}}$ from Definition 3.2, we consider the estimation of the Hawkes graph $\mathcal{G}_{\mathbf{N}}$; see Definition 2.6. We aim to estimate vertex as well as edge weights, and to calculate corresponding confidence bounds for both. That is, after the more structural Hawkes skeleton estimation from Section 3.2, we now *quantify* the various interactions between the observed event-streams. Typically, after the skeleton estimation, we can reduce the effective dimensionality of the model considerably: in a first obvious step, we divide the graph $\widehat{\mathcal{G}}^*_{\mathbf{N}}$ into its strongly-connected subgraphs and treat them separately. In a second step, we identify $\widehat{\mathrm{PA}}_{\mathbf{N}}(j) := \{i \in \mathcal{V}_{\mathbf{N}} : (i,j) \in \widehat{\mathcal{E}}^*_{\mathbf{N}}\}$ for all $j \in \mathcal{V}_{\mathbf{N}}$. From the branching construction of a Hawkes process, respectively, of Hawkes families in Definitions 2.2 and 2.4, we have that any event in component j is either an immigrant stemming from a Poisson random measure with constant intensity η_j or has a direct explanation through an event in one of its parent components $\mathrm{PA}_{\mathbf{N}}(j)$. That is, in a multivariate version of (6), it suffices to regress the bin-counts of component j on the bin-counts in $\mathrm{PA}_{\mathbf{N}}(j)$. The constant term in this regression will represent the j-th immigration intensity. Considering only the parents instead of d other components in the conditional-least-squares regression increases numerical and estimation efficiency. In applications, however, we *do not know* the true parent set $\mathrm{PA}_{\mathbf{N}}(j)$. So, we have to substitute $\mathrm{PA}_{\mathbf{N}}$ with the estimate $\widehat{\mathrm{PA}}_{\mathbf{N}}$. As long as $\mathrm{PA}_{\mathbf{N}}(j) \subset \widehat{\mathrm{PA}}_{\mathbf{N}}(j)$ this is not an issue: from the branching construction, we have that component j, conditional on $\sigma(N_i(A) : A \in \mathcal{B}((-\infty, t]), i \in \mathrm{PA}_{\mathbf{N}}(j))$, is independent of the past of all other components $\sigma(N_i(A) : A \in \mathcal{B}((-\infty, t]), i \notin \mathrm{PA}_{\mathbf{N}}(j))$. Consequently, additional vertices in the estimated parent sets do not introduce additional bias in this graph estimation.

Apart from this restriction of the regression variables on (estimated) parent types, we apply the conditional-least-squares approach as in Definition 3.1. This time however, due to reduction of dimensionality, we will typically *be able to choose a much smaller bin-size* Δ than for the skeleton estimation before. To ease implementation, below we give convenient notations and the necessary calculations.

First, we drop the $\mathbf{N}$ subscript for the parent sets $\mathrm{PA}(j)$. Also, we write $\mathrm{PA}(j)$ instead of $\widehat{\mathrm{PA}}(j)$—keeping in mind that the first has to be substituted by the latter in most applications. For $k = 1, 2, \ldots, n$, $j = 1, 2, \ldots, d$ and some $\Delta > 0$ let $\mathbf{X}_{k,j}^{(\Delta_{\mathrm{graph}})}$ denote the j-th component of the d-variate bin-count sequence at time k with respect to $\mathbf{N}$; see (7). Furthermore, set $d_j := |\mathrm{PA}(j)|$ and

$$\mathbf{X}_{k,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})} := \left(\mathbf{X}_{k,i_1}^{(\Delta_{\mathrm{graph}})}, \mathbf{X}_{k,i_2}^{(\Delta_{\mathrm{graph}})}, \ldots, \mathbf{X}_{k,i_{d_i}}^{(\Delta_{\mathrm{graph}})}\right)^\top. \tag{15}$$

In (15) and in all definitions below, we denote $\{i_1, i_2, \ldots, i_{d_j}\} := \mathrm{PA}(j)$ such that $i_1 < i_2 < \cdots < i_{d_j}$. The idea is to regress all d event types separately on the past of their parents with Ansatz

$$\mathbb{E}\left[\mathbf{X}_{n,j}^{(\Delta_{\mathrm{graph}})} \middle| \mathbf{X}_{n-k,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})},\, k = 1, 2, \ldots, p\right] = \alpha_{0,j}^{(\Delta_{\mathrm{graph}})} + \sum_{i \in \mathrm{PA}(j)} \sum_{k=1}^{p} \alpha_{k,i,j}^{(\Delta_{\mathrm{graph}})} \mathbf{X}_{n-k,i}^{(\Delta_{\mathrm{graph}})}, \quad j = 1, 2, \ldots, d. \tag{16}$$

Ansatz (16) should be compared with (6). Note that j itself may or may not be an element of $\mathrm{PA}(j)$. In other words, there might be an autoregressive part in (16) or not.

Definition 3.3: Let $\mathcal{G}_{\mathbf{N}}^*$ be a Hawkes skeleton (estimate) with respect to some d-variate Hawkes process (data) $\mathbf{N}$. Given $(\mathrm{PA}(j))_{j=1,2,\ldots,n}$, a bin size $\Delta_{\mathrm{graph}} > 0$, and a support s with $0 < \Delta_{\mathrm{graph}} \leq s < T$, $p := \lfloor s/\Delta_{\mathrm{graph}} \rfloor$ calculate the conditional-least-squares estimates

$$\widehat{\mathbf{H}}_j^{(\Delta_{\mathrm{graph}}, s)} := \frac{1}{\Delta_{\mathrm{graph}}} \left(\mathbf{Z}_j^\top \mathbf{Z}_j\right)^{-1} \mathbf{Z}_j^\top \mathbf{Y}_j \in \mathbb{R}^{(pd_j+1)\times 1}, \quad j = 1, 2, \ldots, d_j, \tag{17}$$

with design matrices

$$\mathbf{Z}_j := \begin{pmatrix} (\mathbf{X}_{p,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & (\mathbf{X}_{p-1,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & \ldots & (\mathbf{X}_{1,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & 1 \\ (\mathbf{X}_{p+1,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & (\mathbf{X}_{p,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & \ldots & (\mathbf{X}_{2,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (\mathbf{X}_{n-1,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & (\mathbf{X}_{n-2,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & \ldots & (\mathbf{X}_{n-p,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})})^\top & 1 \end{pmatrix} \in \mathbb{N}_0^{(n-p)\times(d_j p+1)}, \quad j = 1, 2, \ldots, d, \tag{18}$$

and vectors of responses

$$\mathbf{Y}_j := \left(\mathbf{X}_{p+1,j}^{(\Delta_{\mathrm{graph}})}, \mathbf{X}_{p+2,j}^{(\Delta_{\mathrm{graph}})}, \ldots, \mathbf{X}_{n,j}^{(\Delta_{\mathrm{graph}})}\right)^\top \in \mathbb{N}_0^{(n-p)\times 1}, \quad j = 1, 2, \ldots, d.$$

