*Bioinformatics*, 38(7), 2022, 1846–1856
https://doi.org/10.1093/bioinformatics/btac027
Advance Access Publication Date: 18 January 2022
Original Paper

Phylogenetics

# From viral evolution to spatial contagion: a biologically modulated Hawkes model

**Andrew J. Holbrook[1,*], Xiang Ji[2] and Marc A. Suchard[1,3,4]**

[1]Department of Biostatistics, University of California, Los Angeles, CA 90095, USA, [2]Department of Mathematics, Tulane University, New Orleans, LA 70118, USA, [3]Department of Biomathematics and [4]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on September 10, 2021; revised on December 11, 2021; editorial decision on January 5, 2022; accepted on January 12, 2022

## Abstract

**Summary:** Mutations sometimes increase contagiousness for evolving pathogens. During an epidemic, scientists use viral genome data to infer a shared evolutionary history and connect this history to geographic spread. We propose a model that directly relates a pathogen's evolution to its spatial contagion dynamics—effectively combining the two epidemiological paradigms of phylogenetic inference and self-exciting process modeling—and apply this *phylogenetic Hawkes process* to a Bayesian analysis of 23 421 viral cases from the 2014 to 2016 Ebola outbreak in West Africa. The proposed model is able to detect individual viruses with significantly elevated rates of spatiotemporal propagation for a subset of 1610 samples that provide genome data. Finally, to facilitate model application in big data settings, we develop massively parallel implementations for the gradient and Hessian of the log-likelihood and apply our high-performance computing framework within an adaptively pre-conditioned Hamiltonian Monte Carlo routine.

**Contact:** aholbroo@g.ucla.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

## 1 Introduction

The coronavirus disease 2019 (COVID-19) pandemic has demonstrated the need for new scientific tools for the analysis and prediction of viral contagion across human landscapes. The mathematical characterization of the complex relationships underlying pathogen genetics and spatial contagion stands as a central challenge of 21st century epidemiology. We approach this task by unifying two distinct probabilistic approaches to viral modeling. On the one hand, Bayesian phylogenetics (Mau *et al.*, 1999; Sinsheimer *et al.*, 1996; Suchard *et al.*, 2001; Yang and Rannala, 1997) uses genetic sequences from a limited collection of viral samples to integrate over high-probability shared evolutionary histories in the form of *phylogenies* or family trees. On the other hand, self-exciting, spatiotemporal Hawkes processes (Reinhart, 2018) model spatial contagion by allowing an observed event to increase the probability of additional observations nearby and in the immediate future.

Both modeling paradigms come with their own advantages. For Bayesian phylogenetics, the past twenty years have witnessed a slew of high-impact scientific studies in viral epidemiology (Boni *et al.*, 2020; Dudas *et al.*, 2017; Faria *et al.*, 2014; Gire *et al.*, 2014; Rambaut *et al.*, 2008; Smith *et al.*, 2009) and the rise of powerful computing tools facilitating inference from expressive, hierarchical models of phylogenies and evolving traits (Ronquist *et al.*, 2012; Suchard *et al.*, 2018). Unfortunately, the number of evolutionary trees to integrate over explodes with the number of viral samples analyzed (Felsenstein, 1978), so Bayesian phylogenetic analyses typically restrict to a relatively small number of viral samples, at most totaling a few thousand. The fact that viral cases that undergo genetic sequencing usually represent a small subset of the total case count exacerbates this issue. Thus, failure to detect phylogenetic clades that represent novel strains on account of computational and surveillance limitations always remain a possibility. Until now, these weaknesses have also held for the sub-discipline of Bayesian phylogeography, which attempts to relate viral evolutionary histories to geographic spread as represented by (typically Brownian) phylogenetic diffusions. These models describe viral spread through either discretized (Lemey *et al.*, 2009) or continuous (Lemey *et al.*, 2010) space, but both approaches induce their own form of bias (Holbrook *et al.*, 2021a). In the face of these shortcomings, Bayesian phylogeography needs new tools for directly modeling spatial contagion (Table 1).

Hawkes processes (Hawkes, 1971a,b, 1972, 2018; Hawkes and Adamopoulos, 1973) are widely applicable point process models for generally viral or contagious phenomena, such as earthquakes and aftershocks (Fox *et al.*, 2016; Hawkes and Adamopoulos, 1973; Ogata, 1988; Zhuang *et al.*, 2004), financial stock trading (Bacry

**Table 1.** Comparison of two probabilistic modeling paradigms within viral epidemiology, the combination of which represents a new tool for Bayesian phylogeography

| | Traditional Bayesian phylogenetics | Hawkes processes |
|---|---|---|
| Observational limit | $N$ in low thousands | $N$ in high tens-of-thousands |
| Biological insight | Evolutionary history | None |
| Genetic sequencing | Required | Not required |
| Spatiotemporal data | Not required | Required |
| Geographic spread | Not modeled | Modeled |
| Large-scale transport | Does not induce bias | Induces bias |

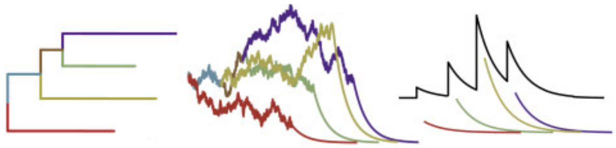

**Fig. 1.** The phylogenetic Hawkes process relates the evolutionary history of a virus to the rate at which subsequent viral cases occur nearby. (Left) A phylogenetic tree characterizes the evolution of a viral strain. (Middle) A 1D Brownian motion 'along the tree' describes the evolution of infinitesimal rates as a function of branch lengths and tree topology (Section 2.2). Branch lengths influence this evolution insofar as they dictate the length of time over which the individual rates evolve; tree topology influences this evolution insofar as a parent's trajectory fixes the beginning of the child's trajectory. (Right) Virus-specific rates additively contribute to the Hawkes process' rate function for future cases (Section 2.1)

*et al.*, 2015; Hawkes, 2018), viral content on social media (Kobayashi and Lambiotte, 2016; Rizoiu *et al.*, 2017), gang violence (Holbrook *et al.*, 2021b; Loeffler and Flaxman, 2018; Mohler, 2013, 2014; Park *et al.*, 2019) and wildfires (Schoenberg, 2004). Unsurprisingly, Hawkes processes are natural models for the contagion dynamics of biological viruses as well. Kim (2011) uses spatiotemporal Hawkes processes (Reinhart, 2018), which model viral cases as unmarked events in space and time, to model the spread of avian influenza virus (H5N1). Meyer and Held (2014) incorporate power laws to describe spatial contagion dynamics and model meningococcal disease in Germany from 2001 to 2008. Although Rizoiu *et al.* (2018) do not model epidemiological data, they do draw connections between epidemiological susceptible, infectious or recovered (SIR) models and Hawkes processes, showing that the rate of events in the SIR model is equal to that of a finite-population Hawkes model. Kelly *et al.* (2019) apply a temporal nonparametric Hawkes process to the 2018–2019 Ebola outbreak in the Democratic Republic of the Congo and successfully generate accurate disease prevalence forecasts. Chiang *et al.* (2020) model COVID-19 cases and deaths in the USA at the county level using spatially indexed mobility and population data to modify the process conditional intensity. Most recently, Bertozzi *et al.* (2020) compare the performance of a temporal Hawkes process model with temporally evolving conditional intensity to that of SIR and susceptible, exposed, infectious or recovered (SEIR) models for modeling regional COVID-19 case dynamics.

Because such Hawkes processes do not involve genetic information, one may apply the model to a much larger collection of cases, i.e. those for which a timestamp and spatial coordinates are available. Moreover, recent successes in scaling Hawkes process inference to a big data setting enable inference from observations numbered in the high tens-of-thousands (Holbrook *et al.*, 2021b; Yuan *et al.*, 2021). This ability to interface with an order of magnitude more cases represents a major benefit of the Hawkes process in comparison to the Bayesian phylogenetic paradigm, but the tradeoff is that conclusions drawn from a Hawkes process analysis are devoid of explicit biological insight. It is possible for the model to attribute self-exciting dynamics to nearby viral cases that are only distantly related to the phylogenetic tree. Finally, these processes do not immediately account for viral spread through large-scale transportation networks (Brockmann and Helbing, 2013; Holbrook *et al.*, 2021a) but attribute events resulting from such contagion to a 'background' process.

In the following, we construct a Bayesian hierarchical model that allows both modeling approaches to support each other. This model (Fig. 1) learns phylogenetic trees that describe the evolutionary history of the subset of observations that yield genetic sequencing and uses this history to inform the distribution of a latent relative rate or *productivity* (Schoenberg, 2020; Schoenberg *et al.*, 2019) for each virus in this limited set. In turn, these virus-specific rates modify the rate of self-excitation of a spatiotemporal Hawkes process describing the contagion of all viruses, sequenced or not. We use a Metropolis-within-Gibbs strategy to jointly infer all parameters and latent variables of our phylogenetic Hawkes process and overcome the $O(N^2)$ computational complexity of the Hawkes process likelihood by incorporating the modified likelihood in the hpHawkes open-source, high-performance computing library (Holbrook *et al.*, 2021b) available at https://github.com/suchard-group/hawkes. Within the same library, we also develop multiple parallel computing algorithms for the log-likelihood gradient and Hessian with respect to the model's virus-specific rates. Graphics processing units (GPU)-based implementations of these gradient and Hessian calculations score 100-fold speedups over single-core computing and help overcome quadratic complexity in the context of an adaptively preconditioned Hamiltonian Monte Carlo (HMC; Neal, 2011). These speedups prove useful in our analysis of 23 421 viral cases from the 2014 to 2016 Ebola outbreak in West Africa.

