

A Metric Space for Point Process Excitations

Myrl G. Marmarelis

Greg Ver Steeg

Aram Galstyan

USC Information Sciences Institute

MYRLM@ISI.EDU

GREGV@ISI.EDU

GALSTYAN@ISI.EDU

Abstract

A multivariate Hawkes process enables self- and cross-excitations through a triggering matrix that behaves like an asymmetrical covariance structure, characterizing pairwise interactions between the event types. Full-rank estimation of all interactions is often infeasible in empirical settings. Models that specialize on a spatiotemporal application alleviate this obstacle by exploiting spatial locality, allowing the dyadic relationships between events to depend only on separation in time and relative distances in real Euclidean space. Here we generalize this framework to any multivariate Hawkes process, and harness it as a vessel for embedding arbitrary event types in a *hidden* metric space. Specifically, we propose a Hidden Hawkes Geometry (HHG) model to uncover the hidden geometry between event excitations in a multivariate point process. The low dimensionality of the embedding regularizes the structure of the inferred interactions. We develop a number of estimators and validate the model by conducting several experiments. In particular, we investigate regional infectivity dynamics of COVID-19 in an early South Korean record and recent Los Angeles confirmed cases. By additionally performing synthetic experiments on short records as well as explorations into options markets and the Ebola epidemic, we demonstrate that learning the embedding alongside a point process uncovers salient interactions in a broad range of applications.

1. Introduction

Infectious diseases (Gibson, Streftaris, & Thong, 2018), *news topics* (He, Rekatsinas, Foulds, Getoor, & Liu, 2015), *crime patterns* (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011), *neuronal spike trains* (Pillow, Shlens, Paninski, Sher, Litke, Chichilnisky, & Simoncelli, 2008), and *market trade-level activity* (Morariu-Patrichi & Pakkanen, 2018; Swishchuk & Huffman, 2020) naturally suit the form of a diachronic point process with a network of directed excitations. Understanding their intrinsic dynamics is of immense scientific and strategic value: a particular series of discrete options trades may inform an observer on the fluctuating dispositions of market agents; similarly, temporal news publication patterns may betray an ensuing shift in the public zeitgeist. The spread of a novel pathogen, notably the COVID-19 virus, through disjointed pockets of the globe hints to how it proliferates, and how that might be averted (Drakopoulos, Ozdaglar, & Tsitsiklis, 2017).

Simple Hawkes (1971) processes model excitations as a linear combination of responses to past events. Often in the multivariate setting, a drastic need for data-efficient techniques arises: practically, estimating the $(n \times n)$ possible excitations between each dyad (pair) of event types is often untenable without succinct and interpretable parametrization. How can one possibly disentangle the contributions of hundreds of options trades within each minute, in a myriad of different strike prices and expiration dates, to the future frequency of

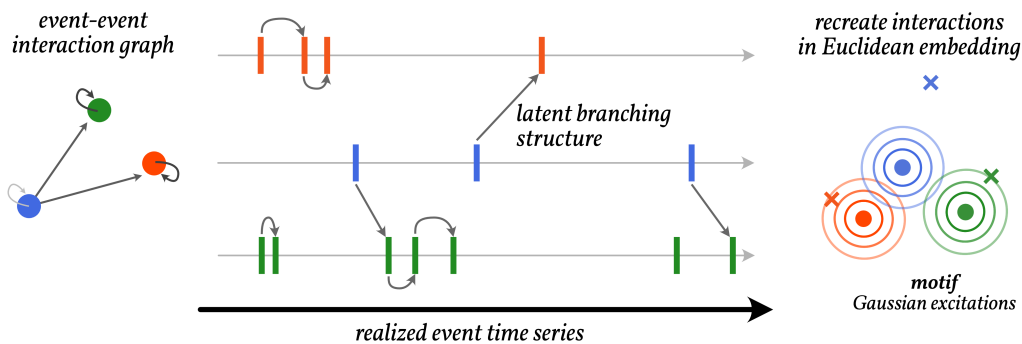


Figure 1: A stylized rendition. Instead of learning all possible interactions, we estimate a compact Euclidean embedding with fewer degrees of freedom and the additional benefit of interpretability.

a particular type of trade? Temporal processes like markets are also highly non-stationary, exacerbating the restriction to learn meaningful models on short windows. In these cases it is necessary to envision a reduced parametric form that is expressive yet cogent. The *manifold hypothesis* (Fefferman, Mitter, & Narayanan, 2016, for possible references) supposes that the vast majority of complex, high-dimensional phenomena reside in a state space of significantly lower dimensionality. The present endeavor is an effort to estimate embeddings of manifold-like structures from indirect measurements—in this case, marked events.

Spatiotemporal domains (Veen & Schoenberg, 2008; Mohler, 2014; Yuan, Li, Bertozzi, Brantingham, & Porter, 2019) exploit physical constraints on interaction locality. In essence, the direct influence one event bears on another is restricted by their separation in space, invariant to absolute positioning. We show that it is feasible to conjure a *latent* metric representation of the event types by harnessing a spatiotemporal approach outside the spatial context. We tailor the latent space toward the precise interactions between events.

Our first contribution is a model that captures multivariate point-process regularities in a compressed metric space, and which eliminates the need to estimate all the pairwise interactions between the event types. As our second contribution, we propose three estimation algorithms founded on the expectation-maximization (EM) technique in identifying Hawkes processes (Yuan et al., 2019; Zhou, Zha, & Song, 2013b; Halpin & Boeck, 2013). A latent structure introduced in the E-step to attribute the source of each event realization (Figure 1) permits tractable solutions in the M-step, which we exploit to optimize the embedding directly. Third, our findings on diverse social datasets demonstrate improvements upon the State of the Art (SotA) models, while also yielding qualitatively intuitive representations (Salehi, Trouleau, Grossglauser, & Thiran, 2019; Xu, Farajtabat, & Zha, 2016; Zhou, Zha, & Song, 2013a). We successfully learn the regional diffusion process for the 2014–2015 Ebola outbreak as well as the influences between options trades in different stocks. We also optimized an embedding for the South Korean early COVID-19 outbreak (Kim, 2020) and the winter stage of the pandemic in Los Angeles (Welsh, 2020).

One active line of work, paralleling ours, embeds each event based on some conceived heuristic like temporal proximity (Torricelli, Karsai, & Gauvin, 2020; Zhu & Xie, 2019; Zuo,

Liu, Lin, Guo, Hu, & Wu, 2018); another direction of inquiry entails the estimation of low-rank multivariate Hawkes processes (Nickel & Le, 2020; Lemonnier, Scaman, & Kalogeratos, 2017; Junuthula, Haghdan, Xu, & Devabhaktuni, 2019). We differ from both in learning an actual metric-space representation vis-à-vis the real Hawkes likelihood. To our knowledge, we are the first to propose a Euclidean embedding scheme driven entirely by an underlying marked Hawkes process.

