Hawkes graphs*

Paul Embrechts, Matthias Kirchner†

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

This version: January 23, 2017

Abstract

This paper introduces the Hawkes skeleton and the Hawkes graph. These objects summarize the branching structure of a multivariate Hawkes point process in a compact, yet meaningful way. We demonstrate how graph-theoretic vocabulary ('ancestor sets', 'parent sets', 'connectivity', 'walks', 'walk weights', . . .) is very convenient for the discussion of multivariate Hawkes processes. For example, we reformulate the classic eigenvalue-based subcriticality criterion of multitype branching processes in graph terms. Next to these more terminological contributions, we show how the graph view may be used 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 specifying and fitting parametric Hawkes models. Our estimation method avoids the a priori assumptions on the model from a straighforward MLE-approach and is numerically more flexible than the latter. Our method has two tuning parameters: one controlling numerical complexity, the other one controlling the sparseness of the estimated graph. A simulation study confirms that the presented procedure works as desired. We pay 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 multivariate 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[1]. We demonstrate how these concepts are 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 computationally beneficial recursive method for the exponential decay case. Liniger (2009) deals especially with the construction of the multivariate and marked case.

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

†Corresponding author: matthias.kirchner@math.ethz.ch

[1] Note that the term 'Hawkes graph' has already been introduced for the graph representation of a specific finite group; see 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.

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 nonzero excitement from event-type i to event-type j. The vertex weights are the corresponding immigration intensities; the weight of an edge (i, j) is the expected number of type-j children events that an type-i event generates. The *Hawkes skeleton* is the Hawkes graph disregarding the weights. The network view on Hawkes processes has been considered in Song et al. (2013), Delattre et al. (2015), Bacry et al. (2015), and Hall and Willett (2016). The graph terminology is convenient to describe many relevant aspects of multivariate Hawkes processes such as 'ancestor and parent sets', 'paths', 'path weights', 'feedback', 'cascades', or 'connectivity'. The graph representation of a Hawkes process also provides additional theoretical insight. For example, in Theorem 1, we give a graph-based criterion for subcriticality which is equivalent to the usual spectral-radius based criterion on the branching matrix. Furthermore, the graph approach turns out to be helpful for the estimation of multivariate 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 reproduction intensities. 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 (2016a). This procedure is based on a limit-representation of the Hawkes process studied in Kirchner (2016b): we discretize the original process and interpret it as an autoregressive model of bin-counts. The latter is statistically estimated using conditional least-squares. In this setup, the asymptotic distribution of the resulting estimators can be obtained. This opens the door to testing. Our procedure is numerically more robust than the standard MLE approach. However, for high-dimensional data our procedure cannot be applied in a straightforward manner either. This is why, in combination with the concept of a Hawkes skeleton and graph, we tackle the numerical difficulties by the following three-step algorithm:

1. Given a large multitype event-stream dataset, 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 edge set. Under the paradigm that the graph of the true underlying multivariate Hawkes model is typically sparse, this estimated superset is still sparse.

2. 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.

3. As a by-product, the calculations in (ii) yield estimates for the values of the nonzero 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: In Section 2, we give definitions and discuss graph attributes that are relevant for the description of multivariate Hawkes processes. In particular, we give results that clarify what kind of information on the Hawkes process a Hawkes graph encodes. In Section 3, we cite earlier results that allow for nonparametric estimation of Hawkes processes. We apply these methods to estimate the Hawkes skeleton and the Hawkes graph. Finally, we show how parametric families for the remaining nonzero reproduction intensities may be specified and calibrated. For an illustration of the new concepts introduced, we present a simulation study in Section 4. In Section 5, we conclude with directions for further research.

2 Definitions

In this section, we recall the branching construction of a multivariate Hawkes process as well as basic graph terminology. After this, we 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 structure. 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 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 the following two conditions hold:*

1. $M(B) \sim \operatorname{Pois}\left(\int_B \lambda(s)\mathrm{d}s\right), \ B \in \mathcal{B}(\mathbb{R})$.
2. *If $B_1, B_2, \dots, B_n \in \mathcal{B}(\mathbb{R})$ with $B_i \cap B_j = \emptyset$, $i \neq j$, then $M(B_1), M(B_2), \dots, M(B_n)$ are mutually independent.*

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

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

A 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**, **G**, and **F** for the corresponding processes.

Definition 2. *Let $d \in \mathbb{N}$ and $[d] := \{1, 2, \dots, d\}$.*

1. *For $(i,j) \in [d]^2$, define* branching coefficients $a_{i,j} \geq 0$, displacement densities $w_{i,j}$ *supported on $\mathbb{R}_{\geq 0}$,* reproduction intensities $h_{i,j} := a_{i,j} w_{i,j}$, *and* reproduction processes $\xi_t^{(i,j)}(\cdot) := \xi^{(i,j)}(\cdot - t) \sim \mathrm{PRM}(h_{i,j}\mathrm{d}s)$, *$t \in \mathbb{R}$, mutually independent over $(i,j,t) \in [d]^2 \times \mathbb{R}$.*
2. *For $i_0 \in [d]$ and $g \in \mathbb{N}_0$, define the g-th* generation process *(generated by a type-i_0 event at time zero) as the d-tuple of random counting measures $\mathbf{G}^{(i_0,g)} := \left(G_1^{(i_0,g)}, \dots, G_d^{(i_0,g)}\right)$ by*

$$G_j^{(i_0,0)}(B) := 1_{\{j=i_0\}} \delta_0(B), \quad B \in \mathcal{B}(\mathbb{R}),\, j \in [d],$$

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

3. *For $i_0 \in [d]$, define the* Hawkes family *(generated by a type-i_0 event at time zero) as the d-tuple of random counting measures*

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

The branching structure of a Hawkes family is encoded in recursion (1). Note that the points of a Hawkes family actually form a *multitype branching random walk*; see Shi (2015). The following definition clarifies how the Hawkes family process is related to the prototypic branching process, the Galton–Watson process:

Definition 3. *For $i_0 \in [d]$, let $\mathbf{F}^{(i_0)}$ be a Hawkes family and let $\{\mathbf{G}^{(i_0,g)}\}_{g\in\mathbb{N}_0}$ be the corresponding generation processes constructed in Definition 2 above. For $g \in \mathbb{N}_0$, define*

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

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

The embedded generation process $(\mathbf{Y}_g^{(i_0)})$ of a Hawkes family is a multitype Galton–Watson process. A multitype Galton–Watson process models the size of a population with individuals of d types, where each individual is alive during exactly one time unit; see Section 2.3 in Haccou et al. (2005). The embedded generation process starts with a single type-i_0 individual in generation 0 and, for $g \in \mathbb{N}$, each type-i individual in generation $g-1$ gives offspring to $\mathrm{Pois}(a_{i,j})$ $a_{i,j} = \int h_{i,j}\mathrm{d}t)$ type-j individuals of type j in generation g. This is why $a_{i,j}$, $(i,j) \in [d]^2$, are called *branching coefficients* and why the matrix $A := (a_{i,j}) \in \mathbb{R}_{\geq 0}$ is called *branching matrix.*

Proposition 1. *Let A be the branching matrix of Hawkes families $\mathbf{F}^{(i_0)}$, $i_0 \in [d]$, respectively, of the corresponding embedded generation processes $(\mathbf{Y}_g^{(i_0)})$, $i_0 \in [d]$. Then we have that*

$$\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}) = \sum_{g\geq 0} \mathbb{E}\, Y_{g,j}^{(i_0)} < \infty, \quad (i_0, j) \in [d]^2, \tag{2}$$

if and only if the spectral radius of A is strictly less than 1. In this case, $(1_{d\times d} - A)$ is invertible and $(\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}))_{(i_0,j)\in[d]^2} = (1_{d\times d} - A)^{-1}$.

Proof. Using

$$\mathbb{E}\,\mathbf{Y}_0^{(i_0)} = \mathbf{Y}_0^{(i_0)} = (0, \ldots, 0, \underbrace{1}_{i_0\text{-th entry}}, 0, \ldots, 0) \text{ and } \mathbb{E}\,\mathbf{Y}_g^{(i_0)} = \mathbb{E}\,\mathbf{Y}_{g-1}^{(i_0)} A,\ g \in \mathbb{N},\ i_0 \in [d],$$

it follows by induction that $(\mathbb{E}\, Y_{j,g}^{(i_0)})_{(i_0,j)\in[d]^2} = A^g$, $g \in \mathbb{N}_0$. By Fubini's theorem, we then get that $(\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}))_{(i_0,j)\in[d]} = \sum_{g\geq 0}(\mathbb{E}\, Y_{g,j}^{(i_0)})_{(i_0,j)\in[d]} = \sum_{g\geq 0} A^g$. Given its entries are finite, the limit matrix $\sum_{g\geq 0} A^g$ is calculated like the limit of a real-valued converging geometric series. The equivalence in Proposition 1 follows from the fact that

$$\sum_{g=0}^{\infty} A^g \text{ converges} \quad \Leftrightarrow \quad \max\left\{|\lambda| : \lambda \text{ eigenvalue of } A\right\} < 1, \quad \text{for } A \in \mathbb{R}^{d\times d}. \tag{3}$$

A detailed proof for (3) can be found in Watson (2015). □

In particular, we get from Proposition 1 that a Hawkes family whose branching matrix satisfies (3) consists of an almost surely finite number of points.