## 2 Materials and methods

We develop the phylogenetic Hawkes process and its efficient inference in the following sections. Importantly, our proposed hierarchical model integrates both sequenced and unsequenced viral case data, representing a significant and clear contribution insofar as:

1. the percentage of confirmed viral cases sequenced during an epidemic is often in the single digits (Wadman, 2021); and
2. previous phylogeographic models have failed to leverage additional information provided by geolocated, unsequenced case data.

We address this shortcoming by constructing a new hierarchical model that *both* models all spatiotemporal data with a Hawkes process (Section 2.1) *and* allows an inferred evolutionary history in the form of a phylogenetic tree to influence dependencies between relative rates of contagion (Section 2.2) for the small subset of viral cases for which genome data are available. We believe that this approach is altogether novel.

### 2.1 Spatiotemporal Hawkes process for viral contagion

Hawkes processes (Hawkes, 1971a,b, 1972, 2018) constitute a useful class of inhomgeneous Poisson point processes (Daley and Jones, 2003) for which individual events contribute to an increased rate of future events. Spatiotemporal Hawkes processes (Reinhart, 2018) are marked Hawkes processes with spatial coordinates for marks (Daley and Jones, 2003). We are interested in spatiotemporal Hawkes processes with infinitesimal rate:

$$\lambda(\mathbf{x},t) = \mu(\mathbf{x}) + \xi(\mathbf{x},t) = \mu(\mathbf{x}) + \sum_{t_n<t} g_n(\mathbf{x}-\mathbf{x}_n, t-t_n)\,,$$

where $\mathbf{x} \in \mathbb{R}^D$, $t \in \mathbb{R}^+$ and the subscript $n$ indicates that the usual triggering function $g(\cdot,\cdot)$ takes on different forms depending on some characteristic associated with event $n$. These non-negative, monotonically non-increasing, event-indexed triggering functions additively contribute to $\xi(\cdot,\cdot)$, the *self-excitatory* rate component, and encourage this rate to increase after each observed event. Here, $\mu(\cdot)$ is the *background* rate and only depends on spatial position $\mathbf{x}$. Conditioned on observations $(\mathbf{x}_n, t_n)$, $n = 1,\dots,N$, we specify the rate components:

$$\mu(\mathbf{x}) = \frac{\mu_0}{\tau_x^D}\sum_{n=1}^N \phi\left(\frac{\mathbf{x}-\mathbf{x}_n}{\tau_x}\right)\mathcal{I}_{[x\neq x_n]} \quad \text{and}$$
$$\xi(\mathbf{x},t) = \frac{\theta_0\omega}{h^D}\sum_{t_n<t}\theta_n\, e^{-\omega(t-t_n)}\phi\left(\frac{\mathbf{x}-\mathbf{x}_n}{h}\right),$$

where $\tau_x > 0$ and $h > 0$ are the background and self-excitatory spatial lengthscales, $\mu_0 > 0$ and $\theta_0 > 0$ are the background and self-excitatory weights, $1/\omega > 0$ is the self-excitatory temporal lengthscale, $\phi(\cdot)$ is the $D$-dimensional standard normal probability density function, and the background rate's indicator function prevents a trivial maximum at $\tau_x \to 0$ (Habbema *et al.*, 1974; Robert, 1976). The inclusion of $\theta_n > 0$ for $n = 1,\dots,N$ within the self-excitatory rate marks a major departure from similar model specifications in Holbrook *et al.* (2021b) and Loeffler and Flaxman (2018). These 'degrees of contagion' or 'productivities' (Schoenberg, 2020; Schoenberg *et al.*, 2019) allow different events to contribute differently to the overall self-excitatory rate of the process: the larger the $\theta_n$, the higher the rate directly following event $n$ (Fig. 1, *right*). Following the connection of the Hawkes process with exponential triggering function to a discret-time SIR model (Rizoiu *et al.*, 2018), Bertozzi *et al.* (2020) refer to these quantities as a reproduction number. In the following, we refer to $\theta_n$ as the *event-specific*, *case-* or *virus-specific rate* for the $n$th event, case or viral observation.

Denoting $\boldsymbol{\theta} = (\theta_1,\dots,\theta_N)^T$, the likelihood of observing data $(\mathbf{X},\mathbf{t}) = ((\mathbf{x}_1,t_1),\dots,(\mathbf{x}_N,t_N))^T$ is (Daley and Jones, 2003):

$$\mathcal{L}(\mathbf{X},\mathbf{t}|\mu_0,\tau_x,\theta_0,\boldsymbol{\theta},\omega,h) = \exp\left(-\int_{\mathbb{R}^D}\int_0^{t_N}\lambda(\mathbf{x},t)\,\mathrm{d}t\,\mathrm{d}\mathbf{x}\right)\prod_{n=1}^N\lambda(\mathbf{x}_n,t_n)$$
$$:= e^{-\Lambda(t_N)}\cdot\prod_{n=1}^N\lambda_n\,.$$

The choice of $\mathbb{R}^D$ for integration domain is popular and often necessary but assumes complete observation over the entirety of $\mathbb{R}^D$ (Schoenberg, 2013). The resulting integral may be written (Appendix A):

$$\Lambda(t_N) = \mu_0 t_N - \theta_0\sum_{n=1}^N\theta_n(e^{-\omega(t_N-t_n)}-1) := \sum_{n=1}^N\Lambda_n\,,$$

leading to a log-likelihood of

$$\begin{aligned}\ell(\mathbf{X},\mathbf{t}|\mu_0,\tau_x,\theta_0,\boldsymbol{\theta},\omega,h) &= -\Lambda(t_N) + \sum_{n=1}^N\log\lambda_n\\ &= \sum_{n=1}^N\Bigg\{\log\Big[\sum_{n'=1}^N\Big(\frac{\mu_0\,\mathcal{I}_{[\mathbf{x}_n\neq\mathbf{x}_{n'}]}}{\tau_x^D}\phi\left(\frac{\mathbf{x}_n-\mathbf{x}_{n'}}{\tau_x}\right)\\ &\quad+\frac{\theta_0\theta_{n'}\omega\,\mathcal{I}_{[t_{n'}<t_n]}}{h^D}e^{-\omega(t_n-t_{n'})}\phi\left(\frac{\mathbf{x}_n-\mathbf{x}_{n'}}{h}\right)\Big)\Big]-\Lambda_n\Bigg\}\\ &:= \sum_{n=1}^N\left[\log\left(\sum_{n'=1}^N\lambda_{nn'}\right)-\Lambda_n\right] := \sum_{n=1}^N\ell_n\,.\end{aligned}\tag{1}$$

Here, we use the following shorthand notations: $\Lambda_n$ is the additive contribution of the $n$th event to the likelihood's integration term; $\lambda_{nn'}$ refers to the additive contribution of the $n'$th event to the rate function evaluated at the $n$th event $\lambda_n = \lambda(\mathbf{x}_n,t_n)$; and $\ell_n$ is the overall additive contribution of the $n$th event to the log-likelihood. We reference these formulas while outlining our inference strategy in Section 2.3 and detailing our massively parallel algorithms for calculating the log-likelihood gradient and Hessian with respect to event-specific rates $\boldsymbol{\theta}$ in Appendix B. We describe our biologically modulated joint prior on event-specific rates $\theta_1,\dots,\theta_N$ in Section 2.2.

## 2.2 Phylogenetic Brownian process prior on rates

We use standard Bayesian phylogenetics hierarchical approaches (Suchard *et al.*, 2003) to model a single molecular sequence alignment $\mathbf{S}$ containing sequences from $M \le N$ evolutionarily related viruses. Let $\mathcal{M}$ denote the ordered index set with cardinality $|\mathcal{M}| = M$ containing every number within the set $\{1,\dots,N\}$ that corresponds to an observed virus for which genome data are present. In the following, we number the elements within $\mathcal{M}$ as $m_1, m_2,\dots,m_M$. Moreover, we make use of the set $\mathcal{M}^+$ with cardinality $|\mathcal{M}^+| = 2M-1$, satisfying $\mathcal{M}\subset\mathcal{M}^+$ and containing elements $m_1,\dots,m_{2M-1}$. Our primary object of interest is the phylogenetic tree $\mathcal{G}$ (Fig. 1, *left*) defined as a bifurcating, directed graph with $M$ terminal degree-1 nodes $(\nu_{m_1},\dots,\nu_{m_M})$ that correspond to the tips of the tree (or sequenced observations), $M-2$ internal degree-3 nodes $(\nu_{m_{M+1}},\dots,\nu_{m_{2M-2}})$, a root degree-2 node $\nu_{m_{2M-1}}$ and edge weights $(w_{m_1},\dots,w_{m_{2M-2}})$ that encode the elapsed evolutionary time between nodes. Here, each $w_m$ communicates the expected number of molecular substitutions per site, which is itself the product between the real-time duration and the evolutionary rate arising from a molecular clock model. For example, we use a relaxed molecular clock model (Drummond *et al.*, 2006) that allows for substitution rates to flexibly vary across branches (Section 3.2). One may either know $\mathcal{G}$ *a priori* or endow it with a prior distribution parameterized by some vector $\phi$. Suchard *et al.* (2001, 2018) develop the joint distribution $p(\mathbf{S},\phi,\mathcal{G})$ in detail.