2. Method

Consider a record of N event occurrences $(k_i, t_i) \in \mathbb{H}$, $i = 1, 2, \dots, N$, with n marked types $k_i \in \{1, 2, \dots, n\}$ at times $t_i \in [0, T)$. We assume that events with certain marks excite future events of either the same or another type. Multiple such interactions may be present, and we desire to identify those that are warranted by the available observations. A compact representation would, in effect, induce a shrinkage prior on the continuum of allowable interactions. Its form should align with our inductive bias.

We begin with preliminaries. The multivariate intensity function $\lambda(k_j, t_j)$, conditional on the events in $[0, t_j)$, dictates the instantaneous frequency of event (k_j, t_j) . Any interval $[t, t + dt)$ witnesses Poisson-distributed instances of event type k with rate $\lambda(k, t)dt$. We decompose this intensity (Bacry & Muzy, 2016) into self- and cross-excitations and an intrinsic background rate, not knowing a priori which event triggered which:

$$\lambda(k_j, t_j) := \sum_{i=1}^N h(k_j, k_i, t_j - t_i) + \mu(k_j). \tag{1}$$

A variety of approaches exists to estimate or infer the $(n \times n)$ separate response functions, $\tau \mapsto h(k_j, k_i, \tau)$, indexed by the *influencing* event type k_i and *influenced* type k_j . Short records and ambiguities in the combinatorial structure of potential causation hinder proper identification of actual interactions. In the presence of competing renditions for the $\mathcal{O}(n^2)$ coefficients, we elect the structure that follows from the manifold hypothesis. Concretely, we envision a Euclidean geometry that adequately captures the interactions between event types. We estimate distinct embeddings for receiving, $x_{k_j} \in X \subset \mathbb{R}^m$, versus influencing, $y_{k_i} \in Y \subset \mathbb{R}^m$, event types because otherwise we inadvertently constrain ourselves to symmetrical interactions. Multivariate Hawkes processes generally admit directed interactions. Accordingly, the finite collections of vectors X and Y constitute the embedding of a putative manifold wherein the multivariate component of the excitations is characterized by $\|y_{k_i} - x_{k_j}\|_2$ —the influence that some event type k_i exerts on k_j . From hereon, we shall express y_{k_i} as y_i and x_{k_j} as x_j to ease the notational burden.

Each response is the result of a dyadic interaction between an event (k_i, t_i) in the past and the potential occurrence of an event with type k_j at the present t_j . The bank of R basis functions combines Gaussian spatial proximities and exponential clocks with support forward in time:

$$g_r(x_j, y_i) := (2\pi\beta_r^2)^{-\frac{1}{m}} \exp\left\{-\frac{\|x - y\|^2}{2\beta_r^2}\right\}, \tag{2}$$

$$f_r(\tau) := \kappa_r e^{-\kappa_r \tau}. \tag{3}$$

Eqs. 2 & 3 form the spatial and temporal basis of $r = 1, 2, \dots, R$ kernels comprising the response function, tentatively produced below and expanded with full nuance in Eq. 8. Evidently, our construction enables multiscale responses in space (with spreads β_r) and time (with decays κ_r). Later we introduce the ingredients $\xi(k_i)$ and γ_r to relieve a constraint on the magnitude of each response.

$$h(k_j, k_i, \tau) = \mathbb{1}[\tau > 0] \sum_{r=1}^R g_r(x_j, y_i) f_r(\tau)$$

The proposed model family is framed as a strict superset of the parametric Hawkes process with exponential kernels: setting $R = 1$, as we did in most of our experiments, collapses back to the exponential temporal response function.

A generalized Poisson process yields a clean log likelihood function, written as follows:

$$\log L = \sum_{j=1}^N \log \lambda(k_j, t_j) - \sum_{k=1}^n \int_0^T \lambda(k, t) dt. \quad (4)$$

However, direct gradient-based optimization is prone to instability and a painstaking selection of hyperparameters. We turn to an E - M procedure with an augmented objective.

Suppose we happened upon the expected *branching structure* (Halpin & Boeck, 2013; Reinhart, 2018) of the realized point process. In other words, we introduced latent variables $[p_{ijr}] \in \mathbb{R}^{N \times N \times R}$ holding expectation estimates of $P_{ijr} \in \{0, 1\}$, indicating whether it was the event instance i that triggered instance j , and attributing responsibility to kernel basis r . Approximate knowledge of the untenable true line of causation endows us with the so-called complete-data log likelihood termed $\log L_c$ (Yuan et al., 2019; Veen & Schoenberg, 2008; Reinhart, 2018), an expectation of the joint log-probability density of the record and the latent variables $[P_{ijr}]$ in terms of their probabilities $[p_{ijr}]$.

$$\log L_c := \sum_{j=1}^N \left(\sum_{i=1}^N \sum_{r=1}^R p_{ijr} \log h_r(k_j, k_i, t_j - t_i) + p_{bj} \log \mu(k_j) \right) - \sum_{k=1}^n \int_0^T \lambda(k, t) dt \quad (5)$$

By abuse of notation let $h_r(\dots)$ denote the r th response kernel. Note that in the above form, what was previously a logarithm of summations (see Eq. 4 and 1) is replaced by a weighted sum of decoupled logarithms. The probability that the event instance was due to the background white Poisson process is p_{bj} ; $\forall j$, $\sum_{i,r} p_{ijr} + p_{bj} = 1$. Concretely, given a model $\lambda(k, t)$, our expectation for the particular branch of event cascades where event i triggered event j via kernel r is the ratio of that particular contribution to the overall intensity:

$$p_{ijr} := \frac{h_r(k_j, k_i, t_j - t_i)}{\lambda(k_j, t_j)}. \quad (6)$$

Lemma 1. *The complete-data log likelihood in Eq. 5 provides a lower bound for the point process log likelihood (Eq. 4): $\log L_c \leq \log L$.*

Proposition 1. *The lower bound of Lemma 1 is maintained when the latent variables' expectations $[p_{ijr}]$ take on any value as long as $\forall j$, $\sum_{i,r} p_{ijr} + p_{bj} = 1$.*

Our proof of Proposition 1 relies on Lemma 1 in the supplementary material. The right-hand term in Eq. 5 simplifies vastly if one assumes that $\forall y_i \in Y, \forall r, \sum_{x_k \in X} g_r(x_j, y_i) = 1$, and that $\forall i \forall r, \int_{t_i}^T f_r(t - t_i) dt \approx 1$. The latter approximation is tenable for large enough T ; the former is not. Only through certain concessions may we gain confidence that the sum is roughly unit (equal to one). First note that by the Gaussian integral, $\forall y \int_{\mathbb{R}^m} g_r(x, y) dx = 1$. Veen & Schoenberg (2008) approximated this sum as a Gaussian integral and demonstrated the viability of such alongside later studies. This approximation holds as long as the spatial occurrence of events is distributed uniformly in \mathbb{R}^m . In our embedding scheme, they are not: events are clumped at discrete locations of their types, and the objective function must also be constrained so that it does not drift into regimes that violate the approximation. Thusly, we are left with a coerced normalization of Eq. 2;

$$\hat{g}_r(x, y) = \frac{g_r(x, y)}{g_r(y)}, \quad g_r(y) := \sum_{x' \in X} g_r(x', y). \quad (7)$$