Definition 4. *Let $\mathbf{I} = (I_1, I_2, \ldots, I_d)$ be a* Hawkes immigration process *with $I_{i_0} \sim PRM(\eta_{i_0}\mathrm{d}s)$, $i_0 \in [d]$, independent, where $\eta_{i_0} \geq 0$, $i_0 \in [d]$, are (constant)* immigration intensities. *Furthermore, let $\mathbf{F}_t^{(i_0)}(\cdot) := \mathbf{F}^{(i_0,t)}(\cdot - t)$, $t \in \mathbb{R}$, where $\mathbf{F}^{(i_0,t)}$, $t \in \mathbb{R}$, $i_0 \in [d]$, are independent copies of the generic Hawkes family processes $\mathbf{F}^{(i_0)}$ from Definition 2 above—also independent from the immigration process $\mathbf{I}$. Set*

$$\mathbf{N}(B) := \big(N_1(B), \ldots, N_d(B)\big) := \sum_{i_0=1}^{d} \int_{\mathbb{R}} \mathbf{F}_t^{(i_0)}(B)\, I_{i_0}(\mathrm{d}t), \quad B \in \mathcal{B}(\mathbb{R}).$$

*The d-tuple of random counting measures $\mathbf{N}$ is a d-*type Hawkes process. *If $N_i(\{T\}) = 1$, for some $i \in [d]$, 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 their 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 Definitions 2 and 4, is a stationary solution to the system of implicit equations

$$\begin{aligned}\Lambda_j(t) &:= \lim_{\delta\downarrow 0}\frac{1}{\delta}\,\mathbb{E}\left[N_j\big((t,t+\delta]\big)\Big|\sigma\Big(\mathbf{N}\left((a,b]\right),\, a<b\leq t\Big)\right]\\ &= \eta_j+\sum_{i=1}^{d}\int\limits_{-\infty}^{t} h_{i,j}(t-s)N_i\left(\mathrm{d}s\right), \quad t\in\mathbb{R},\, j\in[d]. \qquad (4)\end{aligned}$$

We call $\boldsymbol{\Lambda}(t) := (\Lambda_1(t),\Lambda_2(t),\ldots,\Lambda_d(t))$ the *conditional intensity* of $\mathbf{N}$. In terms of intensities, the value of a reproduction intensity 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 introductions (Hawkes, 1971a,b) and also including (Kirchner, 2016a), 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* (4) *with matrix multiplication. In the present graph-driven work, '$a_{i,j}$', '$w_{i,j}$', '$h_{i,j}$', 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; see Pearl (2009) for a recent overview. The notion of causality is subtle. For Hawkes processes, however, the use of the term seems justified. Indeed, 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* $\mathrm{Pois}(a_{i,j})$ 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 on 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 can for instance 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 5. *Let $d\in\mathbb{N}$ and $[d]=\{1,2,\ldots,d\}$. A* graph *$\mathcal{G}$ is a 2-tuple $(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=[d]$ is a set of* vertices *and $\mathcal{E}\subset\mathcal{V}\times\mathcal{V}$ is a set of* edges*. Given such a graph $\mathcal{G}$ we introduce the following definitions:*

i) Vertex i is a parent *of vertex j if $(i,j)\in\mathcal{E}$. We write $\mathrm{PA}(j):=\{i:\ (i,j)\in\mathcal{E}\}$. Vertex i is a* source vertex *if $\mathrm{PA}(i)\setminus\{i\}=\emptyset$. Vertex i is a* sink vertex *if $\{j:\ (i,j)\in\mathcal{E}\}\setminus\{i\}=\emptyset$.*

ii) For $g\in\mathbb{N}$, $(k_0,k_1,\ldots,k_g)\in\mathcal{V}^{g+1}$ is a walk *in $\mathcal{G}$ of length g from vertex i to vertex j if $k_0=i, k_g=j$ and $(k_{l-1},k_l)\in\mathcal{E}$, $l\in[g]$; $(k_0,k_1,\ldots,k_g)\in\mathcal{V}^{g+1}$ is a* closed walk *if it is a walk with $k_0=k_g$. We denote the* set of walks in $\mathcal{G}$ from i to j with length $g\in\mathbb{N}$ *by $\mathcal{W}_g^{(i,j)}$. Furthermore, we set $\mathcal{W}_0^{(i,j)}:=\emptyset$ if $i\neq j$, $\mathcal{W}_0^{(i,j)}:=\{(i)\}$ if $i=j$, $\mathcal{W}^{(i,j)}:=\cup_{g\geq 0}\mathcal{W}_g^{(i,j)}$, and $\mathcal{W}:=\cup_{(i,j)\in[d]^2}\mathcal{W}^{(i,j)}$.*

iii) Vertex i is an ancestor *of j if there exists a walk of length $g\in\mathbb{N}$ from i to j. We denote the* ancestor set *of a vertex i in $\mathcal{G}$ by $\mathrm{AN}(i)$.*

iv) The vertices i and j are weakly connected *if $i=j$ or if there exists a set $\{(k_{l-1},k_l),\, l=1,\ldots,g:\ k_0=i, k_g=j, (k_l,k_{l-1})\in\mathcal{E}\ \text{ or }\ (k_{l-1},k_l)\in\mathcal{E}\}$ for some $g\in\mathbb{N}$. The vertices i and j are* strongly connected *if the sets $\mathcal{W}^{i,j}$ and $\mathcal{W}^{j,i}$ are nonempty. A graph is* weakly (strongly) connected *if all pairs of its vertices are weakly (strongly) connected. A graph is* fully connected *if $(i,j)\in\mathcal{E},\ (i,j)\in[d]^2$.*

Note that in our definition, a graph allows cycles and, in particular, self-loops. A vertex may or may not be an ancestor and, in particular, a parent of itself. Also note that any vertex i is always strongly connected to itself because $\{(i)\}\subset\mathcal{W}^{(i,i)}$, $i\in[d]$—no matter if i is contained in a closed walk or not. Consequently, the singleton graph is always strongly connected. However, it is only fully connected if is a self-loop. Next, we apply the graph terminology from Definition 5 to the Hawkes process:

Definition 6. *Let* $\mathbf{N}$ *be a* d*-type Hawkes process with immigration intensities* $\eta_1, \eta_2, \ldots, \eta_d$ *and branching coefficients* $a_{i,j} (= \int h_{i,j}(t)\mathrm{d}t)$, $(i,j) \in [d]^2$*; see Definitions 2 and 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* $\mathcal{V}^*_{\mathbf{N}} = [d]$ *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 $j \in [d]$*, we denote the* parent*, respectively,* ancestor set *of* j *with respect to the Hawkes skeleton* $\mathcal{G}^*_{\mathbf{N}}$ *by* $\mathrm{PA}_{\mathbf{N}}(j)$ *and* $\mathrm{AN}_{\mathbf{N}}(j)$*. For the* Hawkes graph $\mathcal{G}_{\mathbf{N}} = (\mathcal{V}_{\mathbf{N}}, \mathcal{E}_{\mathbf{N}})$ *of* $\mathbf{N}$*, each vertex, respectively, edge of the corresponding Hawkes skeleton is supplied with a* vertex*, respectively, an* edge weight*:*

$$\begin{aligned}\mathcal{V}_{\mathbf{N}} &:= \Big\{(j;\eta_j) : \quad j \in \mathcal{V}^*_{\mathbf{N}} \text{ and } \eta_j \text{ is the } j\text{-}th \text{ immigration intensity of } \mathbf{N}\Big\},\\ \mathcal{E}_{\mathbf{N}} &:= \Bigg\{(i,j;a_{i,j}) : \quad (i,j) \in \mathcal{E}^*_{\mathbf{N}} \textit{ and } (a_{i,j})_{(i,j)\in[d]^2} \textit{ is the branching matrix of } \mathbf{N}\Bigg\}.\end{aligned}$$