We assume that the event-specific rates $\boldsymbol{\theta}$ defined within our Hawkes model take the form (Fig. 1, *middle*):

$$\begin{cases}\theta_n = \theta_n(z_n) = \exp\left(z_n + \boldsymbol{\beta}^T\mathbf{f}(t_n)\right) & z_n\in\mathbb{R}\,,\quad n\in\mathcal{M}\\ \theta_n = 1 & \qquad\qquad n\notin\mathcal{M}\,,\end{cases}$$

and that the elements of vector $\mathbf{z} = (z_{m_1},\dots,z_{m_M})^T$ follow a Brownian diffusion process along the branches of $\mathcal{G}$ (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1985; Lemey *et al.*, 2010). Here, $\mathbf{f}(\cdot)$ is some fixed vector function and the inclusion of the linear term $\boldsymbol{\beta}^T\mathbf{f}(t_n)$ helps control for global trends resulting from extrinsic events such as mass quarantine or travel restrictions. Under the Brownian process, the latent value of a child node $\nu_c$ in tree $\mathcal{G}$ is normally distributed about the value of its parent node $\nu_{\mathrm{pa}(c)}$ with variance $w_c\times\sigma^2$, where $\sigma^2$ gives the dispersal rate after controlling for correlation in values that are shared by descent through the phylogenetic tree $\mathcal{G}$. We further posit that the latent value of the root node $\nu_{m_{2M-1}}$ is *a priori* normally distributed with mean 0 and variance $\tau_0\times\sigma^2$. The vector $\mathbf{z}$ is then multivariate normally distributed Cybis *et al.* (2015) and has probability density function:

$$p(\mathbf{z}\,|\,\mathbf{V}_{\mathcal{G}},\sigma^2,\tau_0) = (2\pi\sigma^2)^{-M/2}|\mathbf{V}_{\mathcal{G}}|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\mathbf{z}^T\mathbf{V}_{\mathcal{G}}^{-1}\mathbf{z}\right),\tag{2}$$

where $\mathbf{V}_{\mathcal{G}} = \{v_{nm}\}$ is a symmetric, positive definite, block-diagonal $M\times M$ matrix with structure dictated by $\mathcal{G}$. Defining $d_F(u,v)$ to be the sum of edge-weights along the shortest path between nodes $u$ and $v$ in tree $\mathcal{G}$, the diagonal elements $v_{mm} = \tau_0 + d_F(\nu_{m_{2M-1}},\nu_{m_m})$ are the elapsed evolutionary time between the root node and tip node $m_m$, and off-diagonal elements $v_{nm} = \tau_0 + [d_F(\nu_{m_{2M-1}},\nu_{m_n}) + d_F(\nu_{m_{2M-1}},\nu_{m_m}) - d_F(\nu_{m_n},\nu_{m_m})]/2$ are the evolutionary time period between the root node and the most recent common ancestor of tip nodes $m_n$ and $m_m$.

## 2.3 Inference

Due to the complexity of the phylogenetic Hawkes process and the large number of viruses we seek to model, we must use advanced statistical, algorithmic and computational tools to infer the posterior distribution:

$$
\begin{aligned}
&p(\sigma^2, \mathcal{G}, \phi, \mu_0, \tau_x, \theta_0, \boldsymbol{\theta}, \omega, h, \boldsymbol{\beta} \,|\, \mathbf{X}, \mathbf{t}, \mathbf{S}) \\
&\quad \propto \mathcal{L}(\mathbf{X}, \mathbf{t} | \mu_0, \tau_x, \theta_0, \boldsymbol{\theta}(\mathbf{z}), \omega, h) \times p(\mathbf{z} \,|\, \sigma^2, \mathcal{G}) \times p(\mu_0) \times p(\tau_x) \\
&\qquad \times p(\theta_0) \times p(\omega) \times p(h) \times p(\boldsymbol{\beta}) \times p(\sigma^2) \times p(\mathbf{S}, \phi, \mathcal{G}) \,. \qquad (3)
\end{aligned}
$$

We do so using a random-scan Metropolis-with-Gibbs scheme, in which we compute key quantities with the help of adaptively preconditioned HMC (Neal, 2011), dynamic programming and parallel computing on cutting-edge GPUs.

### 2.3.1 Dynamic programming for phylogenetic diffusion quantities

We must evaluate $p(\mathbf{z} \,|\, \sigma^2, \mathcal{G})$ to sample $\mathcal{G}$. The bottleneck within the evaluation of Equation (2) is the ostensibly $\mathcal{O}(M^3)$ matrix inverse $\mathbf{V}_{\mathcal{G}}^{-1}$, but Pybus *et al.* (2012) develop a dynamic programming algorithm to perform the requisite computations in $\mathcal{O}(M)$ with parallelized post-order traversals of $\mathcal{G}$. We use this algorithm, which is closely related to the linear-time algorithms of Freckleton (2012) and Ho and Ané (2014), as all are examples of message passing on a directed, acyclic graph (Cavalli-Sforza and Edwards, 1967; Pearl, 1982). Similar tricks render inference for $\phi$ linear in $M$, and Fisher *et al.* (2021) extend Pybus *et al.* (2012) to compute gradients with respect to $\phi$. Finally, implementing these algorithms on GPUs would lead to additional speedups (Suchard and Rambaut, 2009), but the computational bottleneck we face when applying the phylogenetic Hawkes process arises from the Hawkes process likelihood and its gradients.

### 2.3.2 Massive parallelization for Hawkes model quantities

Sampling the Hawkes process parameters $\mu_0, \tau_x, \theta_0, \omega, h, \beta$ and event-specific rates $\boldsymbol{\theta}$ requires evaluation of the likelihood $\ell(\mathbf{X}, \mathbf{t} | \mu_0, \tau_x, \theta_0, \boldsymbol{\theta}, \omega, h)$ or its logarithm. Unfortunately, the double summation of Equation (1) results in an $\mathcal{O}(N^2)$ computational complexity that makes repeated likelihood evaluations all but impossible for the number of observations considered in this paper. We therefore use the high-performance computing framework of Holbrook *et al.* (2021b) to massively parallelize likelihood evaluations in the context of univariate, adaptive Metropolis-Hastings proposals for parameters $\mu_0, \tau_x, \theta_0, \omega$, $h$ and $\beta$. On the other hand, inference for the $M$-vector $\mathbf{z}$ requires more than fast univariate proposals, so we opt for HMC to sample from its high-dimensional posterior. Even in high dimensions, HMC efficiently generates proposal states by simulating a physical Hamiltonian system that renders the target posterior distribution invariant. Here, we follow standard procedure and specify the system with total energy:

$$
H(\mathbf{z}, \mathbf{p}) = -\log\left(\pi(\mathbf{z})\, \xi(\mathbf{p}|\mathbf{M})\right) \propto -\log \pi(\mathbf{z}) + \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} \,,
$$

where $\pi(\mathbf{z})$ is the density of the marginal posterior for $\mathbf{z}$, $\mathbf{p}$ is a Gaussian distributed 'momentum' variable with density $\xi(\mathbf{p}|\mathbf{M})$, and $\mathbf{M}$ is the system mass matrix and the covariance of $\mathbf{p}$. Within this Newtonian framework, simulating from the posterior distribution $\pi(\cdot)$ is analogous to simulating Hamiltonian dynamics that satisfy the equations:

$$
\dot{z} = \frac{\partial}{\partial \mathbf{p}} H(\mathbf{z}, \mathbf{p}) = \frac{1}{2} \mathbf{M}^{-1} \mathbf{p} \,, \quad \dot{p} = -\frac{\partial}{\partial \mathbf{z}} H(\mathbf{z}, \mathbf{p}) = \nabla \log \pi(\mathbf{z}) \,.
$$

Here, we interpret $\mathbf{z}$ as the position of a physical object with velocity $\dot{z}$ proportional to $\mathbf{M}^{-1}\mathbf{p}$, the momentum divided by mass. In the same way, the instantaneous change in momentum *&pdot;* is proportional to the gradient of the potential energy (or the negative log-posterior) with respect to the 'position' $\mathbf{z}$. But, there is no free lunch: simulation of the physical system requires repeated evaluations of the log-likelihood gradient, and these evaluations may become burdensome in big data contexts. We again follow standard HMC procedure and use the leapfrog algorithm (Leimkuhler and Reich, 2004) to integrate Hamilton's equations. Setting $\epsilon > 0$ small, at any time step $s$ within the numerical discretization scheme we update position and momentum according to the following rules:

$$
\begin{aligned}
\mathbf{p}(s + \epsilon/2) &= \mathbf{p}(s) + \frac{\epsilon}{2} \nabla \log \pi(\mathbf{z}(s)) \\
\mathbf{z}(s + \epsilon) &= \mathbf{z}(s) + \epsilon \mathbf{M}^{-1} \mathbf{p}(s + \epsilon/2) \\
\mathbf{p}(s + \epsilon) &= \mathbf{p}(s + \epsilon) + \frac{\epsilon}{2} \nabla \log \pi(\mathbf{z}(s + \epsilon)) \,.
\end{aligned}
$$

In practice, the step size $\epsilon$ is a crucially important tuning parameter, but we use standard adaptation techniques (Rosenthal, 2011) to avoid tedious hand-tuning. Unfortunately, the gradient of the Hawkes process log-likelihood of Equation (1) with respect to $\boldsymbol{\theta}$—a key term when calculating $\nabla \log \pi(\mathbf{z})$—becomes computationally onerous for large $N$ and $M$. Recalling that $\Lambda_m$ is the additive contribution of the $m$th event to the likelihood's integration term and that $\lambda_{nm}$ is the $m$th event's additive contribution to the rate function evaluated at the $n$th event, the gradient with respect to a single event-specific rate $\theta_m$ takes the form:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta_m} &= \frac{\partial}{\partial \theta_m} \left( -\Lambda_m + \sum_{n=1}^{N} \log \lambda_n \right) = -\frac{\partial \Lambda_m}{\partial \theta_m} + \sum_{t_m < t_n} \frac{1}{\lambda_n} \frac{\partial \lambda_{nm}}{\partial \theta_m} \\
&= \theta_0 \left( e^{-\omega(t_N - t_m)} - 1 \right) + \sum_{t_m < t_n} \frac{1}{\lambda_n} \frac{\theta_0 \omega}{h^D} e^{-\omega(t_n - t_m)} \phi\left( \frac{\mathbf{x}_n - \mathbf{x}_m}{h} \right) .
\end{aligned} \qquad (4)
$$