Given $\widehat{\mathbf{H}}_j^{(\Delta_{\mathrm{graph}}, s)}$, $j = 1, 2, \ldots, d$, we define the *Hawkes-graph estimator* $\widehat{G}_{\mathbf{N}} := (\widehat{\mathcal{V}}_{\mathbf{N}}, \widehat{\mathcal{E}}_{\mathbf{N}})$ with $\widehat{\mathcal{V}}_{\mathbf{N}} := \{(j; \hat{\eta}_j) :\ j = 1, 2, \ldots, d\}$ and

$$\widehat{\mathcal{E}}_{\mathbf{N}} := \bigcup_{j=1,\ldots,d} \left\{(i_l, j; \hat{a}_{i_l,j}) :\ \{i_1, \ldots, i_{d_j}\} = \mathrm{PA}(j),\, \hat{a}_{i_l,j} = b_{l,j}^\top \widehat{\mathbf{H}}_j^{(\Delta_{\mathrm{graph}}, s)}\right\}, \tag{19}$$

where, for $l = 1, 2, \ldots, d_j$, $b(l, j) \in \{0, 1\}^{(d_j p+1)\times 1}$ is a column vector with zeroes in all components but 1s in components $((k-1)d_j + l)$, $k = 1, 2, \ldots, p$. Furthermore, for $\alpha_{\mathrm{graph}} \in (0, 1)$, we define the confidence intervals $[\hat{\eta}_j \pm \hat{\sigma}_j z_{1-\alpha_{\mathrm{graph}}}^{-1}]$ and, for $i_l \in \mathrm{PA}_{\mathbf{N}}(j)$, $[\hat{a}_{i_l,j} \pm \hat{\sigma}_{i_l,j} z^{-1}(1-\alpha_{\mathrm{graph}})]$. We give the calculation of $\hat{\sigma}_{i_l,j}$ and $\hat{\sigma}_j$ in Algorithm **2**, below.

As before, additional notation clarifies what the entries of the $\widehat{\mathbf{H}}_j^{(\Delta_{\text{graph}},s)}$ matrix actually estimate:

$$\begin{pmatrix} \widehat{H}_{\text{PA}(j),j}(\Delta_{\text{graph}}) \\ \widehat{H}_{\text{PA}(j),j}(2\Delta_{\text{graph}}) \\ \dots \\ \widehat{H}_{\text{PA}(j),j}(p\Delta_{\text{graph}}) \\ \hat{\eta}_j \end{pmatrix} := \widehat{\mathbf{H}}_j, \text{ with} \tag{20}$$

$$\widehat{H}_{\text{PA}(j),j}(k\Delta_{\text{graph}}) = \left(\hat{h}_{i_1,j}(k\Delta_{\text{graph}}), \hat{h}_{i_2,j}(k\Delta_{\text{graph}}), \dots, \hat{h}_{i_{d_j},j}(k\Delta_{\text{graph}})\right)^\top,$$

$k = 1, 2, \dots, p$ and $\{i_1, i_2, \dots, i_{d_j}\} = \text{PA}(j)$.

Finally, we provide efficient computations for the covariance estimates that are necessary for the confidence intervals around the estimated edge and vertex weights.

Algorithm 2: Let $j \in \{1, 2, \dots, d\}$ such that $|\text{PA}(j)| > 0$ and let $\{i_1, i_2, \dots, i_{d_j}\} = \text{PA}(j)$ with $i_1 < i_2 < \dots, < i_{d_j}$. For (i_l, j), $l = 1, 2, \dots, d_j$, let $e(i_l, j) \in \{0,1\}^{(d_j p+1)\times 1}$ be a column vector with all entries zero but 1s at components $(k-1)d_j + (l-1)$, $k = 1, 2, \dots, p$. We compute $\hat{\sigma}_{i_l,j}$ in the following way:

(i) Compute $\mathbf{C}_{l,j} := e(i_l, j)^\top \left((\mathbf{Z}_j^\top \mathbf{Z}_j)^{-1} \mathbf{Z}_j^\top\right) \in \mathbb{R}^{1\times(n-p)}$.

(ii) Set $\mathbf{U}_j = (\mathbf{Y}_j - \Delta_{\text{graph}} \mathbf{Z}_j \widehat{\mathbf{H}}_j) \in \mathbb{R}^{(n-p)\times 1}$. Denoting $(U_{p+1,j}, U_{p+2,j}, \dots, U_{n,j})^\top := \mathbf{U}_j$, we have that

$$U_{k,j} = \left(\mathbf{X}_{k,j}^{(\Delta_{\text{graph}})} - \Delta_{\text{graph}}\hat{\eta} - \sum_{m=1}^{p} \Delta_{\text{graph}} \widehat{H}_{\text{PA}(j),j}^\top(m\Delta_{\text{graph}})\, \mathbf{X}_{k-m,\text{PA}(j)}^{(\Delta_{\text{graph}})}\right),$$

for $k = p+1, p+2, \dots, n$.

(iii) Pointwise multiply $\mathbf{C}_j$ and $\mathbf{U}_j$. The sum of the squares of the result yields $\hat{\sigma}_{i_l,j}^2 \in \mathbb{R}_{\geq 0}$.

For the variance estimates corresponding to the j-th vertex weight, consider the last row of $\left((\mathbf{Z}_j^\top \mathbf{Z}_j)^{-1} \mathbf{Z}_j\right) \in \mathbb{R}^{(d_j p+1)\times(n-p)}$, multiply it point-wise with $\mathbf{U}_j$ from above, take the sum of squares of the results and multiply the result with $\Delta_{\text{graph}}^{-2}$; this yields $\hat{\sigma}_j^2$.