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

i) A Hawkes graph $\mathcal{G}_{\mathbf{N}}$ *is* weakly, strongly, respectively, fully connected *if the corresponding skeleton* $\mathcal{G}^*_{\mathbf{N}}$ *is weakly, strongly, respectively, fully connected; see Definition 5.*

ii) Vertex $(j;\eta_j)$ *of a Hawkes graph* $\mathcal{G}_{\mathbf{N}}$ *is a* source*, respectively,* sink vertex*, if it is a source, respectively, sink vertex in the corresponding skeleton* $\mathcal{G}^*_{\mathbf{N}}$*. Furthermore,* $(j;\eta_j)$ *is a* redundant vertex *if* $\eta_j = 0$ *and, in addition,* $\eta_i = 0$ *for all* $i \in \mathrm{AN}_{\mathbf{N}}(j)$*.*

iii) For any walk $w \in \mathcal{W}_{\mathcal{G}_{\mathbf{N}}}$ $(:= \mathcal{W}_{\mathcal{G}^*_{\mathbf{N}}})$ *in a Hawkes graph* $\mathcal{G}_{\mathbf{N}}$*, we define the* walk weights

$$|w| = |(i_0, i_1, \ldots, i_g)| := \begin{cases} 1, & g = 0, \ \textit{and}\\ \prod_{l=1}^{g} a_{i_{l-1},i_l}, & g > 0,\end{cases}$$

where a_{i_{l-1},i_l}, $l = 1, 2, \ldots, g$*, are the edge weights from* $\mathcal{E}_{\mathbf{N}}$*.*

iv) A Hawkes graph is subcritical *if*

$$\sum_{w\in\mathcal{W}^{(i_0,i_0)}} |w| < \infty, \ i_0 \in [d], \textit{ or, equivalently,} \qquad \sum_{\substack{w:\\ w \textit{ closed walk in } \mathcal{G}_{\mathbf{N}}}} |w| < \infty. \tag{5}$$

Note that if a Hawkes graph vertex is redundant, then all its ancestors are also redundant. The notion of a subcritical Hawkes graph in Definition 6 iv) might ask for further explanation. The following theorem clarifies things:

Theorem 1. *Let* $\mathbf{N}$ *be a Hawkes process and let* $\mathcal{G}_{\mathbf{N}}$ *be the corresponding Hawkes graph. Then* $\mathbf{N}$ *is a subcritical Hawkes process (see Definition 4) if and only if* $\mathcal{G}_{\mathbf{N}}$ *is a subcritical Hawkes graph (see Definition 6).*

Proof. First, we prove that

$$\sum_{w\in\mathcal{W}^{(i_0,i_0)}} |w| < \infty, \ i_0 \in [d] \quad \Leftrightarrow \quad \sum_{w\in\mathcal{W}^{(i_0,j)}} |w| < \infty, \ (i_0,j) \in [d]^2. \tag{6}$$

'$\Leftarrow$' is trivial. We show '$\Rightarrow$' by induction over the graph size d: for $d = 1$, the implication is true. For $d > 1$, consider a graph with d vertices and assume that the left-hand side of (6) holds. Pick any $(i_0, j) \in [d]^2$. We split the possible paths from i_0 to j, $\mathcal{W}^{(i_0,j)}$, into *paths excluding* d, $\mathcal{W}^{(i_0,j)}_{\mathrm{excl}.d}$, and *paths including* d, $\mathcal{W}^{(i_0,j)}_{\mathrm{incl}.d}$:

$$\sum_{w\in\mathcal{W}^{(i_0,j)}} |w| = \sum_{w\in\mathcal{W}^{(i_0,j)}_{\mathrm{excl}.d}} |w| + \sum_{w\in\mathcal{W}^{(i_0,j)}_{\mathrm{incl}.d}} |w|. \tag{7}$$

The first sum is finite by the induction hypothesis. Now, assume the case that $i_0 \neq d$ and $j \neq d$. Every walk in the second sum of (7) may be (uniquely) split into the following five subwalks: a d-avoiding walk w_1 from i_0 to some $i_1 \in \mathrm{PA}(d)$, a one-step walk (i_1, d), a walk $w_2 \in \mathcal{W}^{(d,d)}$, another one-step walk (d, j_1), with $d \in \mathrm{PA}(j_1)$, and finally some d-avoiding walk w_3 from j_1 to j. This yields

$$\begin{aligned}
&\sum_{w\in\mathcal{W}^{(i_0,j)}_{\mathrm{incl.}d}} |w| \\
&\quad= \sum_{i_1\in\mathrm{PA}(d)} \sum_{w_1\in\mathcal{W}^{(i_0,i_1)}_{\mathrm{excl.}d}} \sum_{w_2\in\mathcal{W}^{(d,d)}} \sum_{j_1:d\in\mathrm{PA}(j_1)} \sum_{w_3\in\mathcal{W}^{(j_1,j)}_{\mathrm{excl.}d}} |w_1|\, a_{i_1,d}\, |w_2|\, a_{d,j_1}\, |w_3| \\
&\quad\leq \sum_{i_1\in\mathrm{PA}(d)} \sum_{j_1:d\in\mathrm{PA}(j_1)} \max_{(i,j)\in[d]^2} a^2_{i,j} \underbrace{\sum_{w_1\in\mathcal{W}^{(i_0,i_1)}_{\mathrm{excl.}d}} |w_1|}_{<\infty \text{ by ind. hyp.}} \underbrace{\sum_{w_2\in\mathcal{W}^{(d,d)}} |w_2|}_{<\infty \text{ by assumption}} \underbrace{\sum_{w_3\in\mathcal{W}^{(j_1,j)}_{\mathrm{excl.}d}} |w_3|}_{<\infty \text{ by ind. hyp.}} < \infty.
\end{aligned}$$

Note that, by definition, $(i) \in \mathcal{W}^{(i,i)}$ and $|(i)| = 1$, $i \in [d]$, so that the calculation above also covers the cases $\mathrm{PA}(d) = \{i_0\}$ and $\mathrm{PA}(j) = \{d\}$ as well as d-including walks from i to j that touch d exactly once. If $i_0 = d$ or $j = d$, the splitting argument becomes even simpler; we do not give the details. We have proven the finiteness of the second sum in (7) and therefore (6).

Next, note that

$$\sum_{w\in\mathcal{W}^{(i_0,j)}} |w| = \sum_{g\geq 0} \sum_{w\in\mathcal{W}^{(i_0,j)}_g} |w| = \sum_{g\geq 0} \mathbb{E}\, Y^{(i_0)}_{g,j} = \mathbb{E}\, F^{(i_0)}_j(\mathbb{R}), \quad (i_0, j) \in [d]^2, \tag{8}$$

where $(\mathbf{Y}^{(i_0)}_g) = (Y^{(i_0)}_{g,1}, Y^{(i_0)}_{g,2}, \ldots, Y^{(i_0)}_{g,d})$ are the embedded generation processes of the generic family processes $\mathbf{F}^{(i_0)} = (F^{(i_0)}_1, \ldots, F^{(i_0)}_d)$ of $\mathbf{N}$; see Definition 3. Thus, (5) is a complicated way of saying that, for all $(i_0, j) \in [d]^2$, the expected total number of type-j offspring events of a type-i_0 event is finite, i.e., that $\mathbb{E}\, F^{(i_0)}_j(\mathbb{R}) < \infty$, $(i_0, j) \in [d]^2$. By Proposition 1, this in turn is equivalent to the spectral radius of the branching matrix being strictly less than 1—which is the original Hawkes-*process* subcriticality condition from Definition 4 □

Obviously, the Hawkes graph does not fully specify the corresponding Hawkes process; it only captures the structure of the embedded generation processes from Definition 3 together with the immigration intensities. Despite this simplification, the Hawkes graph gives relevant insight into the underlying Hawkes process—especially in the highdimensional case. For example, *connectivity and redundancy of vertices* are two graph-based concepts that become increasingly important the higher the dimension of the model considered is. If a Hawkes graph is not weakly connected, we may consider the *weakly connected* subgraphs separately and correspondingly split the original model into separate, lower-dimensional Hawkes processes. The notion of *redundant vertices* is important because, typically, we only want to consider 'accessible' event types. *Sink* (*source*) *vertices* of a Hawkes graph correspond to Hawkes process components that only receive (give) excitement from (to) the system. The notion of *parent sets* is also helpful: e.g., for the marginal conditional intensity in (4), it is actually enough to sum over $i \in \mathrm{PA}(j)$ instead of $i \in [d]$ which may be computationally beneficial. The *ancestor sets* may be applied if we are only interested in modeling events of a particular type j. In this situation, it suffices to consider a Hawkes model for the event types in $\{j\} \cup \mathrm{AN}(j)$. Finally, we find the formulation of Hawkes graph *subcriticality* in (5) useful. It provides a more concrete meaning to the somewhat abstract eigenvalue-based criterion for the Hawkes process. E.g., (5) can be used when constructing subcritical Hawkes graphs, respectively, models. And—if a given graph is sparse and the closed walks are not too numerous—one can check subcriticality without even calculating any eigenvalue; see Section 4.1. Furthermore, in some cases, the *path weights* $|w|$ themselves might be worth calculating—even apart from criticality conditions; see the discussion in the proof of Theorem 1. 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 graphical and statistical information that may be calculated from the adjacency matrix of a Hawkes graph:

Proposition 2. *For some $d \geq 2$, let $\mathbf{N}$ be a d-type subcritical Hawkes process. Furthermore, let $\mathcal{G}_{\mathbf{N}} = (\mathcal{V}_{\mathbf{N}}, \mathcal{E}_{\mathbf{N}})$ be the corresponding Hawkes graph with adjacency matrix $A = (a_{i,j}) \in \mathbb{R}_{\geq 0}^{d \times d}$. Then we have that*

i) $a_{i,j} > 0 \quad \Leftrightarrow \quad i \in \mathrm{PA}_{\mathbf{N}}(j)$;

ii) $a_{i,j} = 0$, $j \in [d] \setminus \{i\} \quad \Leftrightarrow$ *vertex i is a sink vertex;*

iii) $a_{i,j} = 0$, $i \in [d] \setminus \{j\} \quad \Leftrightarrow$ *vertex j is a source vertex;*

iv) $(A^g)_{i,j} > 0 \quad \Leftrightarrow$ *there is a walk of length g from i to j;*

v) $(A^g)_{i,j} > 0$ *for some* $g \in [d] \quad \Leftrightarrow \quad i \in \mathrm{AN}(j)$;

vi) for all $(i,j) \in [d]^2$, $((A + A^\top)^g)_{i,j} > 0$ for some $g \in \{0\} \cup [d-1]$ $\Leftrightarrow$ the Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is weakly connected;

vii) for all $(i,j) \in [d]^2$, $((A)^g)_{i,j} > 0$ for some $g \in \{0\} \cup [d-1]$ $\Leftrightarrow$ the Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is strongly connected;

viii) $a_{i,j} > 0$, $(i,j) \in [d]^2 \quad \Leftrightarrow$ *the Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is fully connected;*

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

Definition 7. *Let $\mathbf{N}$ be a subcritical d-type Hawkes process and let A be the adjacency matrix of the corresponding Hawkes graph $\mathcal{G}_{\mathbf{N}}$. Consider the limit matrix $\mathbb{R}_{\geq 0}^{d \times d} \ni (e_{i,j}) := (1_{d \times d} - A)^{-1} = \sum_{g \geq 0} A^g \, (= (\mathbb{E}\, F_j^{(i_0)}(\mathbb{R}))_{(i_0,j) \in [d]^2})$ from Proposition 1 and define*

$$c_{i_0} := \frac{\eta_{i_0} \sum_{j=1}^d e_{i_0,j}}{\sum_{i=1}^d \eta_i \sum_{j=1}^d e_{i,j}}, \quad i_0 \in [d], \quad \text{and} \quad f_j := \frac{\eta_j e_{j,j}}{\sum_{i=1}^d \eta_i e_{i,j}}, \quad j \in [d].$$

We call $(c_{i_0})_{i \in [d]}$ the cascade coefficients *and $(f_j)_{j \in [d]}$ the* feedback coefficients.

One way of tuning a specific Hawkes graph may be achieved by 'switching-off' a selected vertex by forcing the corresponding immigration intensity to zero. The coefficients defined above summarize the effect of such a manipulation. In view of Proposition 1, we have the following interpretations. First of all, the *cascade coefficients* (c_i) are important from a *systemic* point of view. The cascade coefficient c_i measures the fraction of events in the system stemming from families with immigrated type-i ancestor. If $c_i > 1/d$, this indicates a relatively large impact of type-i events on the system. Secondly, the *feedback coefficients* (f_j) are more important from an *individual* point of view. They indicate how much of the total intensity that a vertex j *experiences* is due to its own immigration activity including the feedback it experiences by closed walks. 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 (Kirchner, 2016b), 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). 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 (2016a). 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 $X_n^{(\Delta)} := N\big(((n-1)\Delta, n\Delta]\big)$, $k = 1,2,\dots,n := \lfloor T/\Delta \rfloor$, and some $p \in \mathbb{N}$, large, we calculate

$$\left(\hat\alpha_0^{(\Delta)}, \hat\alpha_1^{(\Delta)}, \dots, \hat\alpha_p^{(\Delta)}\right) := \mathrm{argmin}_{\left(\alpha_0^{(\Delta)}, \alpha_1^{(\Delta)}, \dots, \alpha_p^{(\Delta)}\right)} \sum_{k=p+1}^{n} \left(X_k^{(\Delta)} - \alpha_0^{(\Delta)} - \sum_{l=1}^{p} \alpha_l^{(\Delta)} X_{k-l}^{(\Delta)} \right)^2 . \quad (9)$$

Given (9), we estimate the reproduction-intensity values $h(k\Delta)$, $k = 1,2,\dots,p$, 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 (2016a)—modulo transposition as stated in Remark 1.

Definition 8. *Let $\mathbf{N} = (N_1, N_2, \dots, N_d)$ be a subcritical d-type Hawkes process with immigration intensity $\eta \in \mathbb{R}^d_{\geq 0} \setminus \{0_d\}$ and reproduction intensities $h_{i,j} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $(i,j) \in [d]^2$. 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:*

$$\mathbf{X}_k^{(\Delta)} := \mathbf{N}\left(\big((k-1)\Delta, k\Delta\big]\right)^{\top} \in \mathbb{N}_0^{d\times 1}, \quad k = 1,2,\dots,n := \lfloor T/\Delta \rfloor . \quad (10)$$

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

$$\widehat{\mathbf{H}}^{(\Delta,s)} := \frac{1}{\Delta}\left(\mathbf{Z}^{\top}\mathbf{Z}\right)^{-1}\mathbf{Z}^{\top}\mathbf{Y} \quad \in \mathbb{R}^{(dp+1)\times d} . \quad (11)$$

Here,

$$\mathbf{Z}\left(\mathbf{X}_1^{(\Delta)}, \dots, \mathbf{X}_n^{(\Delta)}\right) := \begin{pmatrix} (\mathbf{X}_p^{(\Delta)})^{\top} & (\mathbf{X}_{p-1}^{(\Delta)})^{\top} & \dots & (\mathbf{X}_1^{(\Delta)})^{\top} & 1 \\ (\mathbf{X}_{p+1}^{(\Delta)})^{\top} & (\mathbf{X}_p^{(\Delta)})^{\top} & \dots & (\mathbf{X}_2^{(\Delta)})^{\top} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ (\mathbf{X}_{n-1}^{(\Delta)})^{\top} & (\mathbf{X}_{n-2}^{(\Delta)})^{\top} & \dots & (\mathbf{X}_{n-p}^{(\Delta)})^{\top} & 1 \end{pmatrix} \in \mathbb{R}^{(n-p)\times(dp+1)} \quad (12)$$

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

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 (11) actually estimate:

$$\begin{pmatrix} \widehat{H}_1 \\ \dots \\ \widehat{H}_p \\ \hat\eta \end{pmatrix} := \widehat{\mathbf{H}} \in \mathbb{R}^{(dp+1)\times d}, \quad \text{where} \quad \widehat{H}_k := \begin{pmatrix} \hat h_{1,1}(k\Delta) & \hat h_{1,2}(k\Delta) & \dots & \hat h_{1,d}(k\Delta) \\ \hat h_{2,1}(k\Delta) & \hat h_{2,2}(k\Delta) & \dots & \hat h_{2,d}(k\Delta) \\ \dots & \dots & \dots & \dots \\ \hat h_{d,1}(k\Delta) & \hat h_{d,2}(k\Delta) & \dots & \hat h_{d,d}(k\Delta) \end{pmatrix} . \quad (13)$$

In Kirchner (2016a), 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 $\mathrm{vec}\left(\widehat{\mathbf{H}}^{\top}\right) \in \mathbb{R}^{d(dp+1)}$ ($\mathrm{vec}(\cdot)$ stacks the columns of its argument) can be consistently estimated by

$$\widehat{S^2} := \frac{1}{\Delta^2}\left(\left(\mathbf{Z}^{\top}\mathbf{Z}\right)^{-1} \otimes 1_{d\times d}\right)\mathbf{W}\left(\left(\mathbf{Z}^{\top}\mathbf{Z}\right)^{-1} \otimes 1_{d\times d}\right) \in \mathbb{R}^{d(dp+1)\times d(dp+1)} . \quad (14)$$