Given all observations and model parameters, one may compute the gradient using this formula, but the summation and $\lambda_n$ are both of complexity $\mathcal{O}(N)$. Thus, computing the entire vector $\partial \ell / \partial \boldsymbol{\theta} = (\partial \ell / \partial \theta_1, \ldots, \partial \ell / \partial \theta_M)^T$ requires time $\mathcal{O}(NM)$. Worse still, due to the multiscale nature of the posterior for the relative rates (Fig. 4), we find it necessary to precondition the Hamiltonian dynamics by specifying a diagonal mass matrix with elements:

$$
\mathbf{M}_{mm}^{-1} \approx -\frac{\partial^2 \ell}{\partial \theta_m^2} = \sum_{t_m < t_n} \frac{1}{\lambda_n^2} \frac{\theta_0^2 \omega^2}{h^{2D}} e^{-2\omega(t_n - t_m)} \phi^2 \left( \frac{\mathbf{x}_n - \mathbf{x}_m}{h} \right) . \qquad (5)
$$

Specifically, we maintain a running average of Hessians calculated at a fixed interval and use this as our preconditioner $\mathbf{M}$, thus maintaining asymptotic unbiasedness of Monte Carlo estimates (Haario *et al.*, 2001). Just as with the gradient, the summation and $\lambda_n$ are both of complexity $\mathcal{O}(N)$, and the resulting complexity for the entire Hessian is $\mathcal{O}(NM)$. To overcome these rate-limiting steps, we develop massively parallel central processing unit (CPU) and GPU implementations of both the gradient and the Hessian. In Section B, Algorithms 1 and 2 detail both parallel implementations of the gradient. Although our GPU-based implementations are fastest (Section 3.1), our CPU implementations are competitive, making use of both multi-core processing and SIMD (single instruction, multiple data) vectorization (Holbrook *et al.*, 2021a). Regardless of implementation, all of our high-performance software remains freely available for public use.

## 2.4 Software availability

We use the Bayesian evolutionary analysis by sampling trees (BEAST) software package (Suchard *et al.*, 2018), a popular tool for viral phylogenetic inference that implements Markov chain Monte Carlo (MCMC) methods to explore $p(\mathbf{S}, \phi, \mathcal{G})$ and $p(\mathbf{z} \,|\, \sigma^2, \mathcal{G})$ (Cybis *et al.*, 2015) under a range of evolutionary models. In writing this article, we have contributed to the open-source, stand-alone library hpHawkes http://github.com/suchard-group/hawkes for computing the spatiotemporal Hawkes process log-likelihood (Equation 1), its gradient (Equation 4), and its Hessian (Equation 5). hpHawkes integrates into BEAST with the help of an application programming interface. Within hpHawkes, we combine C++ code with which standard compilers generate vectorized CPU-specific instructions and OpenCL kernels that allow for GPU-specific optimization. Finally, we have used the Rcpp package (Eddelbuettel and François, 2011) to make the same massive parallelization speedups available to users of the R programming language.

# 3 Demonstration

## 3.1 Massive parallelization

Figure 2 shows benchmarking results for evaluating the Hawkes log-likelihood gradient with respect to event-specific rates $\boldsymbol{\theta}$ (Equation 4). For the GPU results, we use an NVIDIA Quadro GV100, which has 5120 CUDA cores (at 1.13 GHz) and reaches an (unboosted) 2.9 teraflops peak double-precision floating-point performance (or 5.8 teraflops for fused operations such as fused multiply-add). We use a Linux machine with two 26-core Intel Xeon Gold processors (2.1 GHz) for CPU results. Each physical core supports two threads or logical cores, and the machine achieves a peak performance of 874 gigaflops with double-precision floating point enhanced with advanced vector extensions (AVX) vectorization (again, double this for fused operations). Based on peak double-precision floating-point operations, our *a priori* expectation is for fully parallelized GPU-based gradient evaluations to be roughly 3.3 times faster than 104-threaded AVX evaluations on the CPU.

On the left of Figure 2, we compare relative efficiency for GPU and various CPU implementations of the log-likelihood gradient and Hessian for 25 000 simulated data points using single-threaded AVX computing (15.7 and 16.3 s per gradient and Hessian evaluations) as baseline. Using streaming SIMD extension (SSE) or non-vectorized single-threaded computing results in 1.5- and 2.2-fold slowdowns for the gradient and 1.3- and 2.1-fold slowdowns for the Hessian. Sticking with AVX processing, we see diminishing returns as we increase the number of threads. For both the gradient and the Hessian, the 14-, 54- and 104-thread AVX implementations are roughly 13, 33 and 41 times faster than single-threaded AVX. Agreeing with our *a priori* expectations, the GPU implementation is 140.4 times faster than single-threaded AVX and 3.5 times faster than 104-threaded AVX for the gradient and 120.0 times faster than single-threaded AVX and 2.9 times faster than 104-threaded AVX for the Hessian. The right of Figure 2 demonstrates the $\mathcal{O}(N^2)$ computational complexity for the same gradient and Hessian evaluations by varying the number of data points from 10 000 to 90 000. Although parallelization does not overcome this quadratic scaling, it does reduce computational costs for finite observation counts.

## 3.2 2014–2016 Ebola outbreak in West Africa

During the 2014–2016 outbreak in Guinea, Sierra Leone and Liberia, Ebola viral fever resulted in over 28 000 known cases and 11 000 known deaths (World Health Organization, 2015). First reports of the virus in Guinea emerged during March of 2014 (Baize *et al.*, 2014). At around the same time, viral cases with the same Guinean origin (Gire *et al.*, 2014) emerged in Sierra Leone and Liberia. In May 2014, the virus crossed from Guinea to Kailahun, Sierra Leone. From there, it spread to multiple counties of Liberia and Guinea (Dudas *et al.*, 2017), and the same strain reached Freetown, the capital of Sierra Leone, by July 2014. In the fall of 2014, Sierra Leone and Liberia were detecting 500 and 700 new cases a week. Only by the end of 2014 did case numbers begin to abate in most areas due to control measures. By March of 2015 sustained transmission of the virus only continued in western Guinea and western Sierra Leone (Dudas *et al.*, 2017). Figure 3 shows the spatiotemporal distribution of the majority of known Ebola virus cases during the epidemic. The right-hand side of Figure 3 is a stacked histogram, displaying the relative contribution of sequenced and unsequenced viral cases to the total case count.

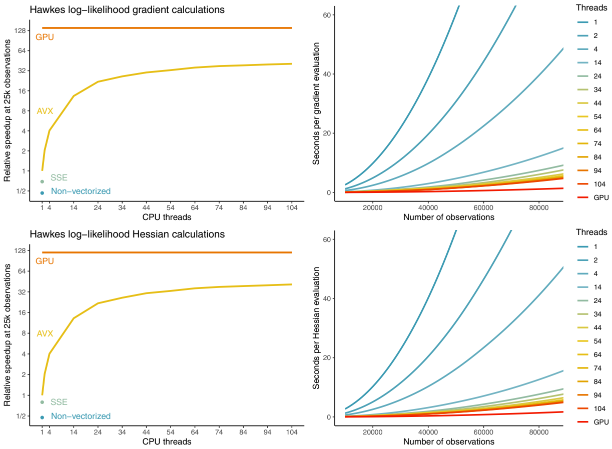

**Fig. 2.** Spatiotemporal Hawkes process log-likelihood gradient and Hessian calculations with respect to event-specific rates $\boldsymbol{\theta}$ with CPUs and GPUs. (Left) Multiplicative speedups over single-threaded AVX vectorization for single-threaded non-vectorized and SSE, multi-threaded AVX and many-core GPU processing for 25 000 randomly generated data points. (Right) Seconds per gradient and Hessian calculations for multi-threaded AVX and GPU implementations from 10 000 to 90 000 data points

Using our high-performance computing framework, we apply the phylogenetic Hawkes process to the analysis of 23 421 viral cases. Dudas *et al.* (2017) provide a total of 1610 cases furnishing genomic sequencing, 1367 of which come with date and location data (https://github.com/ebov/space-time). We supplement this sequenced data with 21 811 date and location pairs from unsequenced cases documented by the World Health Organization (https://apps.who.int/gho/data/node.ebola-sitrep). The precision of the spatial data is district or county level. To leverage spatial information as much as practically possible within our Hawkes model, we assume the locations follow a Gaussian distribution at district population centroids and with variance guaranteeing a 95% probability of the case occurring within the circle of equal area to the district and centered at the population centroid. We then integrate over uncertainty with respect to these locations by periodically sampling new locations according to the assumed Gaussian distribution throughout the MCMC run and with a period of roughly 100 iterations. That said, sensitivity analyses show that model inference is robust to fixing randomly generated locations for the entire MCMC chain. We make the combined data and documentation for our entire BEAST analysis available within the single file Final.xml and

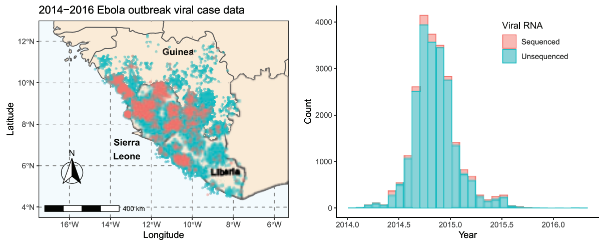

**Fig. 3.** Spatiotemporal distribution of 23 178 viral cases during the 2014–2016 Ebola outbreak in Guinea, Sierra Leone and Liberia. On the one hand, this number consists of 1367 viral samples that yield RNA sequence data and interface directly with the prior over phylogenetic trees. On the other hand, all 23 178 cases for which spatiotemporal data are available—including 21 811 unsequenced cases—interface with the Hawkes process likelihood. This leaves 243 sequenced cases for which spatiotemporal data are not available that interface with the phylogenetic but *not* the Hawkes process model