Remark 2: The bin size Δ_{graph} for the graph estimation in Definition 3.3 will typically be much smaller than the bin-size Δ_{skel} for the skeleton estimation in Definition 3.2. Depending on the application, after the graph estimation one might also want to delete edges with edge weight estimates non-significantly different from zero, or treat vertex weight estimates, respectively, immigration intensities, that are non-significantly different from zero as zero; see Figure 2. Also note that the latter could possibly be tested with a different significance parameter α_{vertex} than the significance parameter α_{graph} from the edge-weight estimation. In any case, the resulting Hawkes graph estimations ought to be checked for *redundant vertices*; see Definition 2.5. If the estimate has redundant vertices, the results are typically inconsistent with the data—as we typically observe data in all components. Therefore, if a fitted model has redundant vertices, we increase α_{skel}, α_{graph}, and/or α_{vertex}. Thus, we obtain more estimated non-zero immigration intensities and/or larger estimated edge sets. We proceed with increasing the significance parameters until there are no redundancies left.

Given a Hawkes-graph estimate as in Definition 3.3, one may examine connectivity issues, path-weights, graph distances, feedback and cascade coefficients, exploit graphical representations, etc.; see the example in Section 4.

3.4 *Estimation of the non-zero excitement functions*

For many applications, the results discussed above may already suffice. In other applications however, the graph estimation will only be a preliminary step and one would like to examine how the various excitements are distributed *over time*. In other words, one would like to explicitly estimate the reproduction intensities.

Parametric estimation Given the Hawkes estimator from Definition 3.1, the Hawkes model is not yet completely specified. In particular, (17) only yields estimates of the reproduction intensities on a grid:

$$\left\{\Big(k\Delta\Big), \hat{h}_{i,j}(k\Delta)\right\}_{k=1,2,\ldots,p}, \quad i \in \widehat{\mathrm{PA}}(j),\ j = 1, 2, \ldots, d. \tag{21}$$

One obvious possibility to complete the model specification would be the application of any kind of smoothing method on (21). We want to consider another approach: we exploit (21) graphically (examine log/log-plots, id/log-plots, check for local maxima, convex/concave regions, etc.) and identify appropriate parametric families. The parameters can then be fitted to the estimates (21) via non-linear least-squares:

Definition 3.4: Consider a Hawkes-graph estimation as in Definition 3.3 with respect to some d-variate Hawkes process data and a bin size $\Delta_{\mathrm{graph}} > 0$. For $j = 1, 2, \ldots, d$ and $i \in \widehat{\mathrm{PA}}(j)$, let $(h_{i,j}^{(\theta_{i,j})} : \mathbb{R} \to \mathbb{R}_{\geq 0})$ be function families parametrized by $\theta_{i,j} \in I_{i,j} \subset \mathbb{R}^{d_{i,j}}$. With the notation from (21), let

$$\hat{\theta}_{i,j} := \mathrm{argmin}_{\theta \in I_{i,j}} \sum_{k=1}^{p} \Big(h_{i,j}^{(\theta)}\big(k\Delta_{\mathrm{graph}}\big) - \hat{h}_{i,j}(k\Delta_{\mathrm{graph}})\Big)^2, \quad (i,j) \in \widehat{\mathcal{E}}^*, \tag{22}$$

and define the *parametric reproduction-intensity estimates*

$$\hat{h}_{i,j}^{(\mathrm{par})}(t) := \begin{cases} h_{i,j}^{(\hat{\theta}_{i,j})}(t), & (i,j) \in \widehat{\mathcal{E}}^*, \quad t \in \mathbb{R}, \\ 0, & (i,j) \notin \widehat{\mathcal{E}}^*, \quad t \in \mathbb{R}, \end{cases}$$

the *parametric branching-matrix estimate*

$$\widehat{A}^{(\mathrm{par})} := \left(\int \hat{h}_{i,j}^{(\mathrm{par})}(t)\mathrm{d}t\right)_{1 \leq i,j \leq d},$$

and the *parametric immigration-intensity estimates*.

$$\hat{\eta}^{(\mathrm{par})} := \Big(\hat{\eta}_1^{(\mathrm{par})}, \ldots, \hat{\eta}_d^{(\mathrm{par})}\Big)^\top := \Big(1_{d\times d} - (\widehat{A}^{(\mathrm{par})})^\top\Big)\,\lambda^{(\mathrm{emp})}, \tag{23}$$

where $\lambda^{(\mathrm{emp})}$ denotes the observed empirical intensity $\lambda^{(\mathrm{emp})} := \mathbf{N}\big((0,T]\big)/T \in \mathbb{R}_{\geq 0}^{d\times 1}$.

Remark 3 : The definition of $\eta^{(\mathrm{par})}$ in (23) is motivated by the equality

$$\Big(1_{d\times d} - (\widehat{A}^{(\mathrm{par})})^\top\Big)^{-1} \eta^{(\mathrm{par})} = \lambda^{(\mathrm{emp})}.$$

In other words, with this choice of $\hat{\eta}^{(\mathrm{par})}$, the observed unconditional intensity exactly equals the estimated unconditional intensity. This might be relevant in some applications (e.g., simulation

from a fitted model). The results from the simulation study below suggest that the parameter-estimates obtained from above are asymptotically normally distributed. Finally note that it might often be more efficient to consider weighted least-squares in (22).

The above may be the only feasible way to retrieve parametric estimates. MLE—at least if it is applied in a straightforward manner—might require far too much computing time if the data contains many event streams with many events.

4. Example

We illustrate the concepts introduced above with a ten-dimensional Hawkes model. We perform a simulation study and apply the estimation methods from Sections 3.2, 3.3, and 3.4 to the Hawkes skeleton, the Hawkes graph, and the reproduction intensity parameters.