Here, $\otimes$ denotes the Kronecker product, $\mathbf{Z}$ is the design matrix from (12) 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 $k = p+1, p+2, \dots, n$,

$$\begin{aligned} \mathbf{w}_k \quad := \quad & \left(\left(\left(\mathbf{X}_{k-1}^{(\Delta)}\right)^{\top}, \left(\mathbf{X}_{k-2}^{(\Delta)}\right)^{\top}, \dots, \left(\mathbf{X}_{k-p}^{(\Delta)}\right)^{\top}, 1\right)^{\top} \otimes 1_{d\times d}\right) \quad (15) \\ & \cdot \left(\mathbf{X}_k^{(\Delta)} - \Delta\hat\eta - \sum_{l=1}^{p} \Delta \widehat{H}_l^{\top}\mathbf{X}_{k-l}^{(\Delta)}\right) \in \mathbb{R}^{d(dp+1)\times 1} . \end{aligned}$$

In Definition 8, we consider $\mathrm{vec}(\mathbf{H}^\top)$ instead of $\mathrm{vec}(\mathbf{H})$ in order to apply the results from Kirchner (2016b) more directly; see Remark 1. We will discuss below how one retrieves specific values from the covariance matrix estimation in (14). The estimator from Definition 8 above depends on a support s, $0 < s << T$, and on a bin size Δ, $0 < \Delta \leq s$. Automatic methods for the choice of these estimation parameters are discussed in Kirchner (2016b). In the present paper, we assume s given. Often, an upper bound for the support of the reproduction intensities can be guessed from the data context. The choice of Δ, however, will be crucial in high-dimensional settings. 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 6. The idea is simple: for $(i, j) \in [d]^2$, 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 (13) for the notation. Calculating the covariance estimate (14), 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 8 and discuss numerical issues.

Definition 9. *Given d-type event-stream data on $(0, T]$, calculate the Hawkes estimator $\mathbf{H}^{(\Delta_{skel},s)}$ from Definition 8 with respect to some s, $0 < s < T$, and some Δ_{skel}, $0 < \Delta_{skel} \leq s$. For $j \in [d]$, let $b_j \in \{0, 1\}^{(dp+1)\times 1}$ be column vectors with all entries zero but 1s at entries $(k - 1)d + j$, $k = 1, 2, \dots, p = \lceil s/\Delta_{skel}\rceil$. Let $B := (b_1, b_2, \dots, b_d)^\top$, and calculate*

$$(\hat{a}_{i,j})_{1\leq i,j\leq d} = \Delta_{skel} B\, \mathbf{H}^{(\Delta_{skel},s)} \,. \tag{16}$$

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

$$\widehat{\mathcal{E}}^* := \Big\{(i, j) \in [d]^2 :\ \hat{a}_{i,j} > \hat{\sigma}_{i,j} z^{-1}_{1-\alpha_{skel}}\Big\}. \tag{17}$$

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

The main point of this first estimation step is that we hope that the edge set $|\mathcal{E}^*|$ and, consequently $|\widehat{\mathcal{E}}^*|$ are typically much smaller than d^2, respectively, that $\mathrm{PA}_{\mathbf{N}}(j)$, $j \in [d]$, and, consequently, $\widehat{\mathrm{PA}}_{\mathbf{N}}(j)$, $j \in [d]$, 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 Δ_{skel}

On the one hand, the smaller we choose the bin size Δ, 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 8 become increasingly large when Δ decreases. More specifically, (11) involves the construction and multiplication of matrices with about ds/Δ rows and about T/Δ columns, where $T > 0$ denotes the sample window size, $d \in \mathbb{N}$ the number of event-types, and s, $\Delta \leq s << T$, the support parameter from Definition 8. 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 test involved in (17) does not depend on Δ_{skel} too heavily. The false positive rate (that is, the probability of *including a false edge*) is well controlled by α_{skel}, because, under $H_0 : h_{i,j} \equiv 0$, discretizations as in (9) stay 'correct' even for very coarse Δ_{skel}; see (18) below. The false negative rate (probability of *missing a true edge*) naturally depends strongly on the true underlying edge weights. However, if there is truly considerable direct excitement from one component to another, then typically the effect from some bin to future bins will also be of some significance—which is exactly what our skeleton estimator tests. Our simulation study in Section 4.2 confirms these arguments.

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 (hoped-for) 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 a somewhat rough discretization in the first (skeleton-estimation) step.

By choosing $\Delta_{\mathrm{skel}} = s/k$ for some small $k \in \mathbb{N}$ in the calculations of Definition 9 above, even Hawkes-skeleton estimates of very high-dimensional models (such as $d > 20$) become computationally tractable.

The role of $\alpha_{\mathbf{skel}}$

Note that under $H_0 : \ a_{i,j} \equiv 0$, we have that

$$\mathbb{P}_{H_0}[\hat{a}_{i,j} > \hat{\sigma}^2_{i,j} z^{-1}_{1-\alpha_{\mathrm{skel}}}] \approx \alpha_{\mathrm{skel}}. \tag{18}$$

Still, the parameter $\alpha_{\mathrm{skel}} \in (0, 1)$ should not so much be thought of as an actual significance level—due to the multiple testing setup over $(i, j) \in [d]^2$, and because of the dependence between the different edge tests. Despite this warning, note that in the simulation study from Section 4.2, the corresponding empirical false positive rates are very close to our (varying) choices of α_{skel}. In any case, α_{skel} is a flexible tuning parameter that allows for controlling the degree of sparseness in the estimated graph. A value of $\alpha_{\mathrm{skel}} = 1$ will yield a fully connected estimated graph as Hawkes skeleton. When α_{skel} decreases, the skeleton estimate becomes sparser and sparser. For $\alpha_{\mathrm{skel}} \geq 0.01$, we typically still *overestimate* the true edge set. In other words, for $j \in [d]$, we typically have that $\mathrm{PA}_{\mathbf{N}}(j) \subset \widehat{\mathrm{PA}}_{\mathbf{N}}(j)$ with high probability.

Variance estimate calculation

The most elaborate step from a computational point of view in Definition 8 is the calculation of the covariance estimator in (14). 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 9, we simplify the calculation. First of all, we note that in the matrix $\widehat{S}^2$ from (14), 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 (14) 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 $(i, j) = [d]^2$, in row $(i-1)d+j$ we have 1s at entries $(k-1)d^2 + (i-1)d + j$, $k = 1, 2, \ldots, p$. Let $\mathbf{E}_{l,\cdot}$ denote the l-th row of $\mathbf{E}$. With $\widehat{S}^2$ from (14) and for $(i, j) \in [d]^2$, 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 (16). These estimates can be computed in the following way:*

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

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

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

iii) Multiply $\mathbf{C}$ from (i) pointwise with $\mathbf{U}^{(rep)}$ 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 9, we consider the estimation of the Hawkes graph $\mathcal{G}_{\mathbf{N}}$; see Definition 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 skeleton $\widehat{\mathcal{G}}_{\mathbf{N}}^*$ into its weakly-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 and 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 (9), *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 all of the d other components in the conditional-least-squares regression increases numerical efficiency and decreases estimation variance. 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 the intensity at time t of 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 8. This time however, due to reduction of dimensionality, we will typically *be able to choose a much smaller bin size* Δ_{graph} 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 $k = 1, 2, \ldots, n$, $j \in [d]$ and some $0 < \Delta_{\mathrm{graph}} << \Delta_{\mathrm{skel}}$, let $\mathbf{X}_{k,j}^{(\Delta_{\mathrm{graph}})} := N_j\big(((k-1)\Delta_{\mathrm{graph}}, k\Delta_{\mathrm{graph}}]\big)$, $d_j := |\mathrm{PA}(j)|$, and

$$\mathbf{X}_{k,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})} := \Big(\mathbf{X}_{k,i_1}^{(\Delta_{\mathrm{graph}})}, \mathbf{X}_{k,i_2}^{(\Delta_{\mathrm{graph}})}, \ldots, \mathbf{X}_{k,i_{d_j}}^{(\Delta_{\mathrm{graph}})}\Big)^\top. \tag{19}$$

In (19) and in what follows, 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 the bin counts of all d event types separately on the past of their parents with Ansatz

$$\mathbb{E}\Big[\mathbf{X}_{n,j}^{(\Delta_{\mathrm{graph}})} \Big| \mathbf{X}_{n-k,\mathrm{PA}(j)}^{(\Delta_{\mathrm{graph}})}, k = 1, 2, \ldots, p\Big] = \alpha_{0,j}^{(\Delta_{\mathrm{graph}})} + \sum_{i \in \mathrm{PA}(j)} \sum_{k=1}^{p} \alpha_{k,i,j}^{(\Delta_{\mathrm{graph}})} \mathbf{X}_{n-k,i}^{(\Delta_{\mathrm{graph}})}, \quad j \in [d]. \tag{20}$$

Ansatz (20) should be compared with (9). Note that j itself may or may not be an element of $\mathrm{PA}(j)$.