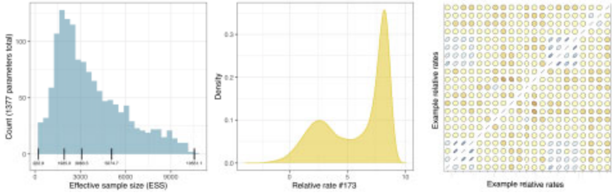

**Fig. 4.** The posterior distribution presents multiple challenges: it is high-dimensional; it takes on different scales for different parameters; it is multimodal in some parameters; and it exhibits complex correlation structures between parameters. (Left) Histogram and quartiles from 100 million MCMC samples for the ESS of all 1377 model parameters. (Middle) Multimodal marginal posterior for a single relative rate. (Right) Posterior correlations between 21 relative rates

**Table 2.** Posterior means and 95% HPD credible intervals from the application of the phylogenetic Hawkes process to the 2014–2016 Ebola outbreak in Guinea, Sierra Leone and Liberia

| | | | Posterior mean | |
|---|---|---|---|---|
| Hierarchical model module | Model parameter | Symbol | (95% HPD credible intervals) | Unit |
| Hawkes process | Background spatial lengthscale | $\tau_x$ | 194 (147, 243) | km |
| | Self-excitatory temporal lengthscale | $1/\omega$ | 29.8 (28.5, 30.9) | days |
| | Self-excitatory spatial lengthscale | $h$ | 7.37 (7.13, 7.62) | km |
| | Normalized self-excitatory weight | $\theta_0/(\theta_0+\mu_0)$ | 0.96 (0.95, 0.97) | — |
| | Temporal trend coefficient | $\beta$ | −2.22 (−2.37, −2.06) | — |
| Phylogenetic diffusion | SD | $\sigma$ | 51.0 (46.4, 55.7) | log rate |

place this as well as other project scripts together at the repository https://github.com/suchard-group/EBOVPhyloHawkes. In addition to the software mentioned in the previous section, we make use of the ggplot2 and ggmap R packages for data and results visualization (Kahle and Wickham, 2013; Wickham, 2016).

For the phylogenetic prior specification $p(\mathbf{S}, \phi, \mathcal{G})$, we follow the phylogeographic analysis of Dudas *et al.* (2017) and use a mixture of 1000 phylogenetic trees obtained as high-probability posterior samples from their purely phylogenetic analysis of the 1610 sequenced viral samples. In that preceding Bayesian analysis, Dudas *et al.* (2017) combine an HKY+$\Gamma_4$ substitution model prior for molecular evolution (Hasegawa *et al.*, 1985; Yang, 1994), a relaxed molecular clock prior on rates (Drummond *et al.*, 2006), a nonparametric coalescent 'Skygrid' prior on effective population size dynamics (Gill *et al.*, 2013) and a continuous time Markov chain reference prior for overall rate (Ferreira and Suchard, 2008). We assume *a priori* that the background lengthscale $\tau_x$ follows a diffuse inverse gamma distribution with Shape 1 and Scale 10, where distance units are latitudinal and longitudinal degrees. An inverse gamma distribution with shape and scale parameters equal to 2 and 0.5 for both $h$ and $1/\omega$ encodes our beliefs that self-excitatory dynamics occur at finer spatiotemporal scales, where years are the temporal units. We upweight self-excitatory dynamics by giving $\theta_0$ and $\mu_0$ gamma priors with shape parameters 1 and 2 and scale parameters 0.001 and 2, respectively. We absorb $\tau_0$ into $\sigma$ and place a tight inverse gamma prior on $1/\sigma$ with shape and scale parameters of 2 and 0.5. Finally, we set $\mathbf{f}(t_n) = t_n$ and place a normal prior on the univariate coefficient $\beta$ with mean 0 and SD of 10. We find all parameters robust to prior specification due to the large number of observations considered.

We generate 100 million MCMC samples according to the routine outlined in Section 2.3 and discard the first 500 000 as burn-in. Using our parallel computing algorithms and a single NVIDIA GV100 GPU (Section 3.1), the routine requires 6.77 h to generate 1 million samples and 28 days to generate all 100 million samples. Figure 4 shows the distribution of effective samples sizes (ESS) across all model parameters and illustrates some of the challenges facing any MCMC routine for the phylognetic Hawkes model. Namely, the posterior distribution is high-dimensional, multimodal, multiscale and has complex correlation structures.

Table 2 shows posterior means and 95% highest posterior density (HPD) credible intervals for the phylogenetic Hawkes process parameters. The posterior mean for the spatial bandwidth $\tau_x$ of the Hawkes background process is 194 km (147, 243), allowing the model to incorporate and adapt to large scale geographic movement. On the other hand, the Hawkes process self-excitatory spatial bandwidth $h$ has a posterior mean of 7.4 km (7.1, 7.6), indicating the smaller local scale for which the model attributes viral contagion. The self-excitatory temporal bandwidth $1/\omega$ has a posterior mean of 29.8 days (28.5, 30.9), indicating the timescale for which the model attributes the same viral contagion. The normalized self-excitatory weight $\theta_0/(\theta_0 + \mu_0)$ indicates the proportion of events the model attributes to self-excitatory (compared with background) dynamics and has a posterior mean of 0.96 (0.95, 0.97). The posterior mean of the self-excitatory rate's temporal trend coefficient $\beta$ is −2.22 (−2.37, −2.06) indicates that, for every additional year and *ceteris paribus*, one should expect a multiplicative decrease of $1 - \exp(-2.22) \times 100 \approx 90\%$ to the process self-excitatory rate. In this way, the model adjusts for downward trends arising from epidemiological control (e.g. mass quarantine and travel restrictions) and controls for these factors when inferring virus-specific relative rates.

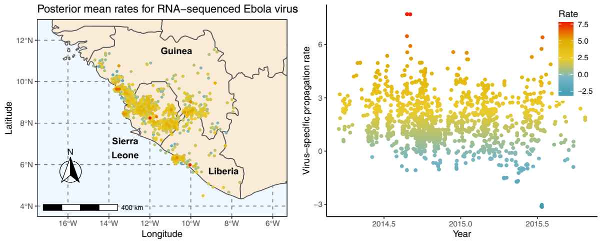

**Fig. 5.** Hawkes model posterior mean rates $\boldsymbol{\theta}$ for the 1367 (of 1610) RNA-sequenced viral samples for which date/location data are available. Unsurprisingly, the largest relative rates occur within or nearby major clusters of events. Adjusting for downward trends in case data with a negative coefficient $\beta$ (Table 2) allows detection of higher relative rates after peak outbreak (late 2014) including a jump in infections mid-2015 (Figure 3)

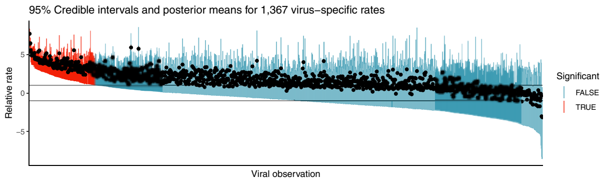

**Fig. 6.** 95% credible intervals and posterior means for virus-specific rates $\boldsymbol{\theta}$ corresponding to the 1367 sequenced viruses that come with date/location data and therefore appear in the Hawkes process module. We call those 183 intervals which do not include 1 'significant' and color the 177 intervals that are above 1 red

Next, we consider posterior inference of the virus-specific rates $\boldsymbol{\theta}$ for those viral observations that provide RNA sequences. When interpreting these results, it is important to understand that the phylogenetic Hawkes process implicitly assumes that such samples spark nearby contagion as described by spatial and temporal bandwidths $h$ and $1/\omega$. Recall that the posteriors for these two parameters concentrate at over 7 km and 4 weeks, respectively. Figure 5 depicts the relationship between posterior mean values of $\boldsymbol{\theta}$ and the spatiotemporal distribution of corresponding viruses. Since these rates represent multiplicative factors of the global self-excitatory weight $\theta_0$, a null value would be 1. Posterior means range from approximately −2.5 to 7.5 and increasingly vary as a function of time. As one might expect, the highest rates appear near or within larger clusters. Thanks to the negative temporal trend coefficient $\beta$ and the increase of uncertainty with time, larger rate values do obtain for some viral cases occurring in 2015, despite following after peak epidemic. Figure 6 features posterior means and 95% intervals for the same virus-specific rates. Only a small subset of 183 rate intervals does not include 1. Of these, 177 have lower bound >1. We

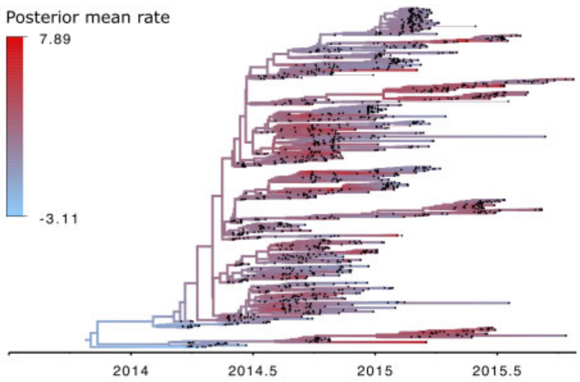

**Fig. 7.** Posterior mean virus-specific relative rates color the posterior maximum clade credibility tree of phylogeny $\mathcal{G}$, a few subtrees of which potentially demonstrate elevated contagiousness

interpret all 183 of the corresponding viruses as having statistically significantly increased or decreased contagiousness.