We rely on the unaltered form in Eq. 2 for the derivation of analytical maximizers during optimization, as is well established in seismology and other spatiotemporal studies (Reinhart, 2018); nevertheless, we found empirically that the intervention in Eq. 7 stymies potential drift towards degeneracy. Granular control of the magnitudes is desirable, so the final touch is the introduction of one more kernel parameter, $\xi(k_i)$, to directly represent the total after summation. The spatial kernel thus handles only the distribution of influence across its receptors. The full response function is therefore Eq. 8. We eliminate redundancies by *post hoc* constraining the *exertion coefficient* $n^{-1} \sum_{l=1}^n \xi(l) = 1$ and scaling the *basis coefficients* γ_r appropriately.

$$h(k_j, k_i, \tau) := \mathbb{1}[\tau > 0] \sum_{r=1}^R \xi(k_i) \gamma_r \hat{g}_r(x_j, y_i) f_r(\tau) \quad (8)$$

2.1 Optimization

Furnished with the expected branching structure in Eq. 6 (the ‘‘Expectation’’ step), we perform projected gradient ascent by setting partial derivatives of the complete-data log-likelihood with respect to each kernel parameter to zero (the ‘‘Maximization’’ step). Eventually, the potential trigger routes p_{ij} are aggregated in certain ways to form coefficient estimates. Omitting the domains of summation over i and j, i as implicitly $\{1, 2, \dots, N\}$ and $\{1, 2, \dots, N\} \times \{1, 2, \dots, N\}$ respectively, the solutions unfold as the following:

$$\kappa_r(\alpha_\delta, \beta_\delta) = \frac{\sum_{j,i} p_{ijr} + \alpha_\delta - 1}{\sum_{i,j} p_{ijr} (t_{k_j} - t_{k_i}) + \beta_\delta}, \quad (9)$$

$$\beta_r^2 = \frac{\sum_{i,j} p_{ijr} \|x_{k_j} - y_{k_i}\|^2}{m \sum_{j,i} p_{ijr}}, \quad (10)$$

$$\gamma_r = \frac{\sum_{j,i} p_{ijr}}{\sum_i \xi(k_i)}, \quad (11)$$

$$\xi(l) = \frac{\sum_{r=1}^R \sum_{j,i} \mathbb{1}[k_i = l] p_{ijr}}{\sum_{r=1}^R \gamma_r \sum_i \mathbb{1}[k_i = l]}, \quad (12)$$

$$\mu(k) = T^{-1} \sum_j \mathbb{1}[k_j = k] p_{bj}. \quad (13)$$

At times, it is necessary to preserve focus on the acceptable time horizons for a particular domain. A $\text{Gamma}(\alpha_\delta, \beta_\delta)$ prior on the decay rate κ_r admits the maximization a posteriori

in Eq. 9, which trivially becomes uninformative at the assignment $\kappa_r(1, 0)$ that we chose in our upcoming experiments. We included those extra parameters simply in case it becomes desirable to bias the model’s time scale in the future. Empirically, we found that one is typically interested in the half-life ($\log 2/\kappa_r$), the prior of which is the reciprocal of the aforementioned gamma distribution and characterized by the aptly named inverse-gamma distribution with expectation $\frac{\beta_\delta}{\alpha_\delta - 1}$. Preserving the mean while increasing both parameters strengthens the prior.

The influences $\Phi = [\varphi(k, l)]$ consist of the kernels with time integrated out, i.e.

$$\varphi(k, l) := \int_0^\infty h(k, l, \tau) d\tau = \sum_{r=1}^R \xi(l) \gamma_r \hat{g}_r(x_k, y_l).$$

There is evidence that this quantity encodes the causal network structure (Achab, Bacry, Gaiffas, Mastromatteo, & Muzy, 2017; Etesami, Kiyavash, Zhang, & Singhal, 2016). Pursuant to the above maximization step, one may alternatively estimate all of these n^2 degrees of freedom: that is what we will later term our baseline.

2.2 Embedding the Dyadic Relationships

We present two candidate approaches for estimating optimal Euclidean embeddings of the Hidden Hawkes Geometry (hence HHG) in a multivariate point process. One is based on the gradients of $\log L_c$, and the other on a diffusion-maps heuristic. Finally, we present a full-rank baseline estimator derived from the same EM algorithm.

2.2.1 MAXIMUM LIKELIHOOD

We learn embeddings directly via the EM objective function, $\log L_c$, rather than some heuristic. Our approach updates both the reception and influence embedding, concurrently with the rest of the parameters, during the M-phase. The influence points maximize their EM objective at tractable solutions to a set of decoupled equations.

Observe the partial gradient with respect to an influence vector y , after expanding the response functions:

$$\begin{aligned} \frac{\partial \log L_c}{\partial y_l} = \sum_{r=1}^R \left[- \sum_{j,i} \mathbb{1}[k_i = l] p_{ijr} \left(\frac{y_l - x_{k_j}}{\beta_r^2} \right) \right. \\ \left. + \xi(l) \gamma_r \sum_i \sum_{k=1}^n \mathbb{1}[k_i = l] (2\pi\beta_r^2)^{-m/2} \left(\frac{y_l - x_k}{\beta_r^2} \right) \exp \left\{ - \frac{\|y_l - x_k\|^2}{2\beta_r^2} \right\} \right] \quad (14) \end{aligned}$$

This expression is difficult to solve analytically. Recall, however, our prior simplifying assumption that $\forall y, r \sum_{x \in X} g_r(x, y) = 1$, also enforced a posteriori by means of Eq. 7. See the discussion above that surrounds this equation. The latter portion of Eq. 14 contains the form $\sum_x (x - y) g_r(x, y)$, equivalent to taking a quantized “expectation” of a Gaussian variable subtracted by its own mean (see Eq. 2). Hence, as long as the Gaussian sum is assumed to be approximately unit, then the entire second part of Eq. 14 vanishes due to the contribution of the pieces involving x_k . We garner the following—rather intuitive—formula for globally

optimal influence points, within the microcosm of the current E-phase:

$$y_l = \frac{\sum_{r=1}^R \sum_{j,i} \mathbb{1}[k_i = l] p_{ijr} x_{k_j}}{\sum_{r=1}^R \sum_{j,i} \mathbb{1}[k_i = l] p_{ijr}}. \tag{15}$$

Evidently each influence point $y_i \in Y$ is attracted to the reception points $\{x_j \in X\}$ that it appears to excite.