Definition 10. *Let $\mathcal{G}_{\mathbf{N}}^*$ be a Hawkes skeleton (estimate) with respect to some d-type Hawkes process (data) $\mathbf{N}$. Given $d_j := |\mathrm{PA}(j)|$, $j \in [d]$, a bin size $\Delta_{graph} > 0$, a support s with $0 < \Delta_{graph} \leq s < T$, and $p := \lceil s/\Delta_{graph} \rceil$, calculate the conditional-least-squares estimates*

$$\widehat{\mathbf{H}}_j^{(\Delta_{graph},s)} := \frac{1}{\Delta_{graph}} \Big(\mathbf{Z}_j^\top \mathbf{Z}_j\Big)^{-1} \mathbf{Z}_j^\top \mathbf{Y}_j \in \mathbb{R}^{(pd_j+1)\times 1}, \quad j \in [d_j], \tag{21}$$

with design matrices

$$\mathbf{Z}_j := \begin{pmatrix} (\mathbf{X}_{p,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & (\mathbf{X}_{p-1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & \ldots & (\mathbf{X}_{1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & 1 \\ (\mathbf{X}_{p+1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & (\mathbf{X}_{p,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & \ldots & (\mathbf{X}_{2,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (\mathbf{X}_{n-1,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & (\mathbf{X}_{n-2,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & \ldots & (\mathbf{X}_{n-p,\mathrm{PA}(j)}^{(\Delta_{graph})})^\top & 1 \end{pmatrix} \in \mathbb{N}_0^{(n-p)\times(pd_j+1)}, \quad j \in [d], \tag{22}$$

and vectors of responses

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

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

$$\widehat{\mathcal{E}}_{\mathbf{N}} := \bigcup_{j=1,\dots,d} \Big\{(i_l, j; \hat{a}_{i_l,j}): \ \{i_1,\dots,i_{d_j}\} = \mathrm{PA}(j),\, \hat{a}_{i_l,j} = b_{l,j}^\top \widehat{\mathbf{H}}_j^{(\Delta_{graph},s)}\Big\}, \tag{23}$$

where, for $l \in [d_j]$*,* $b(l,j) \in \{0,1\}^{(d_j p+1)\times 1}$ *is a column vector with* 0*s in all components but 1s in components* $((k-1)d_j + l)$*,* $k = 1,2,\dots,p$*. Furthermore, for* $\alpha_{graph} \in (0,1)$*, we define the confidence intervals* $\big[\hat{\eta}_j \pm \hat{\sigma}_j z^{-1}_{1-\alpha_{graph}}\big)$*,* $j \in [d]$*, and, for* $i_l \in \mathrm{PA}_{\mathbf{N}}(j)$*,* $\big[\hat{a}_{i_l,j} \pm \hat{\sigma}_{i_l,j} z^{-1}(1-\alpha_{graph})\big)$*. 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 matrices $\widehat{\mathbf{H}}_j^{(\Delta_{\mathrm{graph}},s)}$, $j \in [d]$, actually estimate:

$$\begin{pmatrix} \widehat{H}_{\mathrm{PA}(j),j}(\Delta_{\mathrm{graph}}) \\ \widehat{H}_{\mathrm{PA}(j),j}(2\Delta_{\mathrm{graph}}) \\ \dots \\ \widehat{H}_{\mathrm{PA}(j),j}(p\Delta_{\mathrm{graph}}) \\ \hat{\eta}_j \end{pmatrix} := \widehat{\mathbf{H}}_j, \text{ with} \tag{24}$$

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

$k = 1,2,\dots,p$ and $\{i_1, i_2, \dots, i_{d_j}\} = \mathrm{PA}(j)$. 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* $j \in [d]$ *such that* $|\mathrm{PA}(j)| > 0$ *and let* $\{i_1, i_2, \dots, i_{d_j}\} = \mathrm{PA}(j)$ *with* $i_1 < i_2 < \dots, < i_{d_j}$*. For* (i_l, j)*,* $l \in [d_j]$*, let* $e(i_l, j) \in \{0,1\}^{(d_j p+1)\times 1}$ *be a column vector with all entries 0, 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 $\mathbf{C}_{l,j} := e(i_l,j)^\top \Big((\mathbf{Z}_j^\top \mathbf{Z}_j)^{-1} \mathbf{Z}_j^\top\Big) \in \mathbb{R}^{1\times(n-p)}$.

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

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

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

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

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

Remark 2. *The bin size* Δ_{graph} *for the graph estimation in Definition 10 will typically be much smaller than the bin size* Δ_{skel} *for the skeleton estimation in Definition 9. After the graph estimation, one might again want to delete edges with edge-weight estimates non-significantly different from zero, or treat vertex-weight estimates, respectively, immigration intensities, that are not 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 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 ought to increase* α_{skel}*,* α_{graph}*, and/or* α_{vertex}*. Thus, we obtain more estimated nonzero 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 10, 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 reproduction intensities

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 displacement intensities, respectively, the reproduction intensities from Definition 2.

Parametric estimation

Given the Hawkes estimator from Definition 8, the Hawkes model is not yet completely specified. In particular, (21) only yields estimates of the reproduction intensities on a grid:

$$\left\{\Big(k\Delta\Big), \hat{h}_{i,j}(k\Delta)\right\}_{k=1,2,\dots,p}, \quad i \in \widehat{\mathrm{PA}}(j),\ j \in [d]. \tag{25}$$

One obvious possibility to complete the model specification would be the application of any kind of smoothing method on (25). We want to consider another approach: we exploit (25) 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 (25) via non-linear least-squares (e.g., function `nls` in `R`):

Definition 11. *Consider a Hawkes-graph estimation as in Definition 10 with respect to some d-type event-stream data and a bin size $\Delta_{graph} > 0$. For $j \in [d]$ and $i \in \widehat{\mathrm{PA}}(j)$, let $w_{i,j}^{(\theta_{i,j})} : \mathbb{R} \to \mathbb{R}_{\geq 0}$, $w_{i,j}^{(\theta_{i,j})}(t) = 0$, $t \leq 0$, be density families parametrized by $\theta_{i,j} \in \Theta_{i,j} \subset \mathbb{R}^{d_{i,j}}$. With the notation from* (25)*, let*

$$(\hat{a}_{i,j}, \hat{\theta}_{i,j}) := \operatorname{argmin}_{(a,\theta)\in\mathbb{R}_{\geq 0}\times\Theta_{i,j}} \sum_{k=1}^{p} \Big(a w_{i,j}^{(\theta)}\big(k\Delta_{graph}\big) - \hat{h}_{i,j}(k\Delta_{graph})\Big)^2, \quad (i,j) \in \widehat{\mathcal{E}}^*, \tag{26}$$

and define the parametric reproduction-intensity estimates

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

the parametric branching-matrix estimate

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

and the parametric immigration-intensity estimates*.*

$$\hat{\eta}^{(par)} := \left(\hat{\eta}_1^{(par)}, \dots, \hat{\eta}_d^{(par)}\right) := \lambda^{(emp)}\left(1_{d\times d} - \widehat{A}^{(par)}\right), \tag{27}$$

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

We illustrate this specification and estimation of a fully parametric multivariate Hawkes process in Figure 3. Here, we also see that the parameter estimates from (26) are symmetrically distributed around the true values. Even though the estimator calculations in Definition 11 stand at the end of a long chain of various discretizations and truncations, 'log-likelihood profile' confidence intervals (e.g., from `confint.nls` in `R`) give remarkably good coverage rates for the parameter estimates (not illustrated).

Remark 3. *The definition of $\eta^{(par)}$ in* (27) *is motivated by the desirable equality*

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

In other words, with this choice of $\hat{\eta}^{(par)}$, the observed unconditional intensity exactly equals the estimated unconditional intensity. This might be relevant in some applications (e.g., simulation from a fitted model). Finally note that it might often be more efficient to consider weighted least squares in (26)*.*

4 Example

We illustrate the concepts introduced in the previous sections 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-type Hawkes process $\mathbf{N}$ as in Definition 4 with immigration intensities

$$\eta_i := \begin{cases} 1, & i \in \{1,7,10\}, \\ 0, & i \in \{2,3,4,5,6,8,9\}, \end{cases} \tag{28}$$

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

$$h_{i,j}(t) := \begin{cases} 1.5\,\gamma(t), & (i,j) \in \{(1,2),(2,4),(8,9)\}, \\ 1_{t\in[1,2]}0.5, & (i,j) \in \{(1,1),(2,3),(3,5),(4,3),(4,5),(4,6),(5,3),(7,8),(9,7)\}, \\ 1_{t\in[1,2]}0.1, & (i,j) = (5,7), \\ 0, & \text{else}. \end{cases} \tag{29}$$

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 (29); see the left graph in Figure 1. From this figure, the various direct and indirect dependencies can be read off instantaniously; only the large nodes have nonzero 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 6 and 7:

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 $1,2,3,4,5$ and 6 would be *redundant* as they could not generate events.