Finally, Figure 7 shows how these posterior rates organize as a function of the inferred posterior maximum clade credibility tree $\mathcal{G}$. Generally speaking, shorter branch lengths indicate larger effective populations of viruses, whereas larger branch lengths indicate smaller. For example, the structure of the bottom subtree reflects this intuition as branches are short with many splits during peak outbreak in the second half of 2014 but become mostly long in late 2014 and the remainder of 2015. According to the phylogenetic Brownian process model outlined in Section 2.2, virus-specific rate values are more highly correlated to one another when closely located to one-another on the phylogenetic tree. It is plausible that these correlations allow the phylogenetic Hawkes model to infer higher rates for some strains that survive late into the epidemic despite dropping case counts. The model attributes some of the highest values to strains appearing in Coyah, Conakry and Kindia, Guinea, in late 2014 and early 2015. Interestingly, the model also attributes its lowest values to cases in Kono, Sierra Leone, in early 2015. Taken together, Figures 6 and 7 may provide helpful leads for epidemiologists searching for variants with heightened relative rates of contagion.

## 4 Discussion

We propose the phylogenetic Hawkes process, a Bayesian hierarchical model that relates viral spatial contagion to molecular evolution by uniting the two epidemiological paradigms of self-exciting point process and phylogenetic modeling. Due to difficulties in scaling the model to larger numbers of observations, we advance a computing strategy that combines HMC, dynamic programming and massive parallelization for key inferential bottlenecks. Finally, we apply our novel model and high-performance computing framework to the analysis of over 23 000 viral cases arising from the 2014 to 2016 Ebola outbreak in West Africa, and Ebola strains and subtrees with plausibly higher degrees of contagiousness reveal themselves.

Unfortunately, the current model will fail when applied to the analysis of a global pandemic due to the dominant role of non-local, large-scale transportation networks in propagating viral spread (Holbrook *et al.*, 2020, 2021a). We are particularly interested in developing extensions to the phylogenetic Hawkes process that leverage recent advances in scaling high-dimensional multivariate Hawkes processes (Nickel and Le, 2020) and applying the resulting multivariate phylogenetic Hawkes process to the analysis of global pandemics. In this context, each additional dimension will represent an additional country or province. Prodigious computational challenges are inevitable, and we suspect that non-trivial GPU implementations will be necessary for big data applications.

Moving beyond inference, a major question is whether the phylogenetic Hawkes process can be useful for prediction of spatial contagion and dynamics. Here, recent neural network extensions of the Hawkes process might prove useful (Mei and Eisner, 2017; Zhang *et al.*, 2020; Zuo *et al.*, 2020), but it is unclear what forward simulation of phylogenetic branching dynamics would look like in the context of a Hawkes process. Moreover, generating samples from the posterior predictive distribution of a Hawkes process would be extremely time consuming when one is conditioning on millions of posterior samples. To work around this, one could perhaps parallelize over fixed parameter settings the algorithm of Dassios and Zhao (2011) for simulating Hawkes processes when the temporal triggering function is exponential. Such an implementation would require efficient use of parallel pseudo-random number generators (Salmon *et al.*, 2011).

More broadly, the phylogenetic Hawkes process is a single contribution to the immense scientific project surrounding surveillance of viral populations, inference of viral evolutionary histories and prediction of viral spread. Due to the inherent difficulty of this challenge, we imagine our model will be most useful when used in conjunction with other epidemiological tools, mathematical models and expert intensive lab work. That said, we believe the phylogenetic Hawkes process represents a major step toward automated, holistic and data-centered epidemiological modeling insofar as it fully leverages genomic and spatiotemporal data obtained from both sequenced and unsequenced viral cases. This joint modeling approach stands in contrast to other phylogenetic modeling approaches that also attempt to recover phylogenetically localized measures of contagiousness but do not leverage unsequenced case data. For example, Łuksza *et al.* (2014) combine summary statistics of clade growth with birth–death models to infer clade-specific reproduction numbers but analyze only 81 sequenced cases and no unsequenced cases from the early 2014 Ebola outbreak. On the one hand, the birth–death model allows one to obtain readily interpretable clade-specific reproduction numbers. On the other hand, the phylogenetic Hawkes model allows rates to vary continuously between *and* within clades. Similarly, Stadler *et al.* (2014) and Volz and Pond (2014) apply phylogenetic compartmental models to analyze $<$80 sequenced cases and no unsequenced cases. The fact that these analyses (Łuksza *et al.*, 2014; Stadler *et al.*, 2014; Volz and Pond, 2014) do not take spatial information into account makes it difficult to compare certain results with those from the phylogenetic Hawkes process analysis. For temporal measures, however, we do infer a self-excitatory lengthscale $1/\omega$ with 95% HPD credible interval of (28.5, 30.9) days, which is marginally consistent with Stadler *et al.*'s (2014) 95% HPD credible intervals of (2.11, 23.20) days for the incubation period and (1.24, 6.98) days for the infectious period.

These comparisons raise a broader theoretical question: can the phylogenetic Hawkes process provide phylogenetically localized basic or effective reproduction numbers and, as a result, become more readily interpretable in broader science? Rizoiu *et al.* (2018) show that the rate of events in an epidemiological SIR model is equal to that of their finite-population Hawkes model. Accordingly, crucial next steps in the development of the phylogenetic Hawkes process framework are the incorporation of finite-population dynamics as well as the extension of the temporal model from Rizoiu *et al.* (2018) to the spatiotemporal regime.

## Acknowledgements

We gratefully acknowledge support from NVIDIA Corporation and Advanced Micro Devices, Inc. with the donation of parallel computing resources used for this research.

## Funding

The research leading to these results has received funding through National Institutes of Health grants [K25 AI153816, U19 AI135995 and R01 AI153044] and National Science Foundation grant [DMS1264153].

*Conflict of Interest*: none declared.

## References

Bacry,E. *et al.* (2015) Hawkes processes in finance. *Market Microstruct. Liq.*, **1**, 1550005.

Baize,S. *et al.* (2014) Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.*, **371**, 1418–1425.

Bertozzi,A.L. *et al.* (2020) The challenges of modeling and forecasting the spread of covid-19. *Proc. Natl. Acad. Sci. USA*, **117**, 16732–16738.

Boni,M.F. *et al.* (2020) Evolutionary origins of the SARS-COV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.*, **5**, 1408–1417.

Brockmann,D. and Helbing,D. (2013) The hidden geometry of complex, network-driven contagion phenomena. *Science*, **342**, 1337–1342.

Cavalli-Sforza,L.L. and Edwards,A.W. (1967) Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.*, **19**, 233–257.

Chiang,W.-H. *et al.* (2020). Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *medRxiv. https://doi.org/10.1016/j.ijforecast.2021.07.001.*

Cybis,G. *et al.* (2015) Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.*, **9**, 969– 991.

Daley,D.J. and Jones,D.V. (2003). *An Introduction to the Theory of Point Processes: Elementary Theory of Point Processes*. Springer, New York.

Dassios,A. and Zhao,H. (2011) A dynamic contagion process. *Adv. Appl. Prob.*, **43**, 814–846.

Drummond,A. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.

Dudas,G. *et al.* (2017) Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*, **544**, 309–315.

Eddelbuettel,D. and François,R. (2011) Rcpp: seamless R and C++ integration. *J. Stat. Softw.*, **40**, 1–18.

Faria,N.R. *et al.* (2014) The early spread and epidemic ignition of HIV-1 in human populations. *Science*, **346**, 56–61.

Felsenstein,J. (1978) The number of evolutionary trees. *Syst. Zool.*, **27**, 27–33.

Felsenstein,J. (1985) Phylogenies and the comparative method. *Am. Nat.*, **125**, 1–15.

Ferreira,M.A. and Suchard,M.A. (2008) Bayesian analysis of elapsed times in continuous-time Markov chains. *Can. J. Stat.*, **36**, 355–368.

Fisher,A.A. *et al.* (2021) Relaxed random walks at scale. *Syst. Biol.*, **70**, 258–267.

Fox,E.W. *et al.* (2016) Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Ann. Appl. Stat.*, **10**, 1725–1756.

Freckleton,R.P. (2012) Fast likelihood calculations for comparative analyses. *Methods Ecol. Evol.*, **3**, 940–947.

Gill,M.S. *et al.* (2013) Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol. Biol. Evol.*, **30**, 713–724.

Gire,S.K. *et al.* (2014) Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, **345**, 1369–1372.

Haario,H. *et al.* (2001) An adaptive metropolis algorithm. *Bernoulli*, **7**, 223–242.

Habbema,J. Hermans. J. and *Van Den Bkoek. K.*. (1974). A stepwise discriminant analysis program using density estimation. In: Bruckman (ed), Comjwtat. Physica Verlag, Vienna, pp. 101–110.

Hasegawa,M. *et al.* (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

Hawkes,A.G. (1972) Spectra of some mutually exciting point processes with associated variables. In: *Stochastic Point Processes*, pp. 261–271.

Hawkes,A.G. (1971a) Point spectra of some mutually exciting point processes. *J. R. Stat. Soc. B*, **33**, 438–443.

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

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

Hawkes,A.G. and Adamopoulos,L. (1973) Cluster models for earthquakes-regional comparisons. *Bull. Int. Stat. Inst.*, **45**, 454–461.

Ho,L.S.T. and Ané,C. (2014) A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.*, **3**, 397–402.

Holbrook,A.J. *et al.* (2020). Bayesian mitigation of spatial coarsening for a fairly flexible spatiotemporal Hawkes model. *arXiv preprint arXiv: 2010.02994.*

Holbrook,A.J. *et al.* (2021a) Massive parallelization boosts big Bayesian multidimensional scaling. *J. Comput. Graph. Stat.*, **30**, 11–24.

Holbrook,A.J. *et al.* (2021b) Scalable Bayesian inference for self-excitatory stochastic processes applied to big American gunfire data. *Stat. Comput.*, **31**, 1–15.

Kahle,D. and Wickham,H. (2013) ggmap: spatial visualization with ggplot2. *R J.*, **5**, 144–161.

Kelly,J.D. *et al.* (2019) Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the democratic Republic of the Congo using Hawkes point process models. *Epidemics*, **28**, 100354.