First-order optimizer (HHG-A). All parameters are updated simultaneously in each fixed-point iteration. Solving the optimality conditions leads to a decoupled system of equations comprising functionally distinct blocks of parameters, like reception points, influence points, and kernel decay rates. Unfortunately there is no analogous solution for the reception points that could manifest by a sensible approximation like that of the Gaussian integral, above. The simplest strategy is to submit to regular gradient ascent with learning rate ε : climbing the average log likelihood

$$x_k \leftarrow x_k + \varepsilon \cdot N^{-1} \frac{\partial \log L}{\partial x_k}. \tag{16}$$

We produce the gradient below, in implicit vector notation.

$$a_k := \frac{\partial \log L_c}{\partial x_k} = \sum_{r=1}^R \sum_i \left(\frac{x_k - y_{k_i}}{\beta_r^2} \right) \left[- \sum_j \mathbb{1}[k_j = k] p_{ijr} + \xi(k_i) \gamma_r (2\pi\beta_r^2)^{-m/2} \exp \left\{ - \frac{\|x_k - y_{k_i}\|^2}{2\beta_r^2} \right\} \right]. \tag{17}$$

To gain intuition on the selection of ε , we looked into *entropic impact* as a heuristic. The beautiful findings of McFadden, 1965 allowed us to reason about the contribution of shifting an embedding point to the differential entropy of a doubly stochastic point process:

$$\frac{\partial^2 H(k, t)}{\partial t \partial x_k} = [\log \lambda(k, t) + 1] \sum_i \frac{\partial h(k, k_i, t - t_i)}{\partial x_k},$$

which admitted a simple rule of thumb for adjusting the learning rate of Eq. 16 proportionally to n/N , with other factors pertaining to domain idiosyncrasies. Maintaining this rule ameliorated convergence in our synthetic experiments of §3.1.

Second-order optimizer (HHG-B). Further differentiating Eq. 17 with respect to every reception point leads to a block-diagonal Hessian matrix, \mathbf{B} , decoupled into $(m \times m)$ coordinate blocks indexed by event type, B_k . Assuming each of these is negative definite—more on that below—then it is numerically trivial to invert them. A naïve Newton step could have trouble converging, so we employed Levenberg-Marquardt regularization to introduce a sort of learning rate, ε_1 (Nesterov & Polyak, 2006). Newton-Raphson optimization may be viewed as maximizing a downwards-facing (convex) local quadratic approximation of the objective function. From that perspective, ε_1 plays the role of inserting a parabola centered at the current x_k and strengthening local convexity. We additionally introduce a regularization parameter ε_2 that constrains how far the embedding points may escape the origin. It may be useful in quelling the aimless drift of unconnected event types. Following the previous analogy,

ε_2 superimposes a parabola at the embedding origin, a Gaussian well in the log-likelihood space. High ε_1 grows jump sizes to mimic a learning rate, whereas ε_2 shrinks the jumps as a regularizer.

$$\tilde{a}_k := a_k - 2(N\varepsilon_2)x_k, \quad \tilde{B}_k := B_k - 2N(\varepsilon_1^{-1} + \varepsilon_2)I, \quad x_k \leftarrow x_k + \tilde{B}_k^{-1}\tilde{a}_k. \quad (18)$$

In practice, we chose to execute four such optimization steps during each M-phase to approximately converge to the current global optimum. Recall that a sole global optimum exists for each block of parameters per M-phase; however, we are point-wise maximizing an upper bound to the true likelihood. Each EM iteration concludes at a solution that is not guaranteed to be optimal, much less global. Yet it does converge eventually (Hunter & Lange, 2004).

Convexity. A multivariate function with negative-definite Hessian throughout the domain has one extremum, the global maximum, and is thereby convex. The reception-point Hessians B_k are not negative definite per se. They may, however, be deconstructed into the difference of a diagonal component and an outer product of another matrix with itself: $B_k = \text{diag}(c_k) - D_k^T D_k$. Clearly, if the elements of c_k were all negative, then the whole B_k would be negative definite. The vector c_k is

$$k \mapsto \sum_{r=1}^R \beta_r^{-2} \sum_i \left[\xi(k_i) \gamma_r (2\pi\beta_r^2)^{-m/2} \exp\left\{-\frac{\|x_k - y_{k_i}\|^2}{2\beta_r^2}\right\} - \sum_j \mathbb{1}[k_j = k] p_{ijr} \right],$$

where the positive component inside the sum is bounded above by unit, and the negative component is not analogously bounded below. To avoid undesirable cases, we apply a zero ceiling to each element of c_k before employing the additional regularizers via $(\varepsilon_1, \varepsilon_2)$. Convexity is therefore ensured. Oftentimes, it suffices to only set one of the two—particularly ε_1 —to a nonzero value. Whenever only ε_2 is specified, it is implied that $\varepsilon_1 \rightarrow \infty$.

2.2.2 DIFFUSION-MAPS HEURISTIC (HHG-DM)

Could we posit a diffusion process across event types, and estimate coordinates that recreate diffusion distances on an approximate Riemannian manifold? Random-walk manifold embeddings are helpful in deep representations (Kalatzis, Eklund, Arvanitidis, & Hauberg, 2020; Li, Lindenbaum, Cheng, & Cloninger, 2019; Rey, Menkovski, & Portegies, 2019). Construed as graph affinities, the influences Φ guide a Markovian random walk of which diffusion maps (Soize & Ghanem, 2016; Lian, Talmon, Zaveri, Carin, & Coifman, 2015; Coifman & Lafon, 2006) may be approximated via spectral decomposition. We found that asymmetrical diffusion-maps embeddings in the style of Pham & Chen, 2018 serve as an adequate initial condition for (X, Y) but are not always conducive to stable learning in conjunction with the rest of our iterative procedure. We term the model learned entirely this way as HHG-DM.

We briefly review the technique’s application here; the curious reader is encouraged to peruse the theory presented in Coifman and Lafon’s seminal publication (2006). Casting the influence matrix as edge weights in a bipartite graph flowing between *influence (col.)* \leftrightarrow *reception (row)*, we examine the diffusion process upon it (Pham & Chen, 2018). We first normalize by density to our liking, per our selected value for the parameter $0 \leq \alpha \leq 1$

according to

$$A = \text{diag} \left(\left\langle \sum_{l=1}^n \varphi(l, k) \right\rangle_{k=1}^n \right)^{-\alpha} \cdot \Phi \cdot \text{diag} \left(\left\langle \sum_{l=1}^n \varphi(k, l) \right\rangle_{k=1}^n \right)^{-\alpha}. \quad (19)$$

Consider the row-stochastic version of A , named B_R . Its singular values multiplied by the left (orthonormal) eigenvectors thereof supply manifold embedding coordinates for the reception points, weighted by significance according to the singular values. Likewise, B_I may be constructed as the row-stochastic transformation of A^T from the same Eq. 19, of which the resultant coordinates grant us the influence points. In each set of coordinates, we preserve only those corresponding to the highest m singular values—except for the largest, which is constant by definition.