Connectivity The Hawkes graph $\mathcal{G}_{\mathbf{N}}$ is *not weakly connected.* The graph can be divided in two separate weakly-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 weakly-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 (5) for vertices $i_0 \in \{1,2,3,7\}$. For vertex 1, we find that

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

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

$$\mathcal{W}_1^{(2,2)} = \mathcal{W}_2^{(2,2)} = \emptyset,\ \mathcal{W}_3^{(2,2)} = \{(2,4,6,2)\},\ \mathcal{W}_4^{(2,2)} = \mathcal{W}_5^{(2,2)} = \emptyset,\ \mathcal{W}_6^{(2,2)} = \{(2,4,6,2,4,6,2)\},\ldots$$

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 (5) again follows. For vertices 3 and 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 (29), respectively, Figure 1, we would have to calculate the spectral radius of the adjacency matrix corresponding to the Hawkes (sub-)graph with edges

$$\big\{(3,5;0.5),(5,3;0.5),(5,7;0.1),(7,8;0.5),(8,9;0.5),(9,5;a_{9,5}),(9,7;0.5)\big\};$$

see Theorem 1.

Cascade and feedback coefficients We calculate the coefficients from Definition 7 with respect to the example model; see Table 1. The cascade and feedback coefficients summarize the impact of the driving vertices 1, 4 and 10 (that is, of the vertices with nonzero vertex weights) on the process. The *cascade coefficients* measure the impact of each vertex on the whole system. In our example, the immigrants in the first vertex together with the cascades that they trigger are responsible for about 82% of all events that occur in the system. The *feedback coefficients* measure the impact of the impact of each vertex on itself. In our example, for vertex 8 this means that 76% of its activity are explained by its own immigration activity and by the feedback loops that the immigrants possibly trigger via closed walks. Vertex 1 is only excited by its own activity. For vertex 10 the feedback coefficient is also equal 1—albeit there is no true feedback involved. Still, its intensity would decrease by 100% if it were switched off.

Table 1: Cascade and feedback coefficients

	1	2	3	4	5	6	7	8	9	10
cascade.coefficients	0.82	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.04
feedback.coefficients	1.00	0.00	0.00	0.00	0.00	0.00	0.76	0.00	0.00	1.00

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 and 4 as simulation algorithm. In each realization, 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 9—with respect to different values of Δ_{skel} and α_{skel}. Given these skeleton estimates, we calculate the Hawkes-graph estimator from Definition 10—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 9 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 $\{\widehat{\mathcal{E}}^*(k)\}_{k=1,2,\dots,n_{\text{sim}}}$ and the true edge set by $\mathcal{E}^*$. Using this notation, we summarize the results of the simulation study in Tables 2, 3, 4, and 5 with the following statistics:

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

ii) *total*: fraction of correctly included edges, i.e, of pairs $(i,j) \in \widehat{\mathcal{E}}^*_{\mathbf{N}}(k)$ such that $(i,j) \in \mathcal{E}_{\mathbf{N}}$:

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

Note that $1 - \textit{total}$ is the *false-negative rate.*

iii) *heavy/light/super.light*: more detailed version of ii) above; fractions of correctly estimated edges with heavy (1.5), light (0.5) and super-light (0.1) edge weights.

Table 2: $\Delta_{\text{skel}} = 0.2$

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

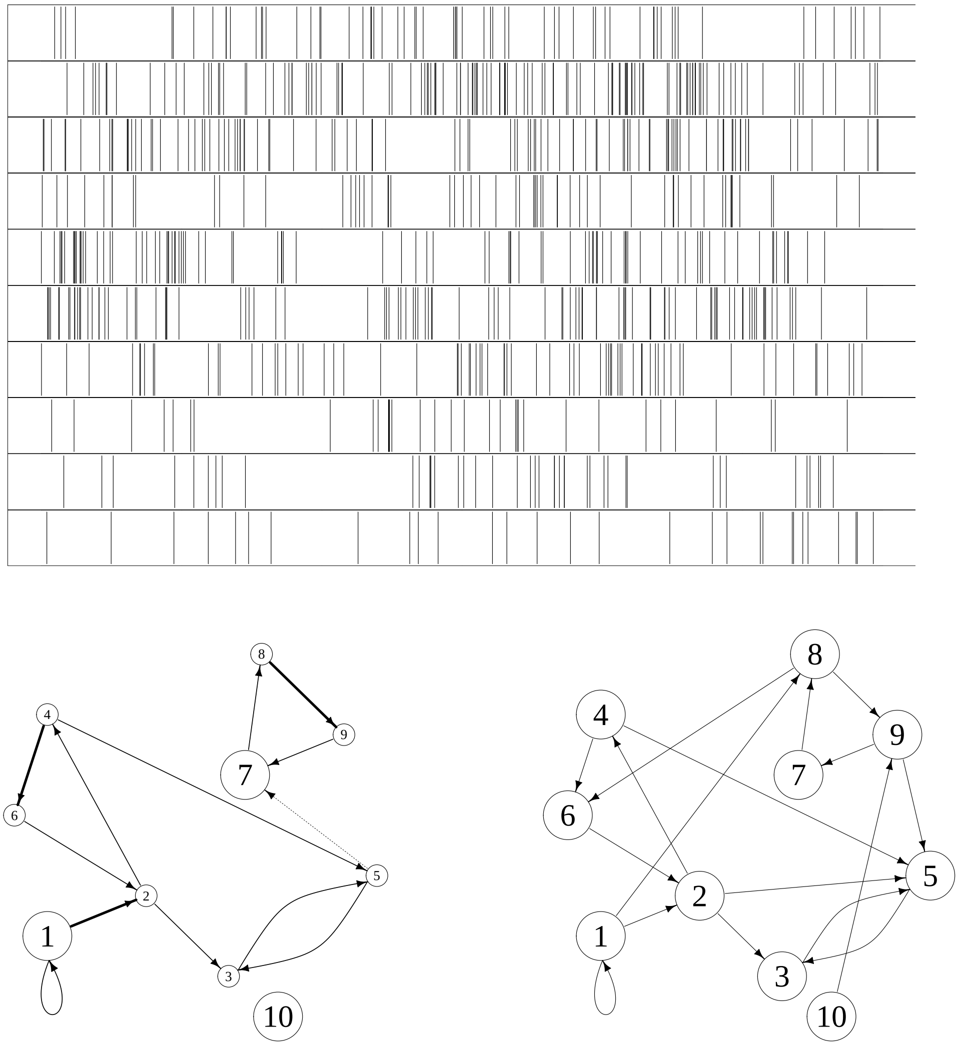

Figure 1: Hawkes process simulation, Hawkes graph, and estimated Hawkes skeleton. The left graph represents the Hawkes graph corresponding to the Hawkes process example from Section 4.1; the graph is a summary of the immigration and branching structure of the model: edges from one vertex to another vertex denote nonzero 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 (29). Large vertices correspond to nonzero immigration-intensities (= 1) and small vertices to the zero-immigration vertices; see (28). 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 10 with respect to a coarse bin size $\Delta_{\mathrm{skel}} = 1$ and a sparseness parameter $\alpha_{\mathrm{skel}} = 0.05$. The right graph illustrates such an 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 twelve of the thirteen true edges. We miss edge $(5, 7)$. Furthermore, the estimate introduces five additional wrong edges $(1, 8)$, $(2, 5)$,$(8, 6)$, $(9, 5)$, and $(10, 9)$. The three crucial points are: (i) These five 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. This considerably reduces the complexity of the graph estimation (with a very fine Δ_{graph}-parameter). See Figure 2, for the Hawkes-graph estimation with respect to the skeleton estimate from above.

Table 3: $\Delta_{\text{skel}} = 0.5$

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 4: $\Delta_{\text{skel}} = 1$

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

iv) *zero*: fraction of correctly excluded edges, i.e., of pairs $(i,j) \notin \widehat{\mathcal{E}}^*_{\mathbf{N}}(k)$ such that $(i,j) \notin \mathcal{E}_{\mathbf{N}}$:

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

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

First, we discuss the estimations with respect to bin size $\Delta_{\text{skel}} = 0.2$; see Table 2. 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 9, 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 3, 4, and 5 look roughly the same as Table 2 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 the skeleton estimation feasible even for much higher dimensions.