Kim,H. (2011). Spatio-temporal point process models for the spread of avian influenza virus (H5N1). PhD Thesis, UC Berkeley.

Kobayashi,R. and Lambiotte,R. (2016). TiDeH: time-dependent Hawkes process for predicting retweet dynamics. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 10.

Leimkuhler,B. and Reich,S. (2004). *Simulating Hamiltonian Dynamics*, Vol. **14**. Cambridge University Press, Cambridge, UK.

Lemey,P. *et al.* (2009) Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, **5**, e1000520.

Lemey,P. *et al.* (2010) Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.*, **27**, 1877–1885.

Loeffler,C. and Flaxman,S. (2018) Is gun violence contagious? A spatiotemporal test. *J. Quant. Criminol.*, **34**, 999–1017.

Łuksza,M. *et al.* (2014). Epidemiological and evolutionary analysis of the 2014 Ebola virus outbreak. *arXiv preprint arXiv:1411.1722. https://doi.org/10.1101/011171.*

Mau,B. *et al.* (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1–12.

Mei,H. and Eisner,J.M. (2017) The neural Hawkes process: a neurally self-modulating multivariate point process. In: *Advances in Neural Information Processing Systems*, pp. 6754–6764.

Meyer,S. and Held,L. (2014) Power-law models for infectious disease spread. *Ann. Appl. Stat.*, **8**, 1612–1639.

Mohler,G. (2013) Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.*, **7**, 1525–1539.

Mohler,G. (2014) Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.*, **30**, 491–497.

Neal,R.M. (2011) MCMC using Hamiltonian dynamics. In: *Handbook of Markov Chain Monte Carlo*, Chapman and Hall / CRC, Boca Raton, Vol. **2**.

Nickel,M. and Le,M. (2020). Learning multivariate Hawkes processes at scale. *arXiv preprint arXiv:2002.12501.*

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

Park,J. *et al.* (2019). Investigating clustering and violence interruption in gang-related violent crime data using spatial-temporal point processes with covariates. J. Am. Stat. Assoc., **116**, 1674–1687.

Pearl,J. (1982). Reverend Bayes on inference engines: a distributed hierarchical approach. In: AAAI-82: Proceedings of the Second National Conference on Artificial Intelligence, pp. 133–136.

Pybus,O.G. *et al.* (2012) Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA*, **109**, 15066–15071.

Rambaut,A. *et al.* (2008) The genomic and epidemiological dynamics of human influenza a virus. *Nature*, **453**, 615–619.

Reinhart,A. (2018) A review of self-exciting spatio-temporal point processes and their applications. *Stat. Sci.*, **33**, 299–318.

Rizoiu,M.-A. *et al.* (2017). A tutorial on Hawkes processes for events in social media. *arXiv preprint arXiv:1708.06401.*

Rizoiu,M.-A. *et al.* (2018). Sir-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web, International World Wide Web Conferences Steering Committee*, pp. 419–428.

Robert,P. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.*, **25**, 1175–1179.

Ronquist,F. *et al.* (2012) Mrbayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.

Rosenthal,J.S. (2011) Optimal proposal distributions and adaptive MCMC. In: *Handbook of Markov Chain Monte Carlo*, Chapman and Hall / CRC, Boca Raton, Vol. 4.

Salmon,J.K. *et al.* (2011). Parallel random numbers: as easy as 1, 2, 3. In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–12.

Schoenberg,F.P. (2004) Testing separability in spatial-temporal marked point processes. *Biometrics*, **60**, 471–481.

Schoenberg,F.P. (2013) Facilitated estimation of ETAs. *Bull. Seismol. Soc. Am.*, **103**, 601–605.

Schoenberg,F.P. (2020). Nonparametric estimation of variable productivity Hawkes processes. *arXiv preprint arXiv:2003.08858.*

Schoenberg,F.P. *et al.* (2019) A recursive point process model for infectious diseases. *Ann. Inst. Stat. Math.*, **71**, 1271–1287.

Sinsheimer,J.S. *et al.* (1996) Bayesian hypothesis testing of four-taxon topologies using molecular sequence data. *Biometrics*, **52**, 193–210.

Smith,G.J. *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122–1125.

Stadler,T. *et al.* (2014) Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Curr.*, **6**.

Suchard,M. and Rambaut,A. (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics*, **25**, 1370–1376.

Suchard,M. *et al.* (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.*, **18**, 1001–1013.

Suchard,M. *et al.* (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.*, **52**, 649–664.

Suchard,M.A. *et al.* (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.*, **4**, vey016.

Volz,E. and Pond,S. (2014) Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr.*, **6**,

Wadman,M. (2021) United states rushes to fill void in viral sequencing. *Science*, **371**, 657–658.

Wickham,H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York.

World Health Organization. (2015). WHO: Ebola situation report 30 December 2015.

Yang,Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.

Yang,Z. and Rannala,B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, **14**, 717–724.

Yuan,B. *et al.* (2021) Fast estimation of multivariate spatiotemporal Hawkes processes and network reconstruction. *Ann. Inst. Stat. Math.*, **73**, 1127–1152.

Zhang,Q. *et al.* (2020). Self-attentive Hawkes process. In: *International Conference on Machine Learning, PMLR*, pp. 11183–11193.

Zhuang,J. *et al.* (2004) Analyzing earthquake clustering features by using stochastic reconstruction. *J. Geophys. Res. Solid Earth*, **109**.

Zuo,S. *et al.* (2020). Transformer Hawkes process. In: *International Conference on Machine Learning, PMLR*, pp. 11692–11702.

## Appendix A

**Algorithm 1** Parallel evaluation of Hawkes process log-likelihood gradient: *Makes use of multiple CPU cores and loop vectorization to calculate Hawkes process log-likelihood gradient with respect to event-specific rates $\boldsymbol{\theta}$. When using double-precision floating point, this algorithm may use either SSE or AVX vectorization to make $J = 2$- or 4-long jumps. We denote the number of CPU cores as B. Symbols $\ell$, $\lambda$ and $\Lambda$ appear in Equations (1) and (4).*

```
1: Compute rates $\lambda_1, \ldots, \lambda_N$:
    parfor $b \in \{1, \ldots, B\}$ do
      if $b \neq B$ then
        $Upper \leftarrow b\lfloor N/B \rfloor$
      else
        $Upper \leftarrow \lceil N/B \rceil$
      end if
      for $n \in \{(b-1)\lfloor N/B \rfloor + 1, \ldots, Upper\}$ do
        copy $\mathbf{x}_n$, $t_n$ to cache
        $\boldsymbol{\lambda} n_n \leftarrow 0$          ▷ vector of length J
        $n' \leftarrow 1$
        while $n' < N$ do
          $J \leftarrow \min(J, N - n')$
            copy $\mathbf{x}_{n':(n'+J)}$, $t_{n':(n'+J)}$ to cache
            $\Delta_{nn'} : \Delta_{nn':(n'+J-1)} \leftarrow (\mathbf{x}_n - \mathbf{x}_{n'}) : (\mathbf{x}_n - \mathbf{x}_{n'+J-1})$ ▷ vectorized subtraction
            calculate $\delta_{nn'} : \delta_{n(n'+J-1)}$ ▷ vectorized multiplication
            calculate $\lambda_{nn'} : \lambda_{n(n'+J-1)}$ ▷ vectorized evaluation
            $\boldsymbol{\lambda} n_n \leftarrow \boldsymbol{\lambda} n_n + \lambda_{nn'} : \lambda_{n(n'+J-1)}$ ▷ vectorized addition
          $n' \leftarrow n' + J$
        end while
      end for
    end parfor
2: Compute $M$ gradients $\frac{\partial \ell}{\partial \theta_n}$:
    parfor $b \in \{1, \ldots, B\}$ do
      if $b \neq B$ then
        $Upper \leftarrow b\lfloor M/B \rfloor$
      else
        $Upper \leftarrow \lceil M/B \rceil$
      end if
      for $n \in \{(b-1)\lfloor M/B \rfloor + 1, \ldots, Upper\}$ do
        copy $\mathbf{x}_n$, $t_n$ to cache
        $\frac{\partial \ell}{\partial \theta_n} \leftarrow 0$
        $n' \leftarrow 1$
        while $n' < N$ do
          $J \leftarrow \min(J, N - n')$
            copy $\mathbf{x}_{n':(n'+J)}$, $t_{n':(n'+J)}$ to cache
            $\Delta_{nn'} : \Delta_{nn':(n'+J-1)} \leftarrow (\mathbf{x}_n - \mathbf{x}_{n'}) : (\mathbf{x}_n - \mathbf{x}_{n'+J-1})$ ▷ vectorized subtraction
            calculate $\delta_{nn'} : \delta_{n(n'+J-1)}$ ▷ vectorized multiplication
            calculate $e^{-\omega(t_{n'}-t_n)}\phi\left(\frac{\delta_{nn'}}{h}\right) : e^{-\omega(t_{n'+J-1}-t_n)}\phi\left(\frac{\delta_{n(n'+J-1)}}{h}\right)$ ▷ vectorized evaluation
          for $j \in n', \ldots, n' + J - 1$ do
            $\frac{\partial \ell}{\partial \theta_n} \leftarrow \frac{\partial \ell}{\partial \theta_n} + \mathcal{I}_{[t_n < t_j]} \frac{1}{\lambda_j} \frac{\partial \lambda_{jn}}{\partial \theta_n}$
          end for
          $n' \leftarrow n' + J$
        end while
        $\frac{\partial \ell}{\partial \theta_n} \leftarrow \frac{\partial \ell}{\partial \theta_n} + \theta_0 (e^{-\omega(t_N - t_n)} - 1)$
      end for
    end parfor
```