4.1 *Example model*

We consider a 10-dimensional Hawkes process $\mathbf{N}$ with immigration intensities

$$\eta_i := \begin{cases} 1, & i \in \{1, 7, 10\}, \\ 0, & i \in \{2, 4, 5, 6, 7, 8, 9\}, \end{cases} \tag{24}$$

and reproduction intensities $h_{i,j}$, $(i, j) \in \{1, 2, \ldots, 10\}^2$, defined, for $t \in \mathbb{R}$, by

$$h_{i,j}(t) := \begin{cases} 1.5\,\gamma(t), & (i,j) \in \{(1,2),(2,4),(8,9)\}, \\ 1_{t\in[1,2]}0.5, & (i,j) \in \{(1,1),(2,3),(3,5),(4,3),(4,5),(4,6),(5,3),(7,8),(9,7)\}, \\ 1_{t\in[1,2]}0.1, & (i,j) = (5,7), \\ 0, & \text{else.} \end{cases} \tag{25}$$

Here, γ denotes a Gamma density with shape parameter 6 and rate parameter 4, i.e., $\gamma(t) = 1_{t\geq 0} t^5 \exp\{-4t\}(4^6)/(5!)$. In Hawkes graph terminology, we have 13 edges supplied with three different kinds of edge weights: a heavy weight (1.5) for three edges, a light weight (0.5) for seven edges, and one edge with a super-light weight (0.1). An illustration of the corresponding graph $\mathcal{G}_{\mathbf{N}}$ is much more meaningful than (25); see the left graph in Figure 1. From this figure, the various direct and indirect dependencies can be read off in a very direct way: only the large nodes have non-zero immigration intensity; a fat edge corresponds to an edge weight of 1.5; a thin edge corresponds to an edge weight of 0.5; the dashed line corresponds to the super-light edge weight (0.1). We examine the Hawkes-graph properties introduced in Definitions 2.6 and 2.8:

Redundancy The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ has no *redundant vertices*: all small vertices have a large vertex as one of their ancestors. If vertex 1 were small, the vertices 1,2,3,4,5,6 would be *redundant* as they could not generate events.

Connectivity The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is *not strongly connected*. The graph can be divided in two separate strongly-connected Hawkes subgraphs with vertex sets $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and $\{10\}$. Deleting edge $(5, 7; 0.1)$ would yield three separate strongly-connected Hawkes subgraphs.

Criticality The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is *subcritical*: all vertices but vertex 10 are part of closed walks. It suffices to check criterion (4) for vertices $i = 1, 2, 7$. For vertex 1, we find that

$$\mathcal{W}_g^{(1)} = \{(\underbrace{1, 1, \ldots, 1}_{g \text{ times}})\},\ g \in \mathbb{N}, \quad \text{and} \quad |(\underbrace{1, 1, \ldots, 1}_{g \text{ times}})| = 0.5^g,\ g \in \mathbb{N}\,.$$

Consequently, $\sum_{g=1}^{\infty} \sum_{w_g^{(1)} \in \mathcal{W}_g^{(1)}} |w_g^{(1}| = \sum_{g=1}^{\infty} 0.5^g < \infty$. For vertex 2, we find that

$$\mathcal{W}_1^{(2)} = \mathcal{W}_2^{(2)} = \emptyset,\ \mathcal{W}_3^{(2)} = \{(2, 4, 6, 2)\},\ \mathcal{W}_4^{(2)} = \mathcal{W}_5^{(2)} = \emptyset,\ \mathcal{W}_3^{(2)} = \{(2, 4, 6, 2, 4, 6, 2)\}, \ldots$$

With $|(2, 4, 6, 2)| = 1.5 \cdot 0.5 \cdot 0.5 = 0.375$, $|(2, 4, 6, 2, 4, 6, 2)| = 0.375^2, \ldots$, criterion (4) again follows. For vertex 7, one argues analogously. In other words, as long as closed walks do not overlap, we can construct large subcritical Hawkes graphs without calculating any eigenvalues. When closed walks overlap, the underlying combinatorics typically become too involved as to proceed in this manner. In this case one could calculate the spectral radius of the adjacency matrix of the involved edges only. For example, if we wanted to introduce another edge $(9, 5; a_{9,5}))$ in model (25), respectively, Figure 1, we would have to calculate the spectral radius of the adjacency matrix corresponding to the Hawkes graph with edges $\{(3, 5; 0.5), (5, 3; 0.5), (5, 7; 0.1), (7, 8; 0.5), (8, 9; 0.5), (9, 5; a_{9,5}), (9, 7; 0.5)\}$; see the discussion after (4).

Cascade and feedback coefficients We calculate the coefficients from Definition 2.8 with respect to the example model:

	1	2	3	4	5	6	7	8	9	10
cascade.coefficients	0.42	0.11	0.03	0.16	0.03	0.07	0.05	0.09	0.04	0.00
feedback.coefficients	1.00	0.08	0.04	0.14	0.05	0.08	0.15	0.24	0.12	

The cascade and feedback coefficients are a summary of the branching dynamics of the process, that is not regarding immigration intensities. *Cascade coefficients*: consider the first vertex. Given exactly one immigrant event in each vertex, approximately 42% of all the subsequent events will stem from the family triggered by vertex 1 whereas vertex 10 does not trigger any cascade at all. *Feedback coefficients*: again, given exactly one immigrant in each vertex, for each vertex, we check which proportion of all events is due to the immigrant in the vertex itself. In our example, for vertex 8 this means that we decrease its activity by 24% if we delete its outgoing edge. For vertex 10, the feedback-coefficient notion makes no sense because it experiences no excitement from any other vertex.

4.2 *Simulation study*

We simulate $n_{\text{sim}} = 1000$ realizations of the Hawkes process $\mathbf{N}$ from Section 4.1. We use the branching construction from Definitions 2.2 and 2.4 as simulation algorithm. In each realization,

Table 1. $\Delta_{\text{skel}} = 0.2$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.324	0.902	1.000	0.956	0.121	0.993
0.010	13.066	0.917	1.000	0.970	0.190	0.987
0.050	17.296	0.946	1.000	0.990	0.379	0.942
0.100	21.995	0.959	1.000	0.995	0.507	0.890
0.250	35.015	0.979	1.000	0.999	0.739	0.744