The main purpose of the skeleton estimation is to lay the ground for the graph estimation which itself depends on a given estimated skeleton; see Definition 10. 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 ($= 1 - \textit{total}$) in the skeleton estimation very small. As a consequence, we need α_{skel} large to include more edges. 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 10. 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.,

Table 5: $\Delta_{\text{skel}} = 2$

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

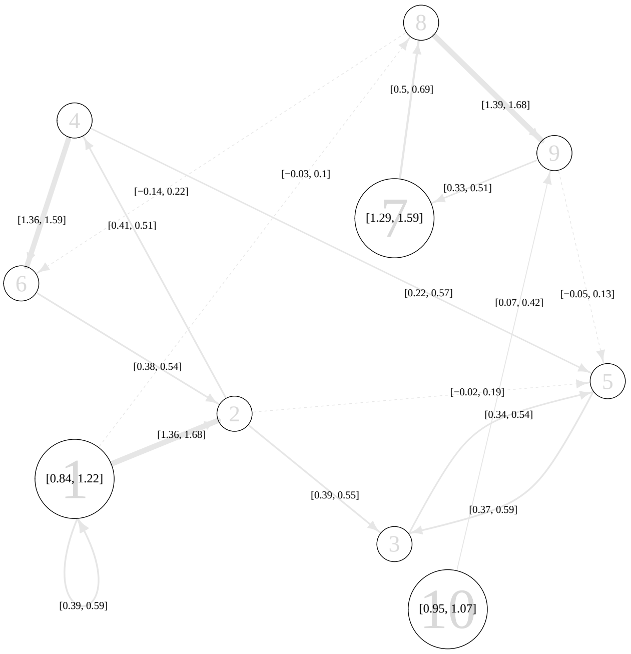

Figure 2: Hawkes-graph estimation. Given a single simulation of length $T = 1000$ from the example Hawkes process in Section 4.1, we calculate the Hawkes-graph estimator from Definition 10 with respect to the Hawkes-skeleton estimation from Figure 1; we apply a bin size $\Delta_{\text{graph}} = 0.025$ 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 in the illustration 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 latter case. Comparing the results with the true Hawkes graph in Figure 1, respectively, with the Hawkes process parametrization in (28) and (29), 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 $(1, 8)$, $(2, 5)$,$(8, 6)$, and $(9, 5)$, 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 true vertex weights but the weight of vertex 7 are also covered by the confidence intervals. The weight of vertex 7 is overestimated because we missed the (light) edge $(5, 7; 0.1)$ in the skeleton estimation; this missing explanatory variable for the events in component 7 is compensated by an extra large vertex weight in the graph estimation. Deleting all insignificant (in figure dashed) edges and setting the vertex weight of the insignificant (in figure small) vertex-weights to zero, we recover the original underlying graph almost perfectly.

the false-negative rate is really small. In the next section, we will see that the graph estimates are not dramatically sensitive to the α_{skel} parameter in the skeleton estimation.

Hawkes-graph estimation

In a further step, we quantify the estimated excitements. That is, given a Hawkes skeleton, we estimate the corresponding graph as in Definition 10; 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 skeletons that were estimated with respect to different α_{skel}-parameters. However, we only consider the skeletons that were estimated with respect to the (rough) bin size $\Delta_{\text{skel}} = 1$. As opposed to the skeleton estimation, we may now use a much smaller bin size $\Delta_{\text{graph}} = 0.1$ for the graph estimation. In the present example, this is approximately 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, no parallelization). 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 from Definition 10. Table 6 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 super-light edge $(5, 7; 0.1)$), the model calibration

Table 6: $\Delta_{\text{graph}} = 0.1$ 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

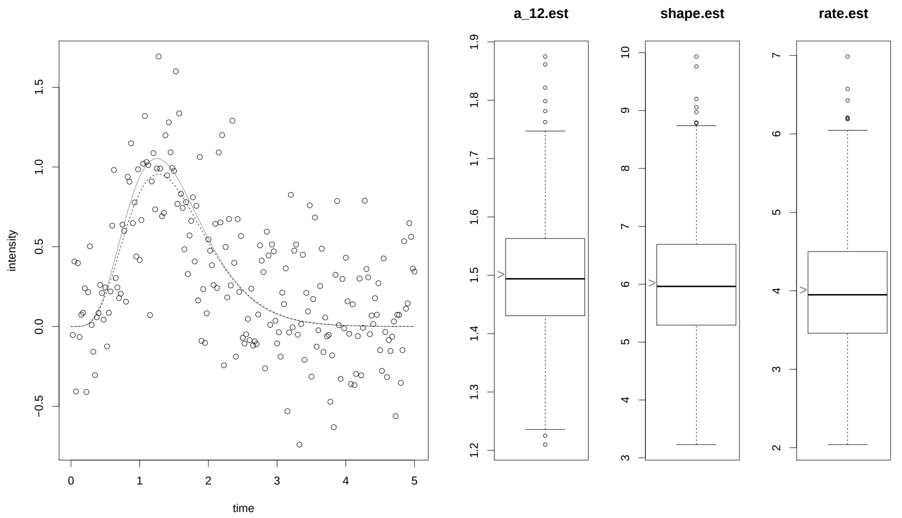

Figure 3: Parametric estimation. *Left:* From a single realization of the example Hawkes model from Section 4.1 with length $T = 1000$, we calculate Hawkes-skeleton and Hawkes-graph estimates from Definitions 9 and 10; see Figures 1 and 2. As a by-product of these calculations, we retrieve pointwise estimates (circles) for the values of the reproduction intensity $h_{1,2}$ on an equidistant grid; see (24). From these estimates, one may guess that $h_{1,2}(t) = a_{1,2}\gamma(t)$, where γ is a Gamma density depending on a shape and on a rate parameter. We fit the three parameters by nonlinear least squares as described in Section 3.4. The dotted black line refers to the corresponding estimated parametric function. It catches the true underlying function (grey solid line) quite well; see (29). *Right:* We apply this parametric estimation of $h_{1,2}$ on 1000 independent realizations of length $T = 500$. The boxplots collect the parameter estimates for each of the 1000 estimations of the simulation study. The grey marks refer to the corresponding true values. Eyeball-examination shows that the estimates are remarkably symmetric distributed and unbiased. QQ-plots (not illustrated) support asymptotic normality.

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 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 reproduction intensity values on an equidistant grid; see (24). Based on the scatter plots of these estimates, we choose appropriate parametrized function families. Given such parametric functions, the parameters are fit to the pointwise estimates via nonlinear least squares; see Figure 3. QQ-plots (not included) support asymptotic normality for the parameter estimates.

5 Conclusion

The Hawkes graph and the Hawkes skeleton describe the immigration and branching structure of a Hawkes process in a graph-theoretical framework. We demonstrate how graph terminology can be very useful for multivariate Hawkes processes. Combining the new concepts with an estimation procedure from earlier work, we develop a statistical estimation method for the Hawkes skeleton and the Hawkes graph. The key idea is that in a preliminary step we only test if there is *at all* excitement from any vertex to another vertex. 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 5 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. Also 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 (2016b). 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—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 factors 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 helped to improve the presentation. We also thank Philippe Deprez for a fertile discussion about Theorem 1.

References

Bacry, E., Gaïffas, S., and Muzy, J. (2015). A generalization error bound for sparse and low-rank multivariate Hawkes processes. *arXiv:1501.00725.*

Bates, D. and Maechler, M. (2015). Matrix: Sparse and dense matrix classes and methods. *R package version 1.1-5.* `http://CRAN.R-project.org/package=Matrix`.

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

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, volume I and II. Springer, New York, 2nd edition.

Delattre, S., Fournier, N., and Hoffmann, M. (2015). Hawkes processes on large networks. *PNAS*, 105(41).

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

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

Hall, E. and Willett, R. (2016). Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, 62(7):4327–4346.

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

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

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

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

Kirchner, M. (2016a). An estimation procedure for the Hawkes process. *Quantitative Finance*. (to appear).

Kirchner, M. (2016b). Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 162:2494–2525.

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

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

Ogata, Y. (1988). Statistical models for earthquake occurences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition.

Shi, Z. (2015). *Branching Random Walks*, volume 2151 of *Lecture Notes in Mathematics*. Springer, Cham.

Song, L., Zha, H., and Zhou, K. (2013). 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)*, 31.

Watson, C. (2015). The geometric series of a matrix. `http://www.math.uvic.ca/~dcwatson/work/geometric.pdf`.