## A1 The likelihood integral with event-specific rates

Without loss of generality, we consider the temporal Hawkes process with constant background rate. To compute the likelihood (Equation 1), we must calculate the integral:

$$\begin{aligned}
\Lambda(t_N) &= \int_0^{t_N} \lambda(t)\,dt = \int_0^{\mathrm{t_N}} \left(\mu + \theta_0 \textstyle\sum_{\mathrm{t_n} < \mathrm{t}} \theta_\mathrm{n} \omega \mathrm{e}^{-\omega(\mathrm{t}-\mathrm{t_n})}\right) \mathrm{dt} \\
&= \int_0^{t_1} \mu\,\mathrm{d}t + \sum_{n=1}^{N-1} \int_{t_n}^{t_{n+1}} \left(\mu + \theta_0 \sum_{t_n < t} \theta_n \omega e^{-\omega(t-t_n)}\right) \mathrm{d}t \\
&= \mu\, t_N + \theta_0 \omega \sum_{n=1}^{N-1} \int_{t_n}^{t_{n+1}} \sum_{n'=1}^{n} \theta_{n'} e^{-\omega(t-t_{n'})}\,\mathrm{d}t \\
&= \mu\, t_N + \theta_0 \omega \sum_{n=1}^{N-1} \sum_{n'=1}^{n} \theta_{n'} \int_{t_n}^{t_{n+1}} e^{-\omega(t-t_{n'})}\,\mathrm{d}t \\
&= \mu\, t_N - \theta_0 \sum_{n=1}^{N-1} \sum_{n'=1}^{n} \theta_{n'} (e^{-\omega(t_{n+1}-t_{n'})} - e^{-\omega(t_n - t_{n'})}),
\end{aligned}$$

and we further simplify the double summation in the following.

Claim 1. *The temporal Hawkes process with rate function:*

$$\lambda(t) = \mu + \theta_0 \sum_{t_n < t} \theta_n \omega e^{-\omega(t-t_n)} \tag{6}$$

*admits the integral*

$$\Lambda(t_N) = \int_0^{t_N} \lambda(t)\,dt = \mu\,\mathrm{t_N} - \theta_0 \sum_{\mathrm{n}=1}^{\mathrm{N}-1} \theta_\mathrm{n} (\mathrm{e}^{-\omega(\mathrm{t_N}-\mathrm{t_n})} - 1)\,. \tag{7}$$

Proof. Proceeding by induction, the assertion is trivial for $N = 1$. If it is true for some $N > 0$, this implies that:

$$\sum_{n=1}^{N-1} \sum_{n'=1}^{n} \theta_{n'} (e^{-\omega(t_{n+1}-t_{n'})} - e^{-\omega(t_n-t_{n'})}) = \sum_{n=1}^{N-1} \theta_n (e^{-\omega(t_N - t_n)} - 1)\,.$$

It follows that:

$$\begin{aligned}
\Lambda(t_{N+1}) &= \mu t_{N+1} - \theta_0 \sum_{n=1}^{N} \sum_{n'=1}^{n} \theta_{n'} \left(e^{-\omega(t_{n+1}-t_{n'})} - e^{-\omega(t_n - t_{n'})}\right) \\
&= \mu(t_{N+1} - t_N) - \theta_0 \sum_{n'=1}^{N} \theta_{n'} \left(e^{-\omega(t_{N+1}-t_{n'})} - e^{-\omega(t_N - t_{n'})}\right) \\
&\quad + \mu t_N - \theta_0 \sum_{n=1}^{N-1} \sum_{n'=1}^{n} \theta_{n'} \left(e^{-\omega(t_{n+1}-t_{n'})} - e^{-\omega(t_n - t_{n'})}\right) \\
&= \mu(t_{N+1} - t_N) - \theta_0 \sum_{n'=1}^{N} \theta_{n'} \left(e^{-\omega(t_{N+1}-t_{n'})} - e^{-\omega(t_N - t_{n'})}\right) + \Lambda(t_n) \\
&= \mu(t_{N+1} - t_N) - \theta_0 \sum_{n'=1}^{N} \theta_{n'} \left(e^{-\omega(t_{N+1}-t_{n'})} - e^{-\omega(t_N - t_{n'})}\right) \\
&\quad + \mu t_N - \theta_0 \sum_{n=1}^{N-1} \theta_n \left(e^{-\omega(t_N - t_n)} - 1\right) \\
&= \mu t_{N+1} - \theta_0 \sum_{n=1}^{N} \theta_n \left(e^{-\omega(t_{N+1}-t_n)} - e^{-\omega(t_N - t_n)}\right) \\
&\quad - \theta_0 \sum_{n=1}^{N-1} \theta_n \left(e^{-\omega(t_N - t_n)} - 1\right) \\
&= \mu t_{N+1} - \theta_0 \sum_{n=1}^{N} \theta_n \left(e^{-\omega(t_{N+1}-t_n)} - 1\right),
\end{aligned}$$

thus completing the proof.

## Appendix B

**Algorithm 2** Parallel evaluation of Hawkes process log-likelihood gradient: *Computes the log-likelihood gradient with respect to event-specific rates $\boldsymbol{\theta}$ using multiple levels of parallelization on a GPU. In this article, we specify $B = 128$ for the size of the GPU work groups. Symbols $\ell$, $\lambda$ and $\Lambda$ appear in Equations (1) and (4).*

```
1: Compute rates $\lambda_1, \ldots, \lambda_N$:
    parfor $n \in \{1, \ldots, N\}$ do
        copy $\mathbf{x}_n$, $t_n$ to local        ▷ $B$ threads
        parfor $N' \in \{1, \ldots, \lfloor N/B \rfloor\}$ do
            $n' \leftarrow N'$
            $\lambda_{nN'} \leftarrow 0$
            while $n' < N$ do
                copy $\mathbf{x}_{n'}$, $t_{n'}$ to local        ▷ $B$ threads
                $\Delta_{nn'} \leftarrow \mathbf{x}_n - \mathbf{x}_{n'}$        ▷ vectorized subtraction
                calculate $\delta_{nn'} = \sqrt{\sum \Delta_{nn'} \circ \Delta_{nn'}}$        ▷ vectorized multiplication
                $\lambda_{nN'} \leftarrow \lambda_{nN'} + \lambda_{nn'}$        ▷ $\lambda_{nn'}$ a function of $\delta_{nn'}$, $t_n$ and $t_{n'}$
                $n' \leftarrow n' + B$
            end while
        end parfor
        $\lambda_n \leftarrow \sum_{N'} \lambda_{nN'}$        ▷ binary tree reduction on chip
    end parfor
2: Compute $M$ gradients $\frac{\partial \ell}{\partial \theta_n}$:
    parfor $n \in \{1, \ldots, M\}$ do
        copy $\mathbf{x}_n$, $t_n$ to local        ▷ $B$ threads
        parfor $N' \in \{1, \ldots, \lfloor N/B \rfloor\}$ do
            $n' \leftarrow N'$
            $\left(\frac{\partial \ell}{\partial \theta_n}\right)_{N'} \leftarrow 0$
            while $n' < N$ do
                copy $\mathbf{x}_{n'}$, $t_{n'}$ to local        ▷ $B$ threads
                $\Delta_{nn'} \leftarrow \mathbf{x}_n - \mathbf{x}_{n'}$        ▷ vectorized subtraction
                calculate $\delta_{nn'} = \sqrt{\sum \Delta_{nn'} \circ \Delta_{nn'}}$        ▷ vectorized multiplication
                $\left(\frac{\partial \ell}{\partial \theta_n}\right)_{N'} \leftarrow \left(\frac{\partial \ell}{\partial \theta_n}\right)_{N'} + \mathcal{I}_{[t_n < t_{n'}]} \frac{1}{\lambda_{n'}} \frac{\partial \lambda_{n'n}}{\partial \theta_n}$
                $n' \leftarrow n' + B$
            end while
        end parfor
        $\frac{\partial \ell}{\partial \theta_n} \leftarrow \sum_{N'} \left(\frac{\partial \ell}{\partial \theta_n}\right)_{N'}$        ▷ binary tree reduction on chip
        $\frac{\partial \ell}{\partial \theta_n} \leftarrow \frac{\partial \ell}{\partial \theta_n} + \theta_0 \left(e^{-\omega(t_N - t_n)} - 1\right)$
    end parfor
```

### B1 Parallelized gradient algorithms

Algorithms 1 and 2 present instructions for computing the Hawkes process log-likelihood gradient (Equation 4) with respect to the $M$-vector $\boldsymbol{\theta}$ of event-specific rates. Figure 2 shows resulting speedups for both Algorithms (1) and (2) on a CPU and GPU, respectively.

## Appendix C

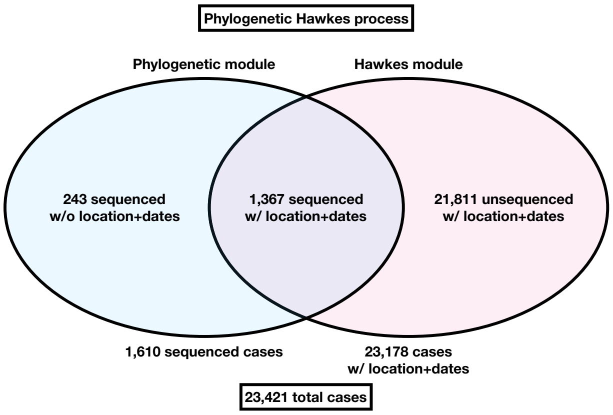
Phylogenetic Hawkes process
Phylogenetic module
Hawkes module
243 sequenced w/o location+dates
1,367 sequenced w/ location+dates
21,811 unsequenced w/ location+dates
1,610 sequenced cases
23,178 cases w/ location+dates
23,421 total cases

**Fig. 8.** The modules of the hierarchical model and the distribution of the data that interface with them

### C1 Data and modules

Figure 8 shows the number of observations that interface with the phylogenetic and Hawkes modules of the phylogenetic Hawkes process model for the Ebola virus analysis of Section 3.2.