2.2.3 FULL RANK BASELINE (FRB)

Did the reduction in degrees of freedom lend its hand to a more generalizable model of the point process? In order to motivate the reason for having an embedding at all—besides the gains in interpretability—we pitted the techniques HHG-A/B and HHG-DM against the following:

$$h(k, l, \tau) = \varphi(k, l) \mathbb{1}[\tau > 0] \sum_{r=1}^R f_r(\tau), \quad (20)$$

having estimated the full-rank matrix entries $\varphi(\cdot, \cdot)$ directly (Zhou et al., 2013b).

2.3 Initialization

We surmised adequate initial conditions for the EM procedure with a fixed empirical protocol. The surmised influence matrix came by summing up correlations between event types.

$$\hat{\varphi}(k, l) = \sum_{i,j} \mathbb{1}[k_j = k, k_i = l] \cdot \hat{\kappa} e^{-\hat{\kappa} \cdot (t_j - t_i)} \quad (21)$$

Notice that it remains unscaled. Initial “coefficient” $\hat{\kappa}$ was computed as the naive reciprocal inter-arrival time between event types $\hat{t} = n(N - 1)^{-1} \sum_{i=1}^{N-1} (t_{i+1} - t_i)$. We justify this construction on basis of Eq. 9, which forms a weighted average over said arrival times to garner an optimal estimate for κ_r^{-1} . We feed the result of Eq. 21 into the diffusion-maps algorithm in order to obtain our initial embeddings $(\hat{x}, \hat{y}) \in \hat{X} \times \hat{Y}$. $\forall r \hat{\beta}_r^2$ is initialized at the mean dyadic squared distance in the embedding; $\hat{\kappa}_r = (r\hat{t})^{-1}$, in which variety is injected to nudge the kernels apart; $\hat{\gamma}_r = R^{-1}$ and finally $\forall x, \hat{\mu}(x) = T^{-1}n^{-1}N$.

3. Results

First we explore simulated estimations to elucidate the comparative behavior of all four techniques: HHG-A, HHG-B, HHG-DM, and FRB. Then we present a number of findings and challenges regarding the characterization, namely, of a recent Ebola epidemic, the COVID-19 epidemic, and an illustrative portion of the options market. As the Euclidean norm, especially under a Gaussian kernel with exponential decay, suffers from the curse of dimensionality, all of the following experiments set $m \in \{2, 3\}$. It remains the topic of future study to incorporate parsimonious yet more robust metrics.

3.1 Synthetic Experiments

We intended to stress-test the learning algorithm under small record sizes and numerous event types; we thereby contrived a few scenarios with known ground-truth parameters sampled randomly. The underlying models had a single kernel $R = 1$ in the response function and conformed to the formulations from §2.2. Each sampled model was simulated with the thinning algorithm (see e.g. Mei & Eisner, 2017 and their supplementary material) in order to generate a time-series record of specified length N . Reception and influence points were realized uniformly from a unit square ($m = 2$). Spatial bandwidths β^2 were granted a gamma distribution with shape $\alpha = 1/\sqrt{n}$ and unit scale. Decay rates κ were standard log-normal as were backgrounds $\mu(k)$, though scaled by $1/n$.

Stability (Bacry & Muzy, 2016) was ensured by setting $\gamma = 1/\sqrt{n}$, constraining the Frobenius norm $\|\Phi\|_F = 1$ that upper-bounds the L^2 -induced norm, which itself upper-bounds the spectral radius of the influences $\rho(\Phi)$, the real criterion.

Settings			Outcomes			
Model (see §2.2)	(n) Types	(N) Events	Test	Train	Div.	Impr.
HHG-B ($\varepsilon_2 = 10^{-1}$, Eq. 18)	15	300	-4.45 ± 2.11	-4.09 ± 1.80	0.15	-0.023
HHG-A ($\varepsilon = 0.75$, Eq. 16)			-4.82 ± 2.80	-4.36 ± 2.59	0.14	*0.028
HHG-DM			-5.21 ± 4.33	-4.52 ± 2.96	0.12	***0.040
FRB			-5.34 ± 3.88	-4.42 ± 2.94	0.20	—
HHG-B ($\varepsilon_2 = 10^0$)		900	-5.03 ± 3.35	-4.79 ± 3.88	0.11	0.023
HHG-A ($\varepsilon = 0.25$)			-5.79 ± 3.98	-5.54 ± 3.78	0.12	-0.012
HHG-DM			-3.94 ± 2.39	-3.78 ± 2.31	0.08	**0.044
FRB			-6.13 ± 3.61	-5.71 ± 3.21	0.14	—
HHG-B ($\varepsilon_2 = 10^0$)	30		-5.51 ± 4.15	-5.47 ± 4.16	0.11	0.009
HHG-A ($\varepsilon = 0.5$)			-4.45 ± 3.96	-4.14 ± 3.48	0.09	0.011
HHG-DM			-7.62 ± 6.59	-7.04 ± 5.92	0.11	*0.020
FRB			-6.87 ± 4.52	-6.46 ± 4.17	0.30	—
—						
HHG-B ($m - 1$)			-7.93 ± 4.29	-8.21 ± 4.78	0.21	-0.010
HHG-B ($m + 1$)			-6.61 ± 4.03	-7.04 ± 4.29	0.12	-0.009
HHG-B ($m + 2$)			-6.17 ± 6.01	-6.60 ± 6.28	0.11	0.035
HHG-B ($m + 3$)			-7.63 ± 4.95	-7.92 ± 5.47	0.12	0.025

Table 1: Agglomerate metrics collected in twenty simulated trials per synthetic experiment. Outcome columns show, respectively, mean and standard deviation of test and train log likelihoods (cols. “Test” & “Train”), divergence from ground-truth branching structure (col. “Div.”), and mean improvement in correlation of the model’s embeddings from GloVe’s, each with respect to the ground-truth space; t-test significance markers are provided as well (col. “Impr.”). ***: $P \leq 0.01$, **: $P \leq 0.05$, *: $P \leq 0.1$. The correlation improvements all passed an Anderson-Darling test (Razali & Wah, 2011) for normality with confidence ≥ 0.95 .