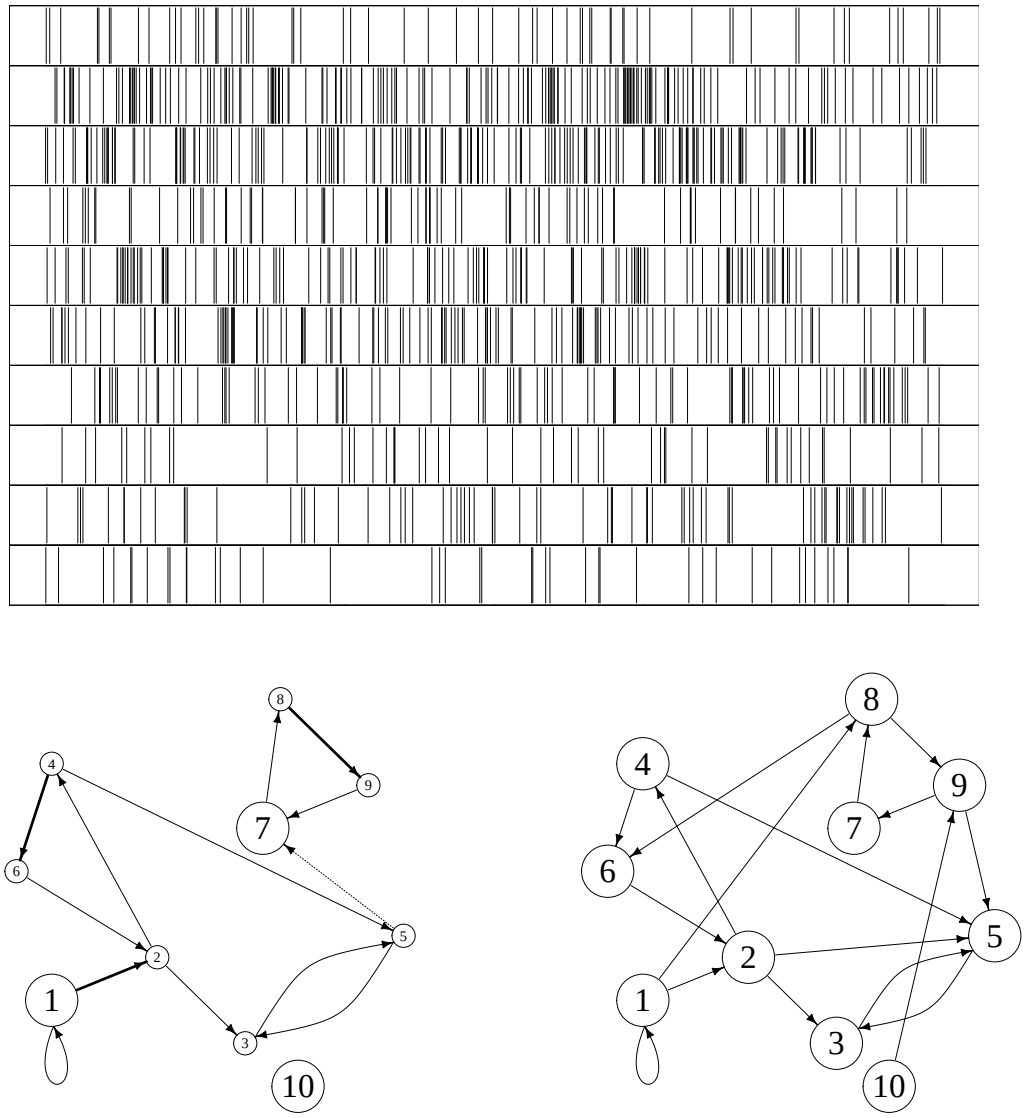

Figure 1. Hawkes process simulation, Hawkes graph, and estimated Hawkes skeleton. The left graph represents the Hawkes graph corresponding to the Hawkes process example from Section 4.1; the graph is a summary of the immigration and branching structure of the model: edges from one vertex to another vertex denote non-zero reproduction intensities, respectively, excitement. Fat edges refer to heavy excitement (1.5 expected children events in branching construction); thin edges to small excitement (0.5 expected children) and the dotted line refers to a very small excitement (0.1 expected children); see (25). Large vertices correspond to nonzero immigration-intensities (1) and small vertices to the zero immigration-vertices; see (24). The barcode plots illustrate a 30 time-units window of a simulated realization of the model (after some burn-in): we observe events of ten types, respectively, in ten components. One goal of our paper is to retrieve the graph on the left from such a realization. As a first step towards this aim, we calculate the Hawkes-skeleton estimate from Definition 3.3 with respect to a coarse bin-size $\Delta_{\text{skel}} = 1$ and a sparseness parameter $\alpha_{\text{skel}} = 0.05$. The right graph illustrates this estimate. This skeleton will be used in a second step to retrieve the Hawkes-graph estimate; see Figure 2. Comparing the skeleton with the true graph on the right, we see that we catch all the twelve of the thirteen true edges. We miss edge $(5, 7)$. Furthermore, we introduce four additional wrong edges $(1, 8)$, $(2, 5)$,$(8, 6)$, $(9, 5)$, and $(10, 9)$. The three crucial points to notice are: (i) The false-positive edges *do not introduce additional bias* in the graph estimation. (ii) Due to the coarse Δ_{skel}-value, the calculation of the skeleton estimate is computationally simple. (iii) The resulting skeleton estimate is nearly as sparse as the true skeleton and thus reduces the complexity of the graph estimation (with a very fine Δ_{graph}-parameter) extremely. See Figure 2, for the Hawkes-graph estimation with respect to the skeleton estimate from above.

Table 2. $\Delta_{\text{skel}} = 0.5$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.353	0.902	1.000	0.957	0.120	0.993
0.010	13.118	0.917	1.000	0.971	0.179	0.986
0.050	17.255	0.945	1.000	0.990	0.375	0.943
0.100	21.952	0.959	1.000	0.995	0.514	0.891
0.250	34.805	0.980	1.000	0.999	0.745	0.746

Table 3. $\Delta_{\text{skel}} = 1$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.476	0.910	1.000	0.967	0.129	0.993
0.010	13.171	0.921	1.000	0.977	0.178	0.986
0.050	17.264	0.949	1.000	0.993	0.400	0.943
0.100	21.806	0.962	1.000	0.997	0.535	0.893
0.250	34.465	0.979	1.000	0.999	0.730	0.750

Table 4. $\Delta_{\text{skel}} = 2$

alpha.skel	nedges	total	heavy	light	super.light	zero
0.005	12.244	0.810	1.000	0.828	0.074	0.980
0.010	13.680	0.846	1.000	0.876	0.119	0.969
0.050	19.709	0.913	1.000	0.957	0.262	0.910
0.100	25.065	0.936	1.000	0.978	0.369	0.852
0.250	38.186	0.966	1.000	0.994	0.605	0.705

we simulate a time window of 500 time units. This typically yields between 500 and 2000 events per component. Given each of these realized event streams, we calculate the Hawkes-skeleton estimator from Definition 3.2—with respect to different values of Δ_{skel} and α_{skel}. Given these skeleton estimates, we calculate the Hawkes-graph estimator from Definition 3.3—including confidence bounds for all vertex and edge weights. Finally, we analyze the scatterplots for branching-intensity estimates, choose parametric function families, and fit the parameters on the estimates by nonlinear least-squares. Figures 1 and 2 illustrate the procedure.

Hawkes-skeleton estimation We fix $s = 5$ and, for each simulated event-stream, we calculate the Hawkes-skeleton estimates from Definition 3.2 with respect to this support parameter s, bin-sizes $\Delta_{\text{skel}} \in \{0.2, 0.5, 1, 2\}$, and various sparseness-parameters $\alpha_{\text{skel}} \in \{0.005, 0.01, 0.05, 0.1, 0.25\}$. We denote the estimated edge sets by $\{\widehat{\mathcal{E}}^*(k)\}_{k=1,2,\ldots,n_{\text{sim}}}$ and the true edge set by $\mathcal{E}^*$. Using this notation, we summarize the results of the simulation study in Tables 1, 2, 3, and 4 with the following statistics:

a) *nedges*: average size of estimated edge-sets (true number is 13), that is, $\sum_{k=1}^{n_{\text{sim}}} |\mathcal{E}^*(k)|/n_{\text{sim}}$.
b) *total*: fraction of correctly estimated edges, that is,

$$\frac{\sum_{k=1}^{n_{\text{sim}}} \sum_{(i,j)\in\mathcal{E}^*} 1_{\{(i,j)\in\widehat{\mathcal{E}}^*(k)\}}}{n_{\text{sim}}|\mathcal{E}^*|}.$$

Note that $1 - \textit{total}$ is the *false-negative rate*.
c) *heavy/light/super.light*: more detailed version of b) above; fractions of correctly estimated edges with heavy (1.5), light (0.5) and super-light (0.1) edge weights.

d) *zero*: fraction of correctly left-out edges, i.e., of pairs $(i,j) \notin \widehat{V}^*_{\mathbf{N}}(k)$ such that $(i,j) \notin V_{\mathbf{N}}$:

$$\frac{\sum_{k=1}^{n_{\text{sim}}} \sum_{(i,j)\notin\mathcal{E}^*} 1_{\{(i,j)\notin\widehat{\mathcal{E}}^*(k)\}}}{n_{\text{sim}}\big(d^2 - |\mathcal{E}^*|\big)}.$$

Note that $1 - \textit{zero}$ is the *false-positive rate*.

First, we consider the estimations with respect to bin-size $\Delta_{\text{skel}} = 0.2$; see Table 1. We note from the last column, *zero*, that the false-positive rate is indeed very close to the value of the chosen theoretical significance level α_{skel}. Going back to Definition 3.2, we see that the larger α_{skel}, the more edges are included in the Hawkes-skeleton estimation. This is reflected in all of the columns. However, even for very small α_{skel}, we detect *all* of the edges with a heavy edge-weight and most of the edges with light edge weight. The edge $(5,7)$ with the super-light weight (0.1) is obviously a hard-to-detect alternative to the zero hypothesis. Note that Tables 2, 3, and 4 look roughly the same as Table 1 one above—though the estimates were calculated with respect to completely different bin-sizes Δ_{skel}.

So, in this first estimation step, we may use a very coarse bin-size Δ_{skel}. This makes the calculations underlying skeleton estimation feasible even for much higher dimensions.

The main purpose of the skeleton estimation is to lay the ground for the graph estimation that itself depends on a given estimated skeleton; see Definition 3.3. Missing edges in the skeleton estimate will typically introduce a bias for the graph-weight estimates. We therefore want to keep the false-negative rate in the skeleton estimation very small. As a consequence, we need α_{skel} large. Note that false-positive edges do *not* add additional bias in the graph estimation; see Section 3.3. So the increase of the false-positive rate (that is, the decrease in the *zero*-column) does not prevent us from increasing the α_{skel}-parameter. Note, however, that the whole reason of the two-step estimation procedure is that in the first step we want to take advantage of the sparseness of the underlying true Hawkes graph and *reduce* the complexity of the a priori fully connected network. Too many additional false-positive edges would hamper this advantage. In this sense, not only Δ_{skel} but also α_{skel} can be understood as a parameter controlling the numerical complexity of the method: the smaller α_{skel}, the sparser the estimated skeleton, the less complex the computations for the Hawkes-graph estimate from Definition 3.3. We see in our tables that, for all choices of Δ_{skel} and all values of α_{skel}, we typically catch all the true edges, i.e., the false-negative rate is really small. In the next section, we will see that the graph estimates are not very sensitive to the α_{skel} parameter in the skeleton estimation.

Simulation study: Hawkes-graph estimation As a next step, we quantify the estimated excitements. That is, given a Hawkes skeleton, we estimate the corresponding graph as in Definition 3.3; see Figure 2.

We do this both with respect to the true skeleton and with respect to the estimated skeletons from the first estimation step. For comparison, we apply the skeletons with respect to different α_{skel}-parameters; we only consider the ones with respect to $\Delta_{\text{skel}} = 1$. As opposed to the skeleton estimation, we may now use a much smaller bin-size $\Delta_{\text{graph}} = 0.1$. In the present example, this is about the lower bin-size bound for tolerable computing time for the simulation study using a 2.3 GHz Intel Core processor (about 10sec for each of the estimations). Furthermore, we apply $s = 5$ and $\alpha_{\text{graph}} = 0.05$ in the calculation. For each simulation, we also calculate the confidence bounds for all vertex and edge weights. Table 4.2 reports the coverage rates.