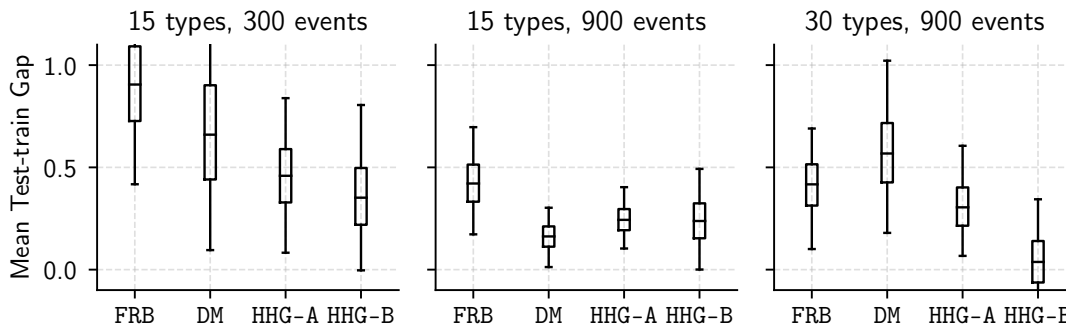


Figure 2: Decreases from in-sample (train) to out-of-sample (test) log likelihoods are indicative of overfitting. Bootstrapped means across the twenty trials reveal quantiles of 95% confidence. Comparative behavior varies by (n, N) configurations.

In line with Goodhart’s Law (1981), different facets of the model apparatus were scrutinized. First, we sought to ascertain whether the baseline tends to reach high in-sample likelihoods yet abysmal out-of-sample likelihoods, including extreme outliers. We bootstrapped the mean difference between the train and test log-likelihoods, arriving at confidence intervals for the test-train gap in Figure 2. This metric is a common indicator of overfitting. Its applicability is more tenuous in the empirical records scrutinized in §3.2, where the process is not guaranteed to be stationary.

There are further questions one could ask than the descriptive means and standard deviations of train and test log L in Table 1. Did our models recover the chain of causation between excited events? We opened up the empirical $[p_{ijr}]$ estimates and computed their Hellinger distance (Campbell & Li, 2019) from those stipulated by the ground truth, as if they were categorical distributions. For each “to be caused” event j , the quantities $(i, r) \mapsto p_{ijr}$ form empirical histograms that are poorly suited numerically for the more conventional KL-divergence measure, unlike the Hellinger distance.

Was the matrix of asymmetric influences $\varphi(k, l)$ recovered correctly? We visualized the squared errors between the ground-truth influences and those of the final estimate in Figure 3.

Digging deeper, it is often pertinent to examine residual distributions. Our point-process model considers some events as (probably) excited by the past and others as purely white-noise background occurrences. These background events innovate the process by possibly triggering new dynamics. Should our model fit the record, then it may be called upon to sample the likely background events from the rest. The interarrival times of that subset ought to follow an exponential distribution with rate parameters $\mu(k)$. By the Poisson superposition principle (Ogata, 1981), we may inspect the distribution of all background events against $\text{Exp}(\mu := \sum_k \mu(k))$. Those quantiles are illustrated in Figure 4.

Finally, as the intrinsic dimensionality m is unknown in practice, we reevaluated HHG-B under mismatched dimensionalities against the static ($m = 2$) ground truth. These results were included for the (30, 900)–case on the bottom of Table 1. $(\varepsilon_1, \varepsilon_2)$ remained unchanged.

Comparison to GloVe. In addition to excitatory point processes, our technique shares commonalities with a dual realm: that of vector embeddings for words and other sequential

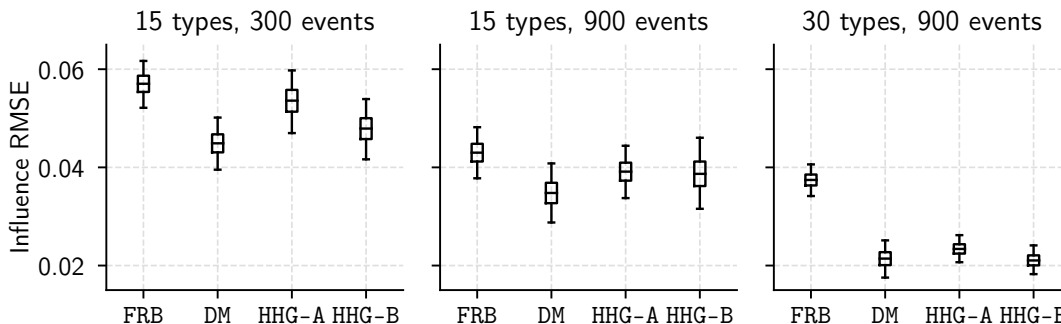


Figure 3: Root Mean Square Error (RMSE) of the recovered $(n \times n)$ influence (triggering) matrix, $[\varphi(k, l)]_{k, l}$. Bootstrapped means across the twenty trials reveal quantiles of 95% confidence. Comparative behavior varies by (n, N) configurations.

entities. We sought to draw comparisons with the latest treatments in that field. Our plan consisted of feeding the ordered sequence of event-type occurrences into a typical application of the GloVe scheme (Pennington, Socher, & Manning, 2014), in an attempt to recover a set of vectors that served as both influence and reception points. We designed a forward-looking window on the three next events to produce an asymmetric co-occurrence matrix, resulting in three embedding dimensions.

The final column of Table 1 displays a systematic evaluation against GloVe’s embeddings with reference to the ground-truth geometry. Concretely, all pairwise distances in each setting (that of our learned model and the newfound GloVe embeddings) were correlated to those of the ground truth by Kendall’s rank-based nonparametric statistic (Newson, 2002). The gap between GloVe’s estimated correlation and the model’s under scrutiny, each in $[-1, 1]$ and where a positive difference means the model correlated more with the ground truth, was collected in each trial, and means along with a t-test significance (Student, 1908) were reported in the last column of Table 1. GloVe is nondeterministic, so we obtained the sample mean of ten embedding correlations per trial.

3.2 Real-World Experiments

Model Variety (see §2.2)					State of the Art, reported in Salehi et al., 2019		
HHG-B	-A	-DM	FRB	GEO	VI-SG (2019)	SGLP (2016)	ADM4 (2013a)
-0.46	-1.18	-1.80	-0.17	-1.15	-2.06	-3.03	-4.61

Table 2: Test $\log L$ at the best train $\log L$ under various conditions from the Ebola dataset. Each proposed model shows the best of 500 epochs; the others were trained until convergence. Log likelihoods are averaged over the record size. GEO refers to interactions modeled over actual geographic coordinates. ($m = 3$, $\varepsilon_1 = 10^{-2}$, $\varepsilon_2 = 10^{-1}$, $\varepsilon = 2$)

Epidemics. Consider an infectious disease within a social apparatus, behaving like a diffusive point process. In 2019, a finely regularized variational approach to learning multivariate