The coverage rates of the graph estimations that were calculated with respect to the true underlying skeleton correspond well with the significance parameter $\alpha_{\text{graph}} = 0.05$. Naturally, the coverage rates for the estimates with respect to the estimated skeleton are smaller: as soon as the estimated skeleton misses an edge (e.g., the edge (5,7;0.1) with the super-light weight), the model calibration balances this missing possibility of excitement by increased baseline intensities or increased edge weights. The larger α_{skel}, the lower the probabilty of missing an edge, the

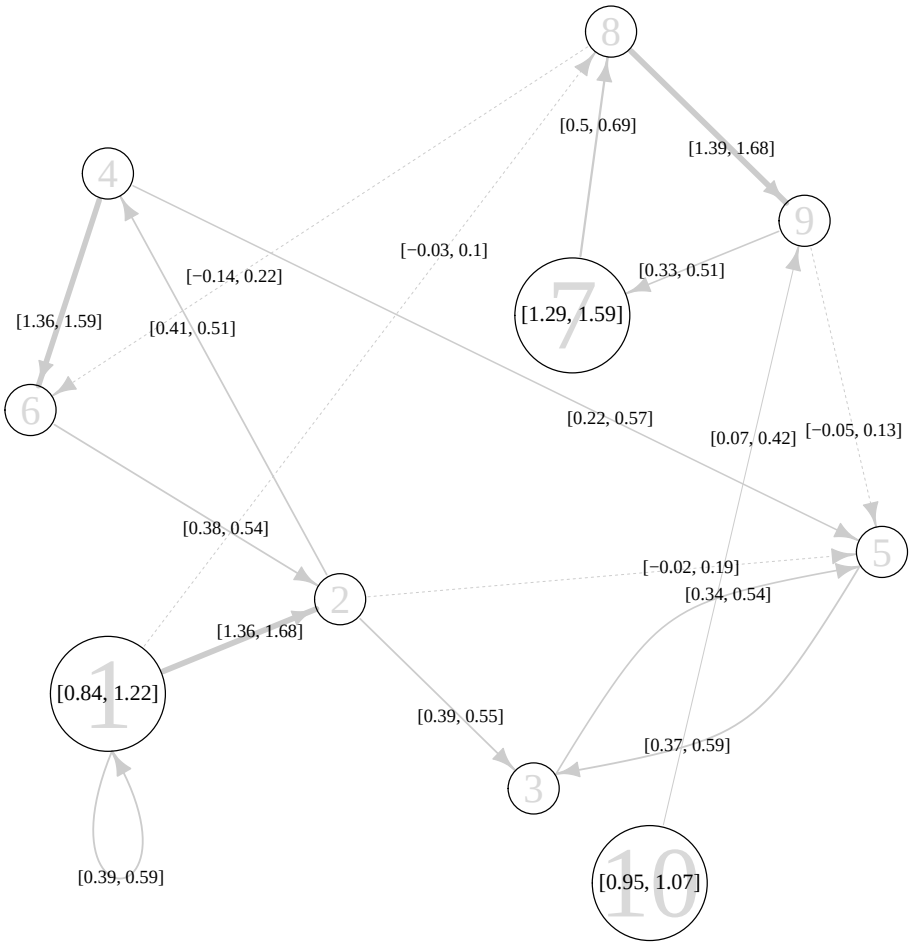

Figure 2. Hawkes-graph estimation. Given a single simulation of length $T = 1000$ from the example model in Section 4.1, we calculate the Hawkes-graph estimator from Definition 3.3 with respect to the Hawkes-skeleton estimation from Figure 1; we apply a bin-size $\Delta_{\text{graph}} = 0.025$ and a significance parameter $\alpha_{\text{graph}} = 0.05$. This calculation allows us to supply each vertex and each node from this estimated skeleton with confidence intervals for their weights in the corresponding Hawkes graph. The edge widths are chosen proportional to the estimated edge-weights. Estimated edge-weights that are not significantly larger than zero are illustrated as a dashed edge. Similarly, vertices where the confidence interval for the vertex weight contains 0 are plotted as smaller circles—the corresponding confidence bounds are left away in this case. Comparing the results with the true Hawkes graph in Figure 1, respectively, with the Hawkes process parametrization in (24) and (25), we see that for all correct edges, the true weights are covered by the confidence intervals. And for the wrong, additional edges from the skeleton estimation $(1, 8)$, $(2, 5)$,$(8, 6)$, $(9, 5)$, we see that their weights are not significantly different from zero ($\alpha_{\text{graph}} = 0.05$). The estimated edge weight for the wrong edge $(10, 9)$ is significantly larger than zero but still small. All vertex weights but the weight of vertex 7 are also covered by the confidence intervals. The weight of vertex 7 is overestimated because in the skeleton estimation we missed the (light) edge $(5, 7; 0.1)$; this missing explanatory variable for the events in component 7 is compensated by a too large vertex weight in the graph estimation. Deleting all insignificant edges and setting the vertex weight of the insignificant vertex-weights to zero, we recover the original underlying graph almost perfectly.

Table 5. $\Delta_{\text{graph}} = 0.1$ and $\alpha_{\text{graph}} = 0.05$

applied.skeleton	vertex.weight.coverage	edge.weight.coverage
alpha.skel = 0.005	0.859	0.907
alpha.skel = 0.01	0.867	0.904
alpha.skel = 0.05	0.896	0.893
alpha.skel = 0.1	0.907	0.900
alpha.skel = 0.25	0.915	0.932
true skeleton	0.947	0.943

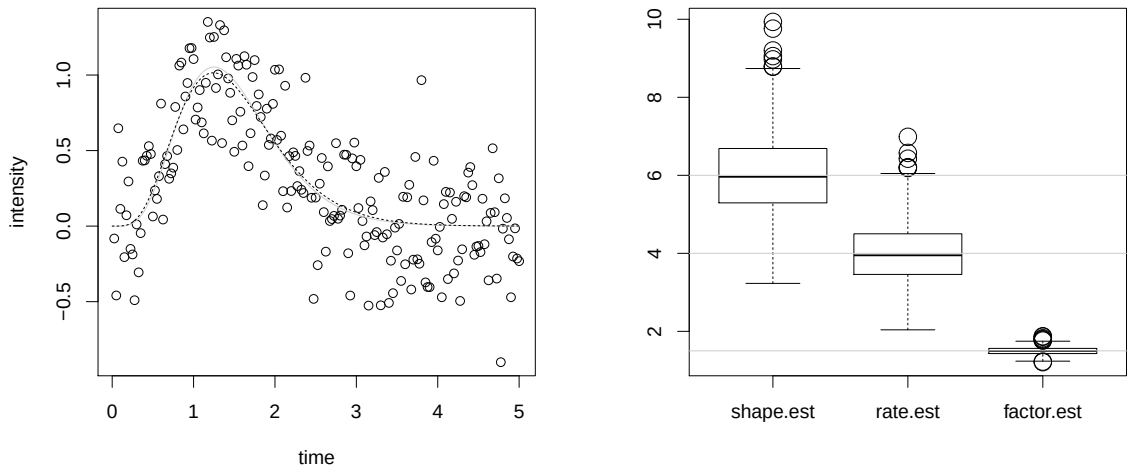

Figure 3. Parametric estimation. From a realization of the example model from Section 4.1, we calculate Hawkes-skeleton and Hawkes-graph estimate from Definitions 3.2 and 3.3; see Figures 1 and 2. As a by-product of these calculations, we retrieve estimates for the values of the reproduction intensities $h_{1,2}$ (circles in left panel) on an equidistant grid; see (20). The grey solid line in the left panel refers to the true underlying functions; see (25). For the pointwise estimates in the top right panel, the underlying parametric Gamma-family can be guessed. We fit the three necessary parameters by non-linear least squares as described in Section 3.4; the dotted line refers to the corresponding estimated parametric functions. The boxplots collect the parameter estimates for each of the 1000 estimations of the simulation study; the grey horizontal lines refer to the true values. Repeating this procedure for all estimated reproduction intensities, we can fully specify a Hawkes model.

better the coverage rates. Note, however, that at the same time, the corresponding skeleton estimate becomes increasingly dense and with it the graph estimation becomes increasingly time-consuming.