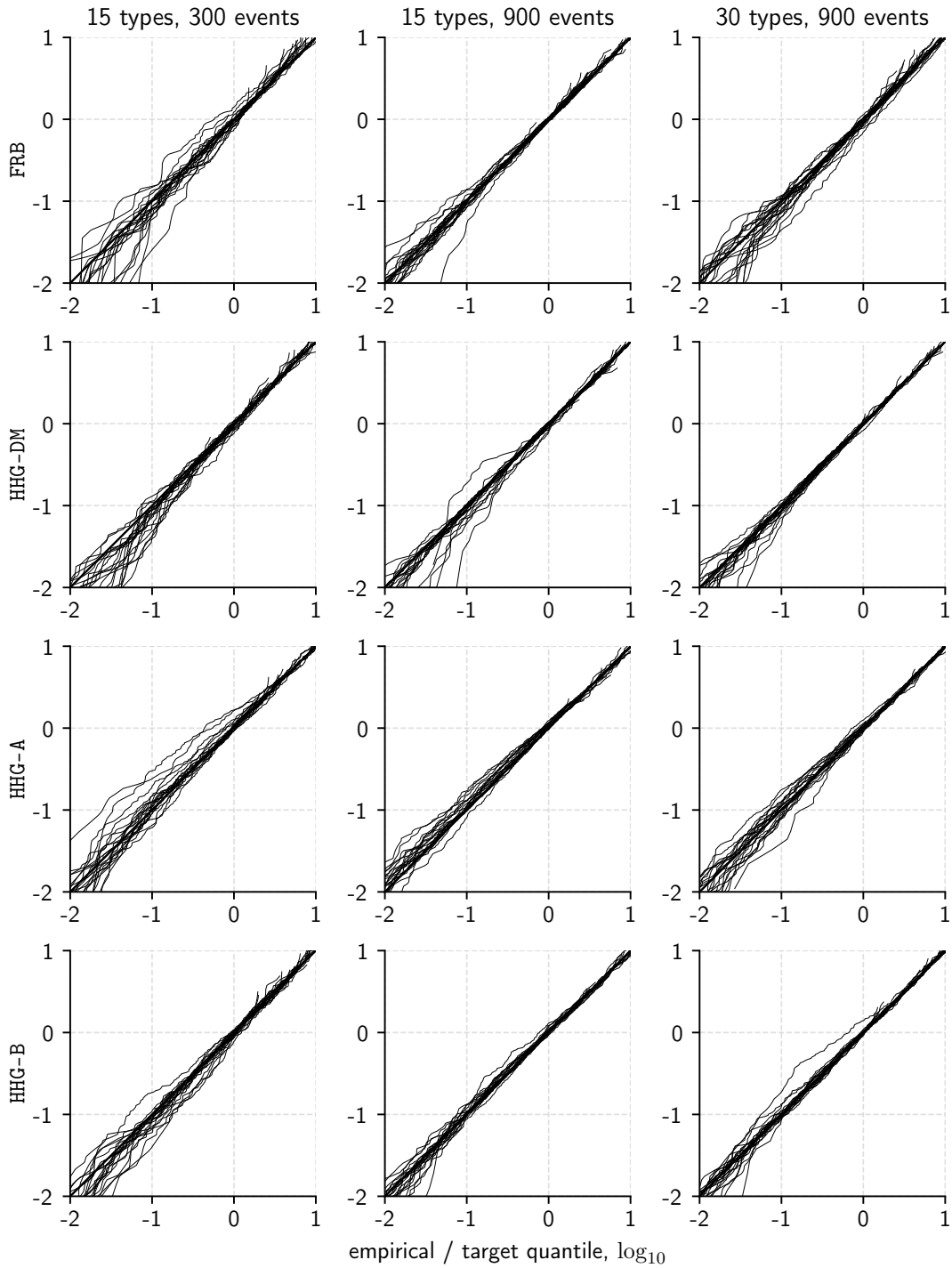
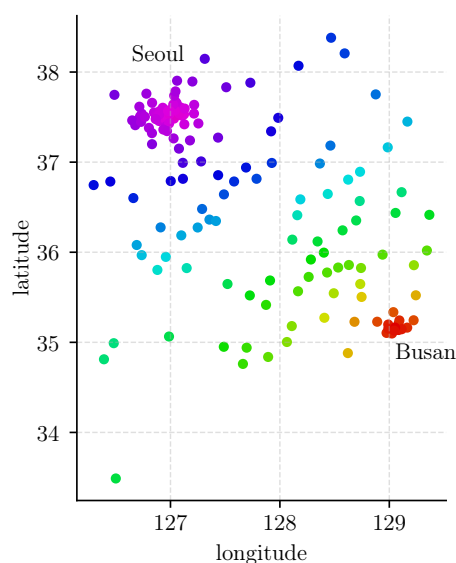


Figure 4: Synthetic results. Aligning quantiles of empirical and theoretical distributions of background-event interarrival times, deduced by the respective modalities. Deviations from the diagonal suggest a model mismatch, particularly in filtering out the background events.

Hawkes processes from relatively short records (Salehi et al., 2019) was demonstrated on a dataset of symptom incidences during the ~2014–2015 Ebola outbreak¹. We gave the record precisely the same treatment the authors did, and obtained significantly higher commensurate likelihoods than their best case. In turn, they outperformed the cutting-edge approaches MLE-SGLP (Xu et al., 2016) and AHHG-DM4 (Zhou et al., 2013a) that regularize towards sparsity.

We also trained a model with an embedding fixed to the geographic coordinates of the 54 West African districts present in the dataset, requiring it only to learn the appropriate response function. Assessing the value of the real spatial component of this process, we found that only HHG-B and FRB successfully outperformed the spatiotemporal model. See Table 2.



Model	Parameters	Test
HHG-B	$(\varepsilon_1 = 10^2, m = 2)$	-3.05
HHG-A	$(\varepsilon = 10, m = 2)$	-3.03
HHG-DM	$(\alpha = 1, m = 2)$	-3.28
FRB		-4.00
GEO		-3.11

Figure 5: **Left.** Geographic locations of South Korean regions (mostly cities). Colors were interpolated by hue on the basis of physical proximity to metropolitan Seoul (violet) versus Busan (red), the two major urban centers. **Right.** Out-of-sample average log-likelihoods attained by each model.