Parametric reproduction intensity estimation Finally, we check how the various excitements are distributed over time. As examples, we examine the reproduction intensity $h_{1,2}$. From the calculation of the Hawkes graph estimate, we retrieve estimates of the excitement function values on an equidistant grid; see (20). Based on the scatter plots of these estimates, we choose appropriate parametrized function families. Given such parametric functions, the parameters are fit to the point-wise estimates via non-linear least-squares; see Figure 3. QQ-plots (not included) support asymptotic normality for the parameter estimates.

5. Conclusion

We introduce the Hawkes graph and the Hawkes skeleton. We describe the immigration and branching structure of a Hawkes process in a graph-theoretical framework. Combining the new concepts with an estimation procedure from earlier work, we develop a statistical estimation method for the Hawkes graph. The key idea is that in a preliminary step we only test if there is excitement from any vertex to another vertex *at all*. We show that this first step is relatively simple to implement. The knowledge of the Hawkes skeleton makes the second step, the estimation of the Hawkes graph, much more efficient—both from a computational and statistical point of view. The simulation study shows that the procedure works as desired. As long as the true underlying graph is sparse (e.g. if the typical number of parents of a node is not larger than five and does not depend on the dimension of the process) the approach may be applied in even higher-dimensional situations. In any case, the method may be a useful tool for preliminary analysis when examining large multi-type event stream data in the Hawkes framework. It might be worthwile to study the distributional properties of the parameter estimates from Section 3.4 in more detail.

Note that the graph representation would also apply for discrete-time event-stream models, i.e., for multivariate time series of counts. More specifically, the present paper could have been developed in complete analogy for multivariate integer-valued autoregressive time series (INAR(∞)) which can be interpreted as discrete-time versions of the Hawkes process; see Kirchner (2015b). In this latter case, all results that we apply in our paper would be valid without taking any discretization error into account. In any case, when applied to real data, the discretization error is *not* the major drawback of our method: our method does indeed solve the important problem of how to decide whether an edge between two components exists at all. But for the specification of a Hawkes process we need to solve another—even more important—issue. We want to be able to decide whether we observe a *complete* Hawkes graph or whether our data lack some non-redundant vertices! In particular, the method presented will also yield reasonable results for data stemming from models with no or less underlying causality. The seeming excitement can then be explained by a confounding factor that we do not observe (and ignore). We believe, in view of the widespread interpretation of the Hawkes model as a causal model, an interpretation we share, it would be of utmost importance to derive tests for the presence of such hidden confounding in the event-stream context.

Acknowledgements

This research has been supported by the ETH RiskLab and the Swiss Finance Institute. The authors wish to express their gratitude to all the R-programmers providing and maintaining powerful statistical software. For our work, the igraph package (Csardi and Nepusz 2006) and the Matrix package (Bates and Maechler 2015) have been particularly helpful. We thank Vladimir Ulyanov for his comments on an earlier version of the paper which improved the presentation a lot.

References

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

Bacry, E., S. Gaïffas and Muzy, J.F., A generalization error bound for sparse and low-rank multivariate Hawkes processes. *ArXiv e-prints*, 2015.

Bates, D. and Maechler, M., Matrix: Sparse and Dense Matrix Classes and Methods. *R package version 1.1-5*, 2015 `http://CRAN.R-project.org/package=Matrix`.

Csardi, G. and Nepusz, T., The igraph software package for complex network research. *InterJournal*, 2006, **Complex Systems**, 1695.

Daley, D. and Vere-Jones, D., *An Introduction to the Theory of Point Processes*, Second Edition , Vol. I and II, , 2003, Springer.

Delattre, S., Fournier, N. and Hoffmann, M., Hawkes processes on large networks. *PNAS*, 2015, **105**.

Gunawardana, A., Meek, C. and Xu, P., A model for temporal dependencies in event streams. *Microsoft Research*, 2014.

Haccou, P., Jagers, P. and Vatutin, V., *Branching Processes*, 2005, Cambridge University Press.

Hall, E.C. and Willett, R.M., Tracking Dynamic Point Processes on Networks. *ArXiv e-prints*, 2014.

Hawkes, A., Point spectra of some mutually-exciting point processes. *Journal of the Royal Statistical Society: Series B*, 1971a, **33**, 438–443.

Hawkes, A., Spectra of some self-exciting and mutually-exciting point processes. *Biometrika*, 1971b, **58**, 83–90.

Hawkes, A., A cluster representation of a self-exciting point process. *Journal of Applied Probability*, 1974, **11**, 493–503.

Hawkes, T., On the class of the Sylow tower groups. *Mathematische Zeitschrift*, 1968, **105**, 393–398.

Jagers, P., *Branching Processes with Biological Applications*, 1975, John Wiley and Sons.

Kirchner, M., An estimation procedure for the Hawkes process. Working Paper, ETH Zurich, 2015a.

Kirchner, M., Hawkes and INAR(∞) processes. Working Paper, ETH Zurich, 2015b.

Liniger, T., Multivariate Hawkes Processes,. PhD thesis, ETH Zurich, 2009.

Meek, C., Toward learning graphical and causal process models. *Microsoft Research*, 2014.

Ogata, Y., Statistical models for earthquake occurences and residual analysis for point processes. *Journal of the American Statistical Association*, 1988, **83**, 9–27.

Pearl, J., *Causality: Models, Reasoning, and Inference*, Second Edition , 2009, Cambridge University Press.

Song, L., Zha, H. and Zhou, K., Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013, **31**.