COVID-19. The novel coronavirus brought the world to its knees almost a year ago, as of this writing. The human suffering and socioeconomic disruptions have persisted since. Identifying harbingers for COVID-19’s spatiotemporal transmission is of paramount importance for proactive policymaking. The most readily available data consist of confirmed cases tallied at the end of each day; timestamps of higher fidelity are both impractical and of questionable value, since a positive test result indicates that virus transmission occurred some days ago, the precise number of which is ambiguous. Further, the inhomogeneous testing rate per community and over time contributes to the nonstationarity of the process. Inherent nonstationarity stems from changing social practices and other exogenous conditions. Reports

1. (Garske, Cori, Ariyaratnam, Blake, Dorigatti, Eckmanns, Fraser, Hinsley, Jombart, Mills, Nedjati-Gilani, Newton, Nouvellet, Perkins, Riley, Schumacher, Shah, Kerkhove, Dye, Ferguson, & Donnelly, 2017)

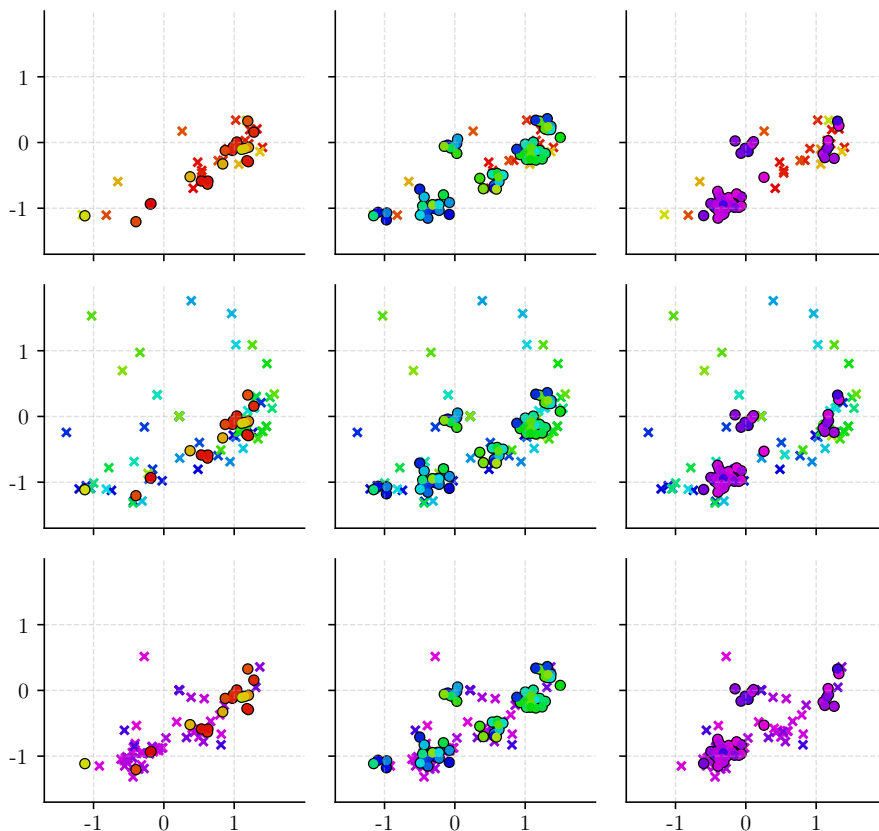


Figure 6: Embeddings of the highest-scoring model for South Korea: our novel HHG-A. Each location is endowed with a pair of an ex and a dot, corresponding to receiving and influencing points respectively. Panels were faceted by qualitative color groups interpreted from Figure 5. Under $m = 2$, the raw hidden-space coordinates were plotted. We added normal noise with standard deviation of five units to each of the influence points (dots) to reveal the various colors stacked on top of each other.

of deaths attributed to COVID-19 are potentially even less reliable as proxy indicators for transmission, due primarily to inconsistent protocols. As before, we only modeled point-process excitations. Incorporating a self-limiting aspect—as in the compartmental models (e.g. Bertozzi, Franco, Mohler, Short, & Sledge, 2020)—remains the topic of future study.

South Korea. Early on in the pandemic, South Korea experienced a transient surge in cases that it swiftly suppressed by strict social controls. We believed we were likely to succeed in interpreting that record as a relatively stable Hawkes process, as opposed to most other countries that did not enforce the same measures. We found the meticulous dataset released by the Korean Centers for Disease Control & Prevention (Kim, 2020). They detail 3,385 incidences from the early outbreaks across the 155 regions of the country, each with at least one infection occurrence. Diligent testing appears to contribute to the embedding’s identifiability.

The test set consisted of the last 30 days in the record, containing 631 incidences. The table in Figure 5 displays the resultant score of each model. It is also worth noting that the full-rank baseline FRB registered a significantly higher in-sample likelihood than all the other models did, indicative of excess overfitting. See the full HHG-A embeddings in Figure 6. The optimal response function (with one kernel, $R = 1$) predicted that each contagious individual infected $\gamma = 0.714$ others on average, with a half life of 3.34 days.

Los Angeles is a dense metropolitan area with diverse demographics. We are interested in recovering the landscape of transmission dynamics. It usually differs from a uniform spatiotemporal diffusion process because the physical separation between cities is not commensurate to how much their residents come into contact. Distant cities connected by common routes between home and work may be nearby in a putative “interaction space.” In the absence of reliable network information on contact rates—for mobility data introduce their own set of biases²—it would be useful to infer this from the infection rates themselves.

Model	Test
HHG-B ($\varepsilon_1 = 10^{-2}$, $m = 3$)	0.694
HHG-A ($\varepsilon = 20$, $m = 2$)	0.655
HHG-DM ($\alpha = 1$, $m = 3$)	0.701
FRB	0.659
GEO	0.684

Table 3: Out-of-sample average log-likelihood for **Los Angeles** attained by each model.

We found a dataset of confirmed cases in Los Angeles County curated by the Los Angeles Times (Welsh, 2020). The sheer volume of infections per city is reflected in daily counts. To curb some of the nonstationary elements we truncated the record to the latest 50,000 cases as of December 12, 2020. To discretize the time series into events, we interpolated the cumulative daily counts of each city into a continuous signal. Specifically, we extended the curves linearly after a logarithmic (because of exponential growth) transformation. We then located the exact timestamp of consecutive increments of a specific threshold size; in this case study, that was ten infections. The resultant dataset consists of 5,000 events beginning roughly on November 9. Each of those events belongs to one of the 64 most impacted cities of Los Angeles County, depicted on the map in Figure 7.

The test set consisted of the last five days in the record, comprising 1,605 events out of the 5,000. Trying $m = 2, 3$ and a handful of choices for the hyperparameter ε , we arrived at the results in Table 3. Our novel parametrization HHG-DM scored the highest, and was visualized in Figure 7 through colorings that reflect topography in the HHG-DM embedding. Its predicted average infection rate was 0.578 and its half life was 0.0934 days: a little troublesome in comparison to the South Korean results, and likely due to the instability of this process.

2. (Oliver, Lepri, Sterly, Lambiotte, Delataille, De Nadai, Letouzé, Salah, Benjamins, Cattuto, Colizza, de Cordes, Fraiberger, Koebe, Lehmann, Murillo, Pentland, Pham, Pivetta, Saramäki, Scarpino, Tizzoni, Verhulst, & Vinck, 2020)

POINT PROCESS METRIC SPACE

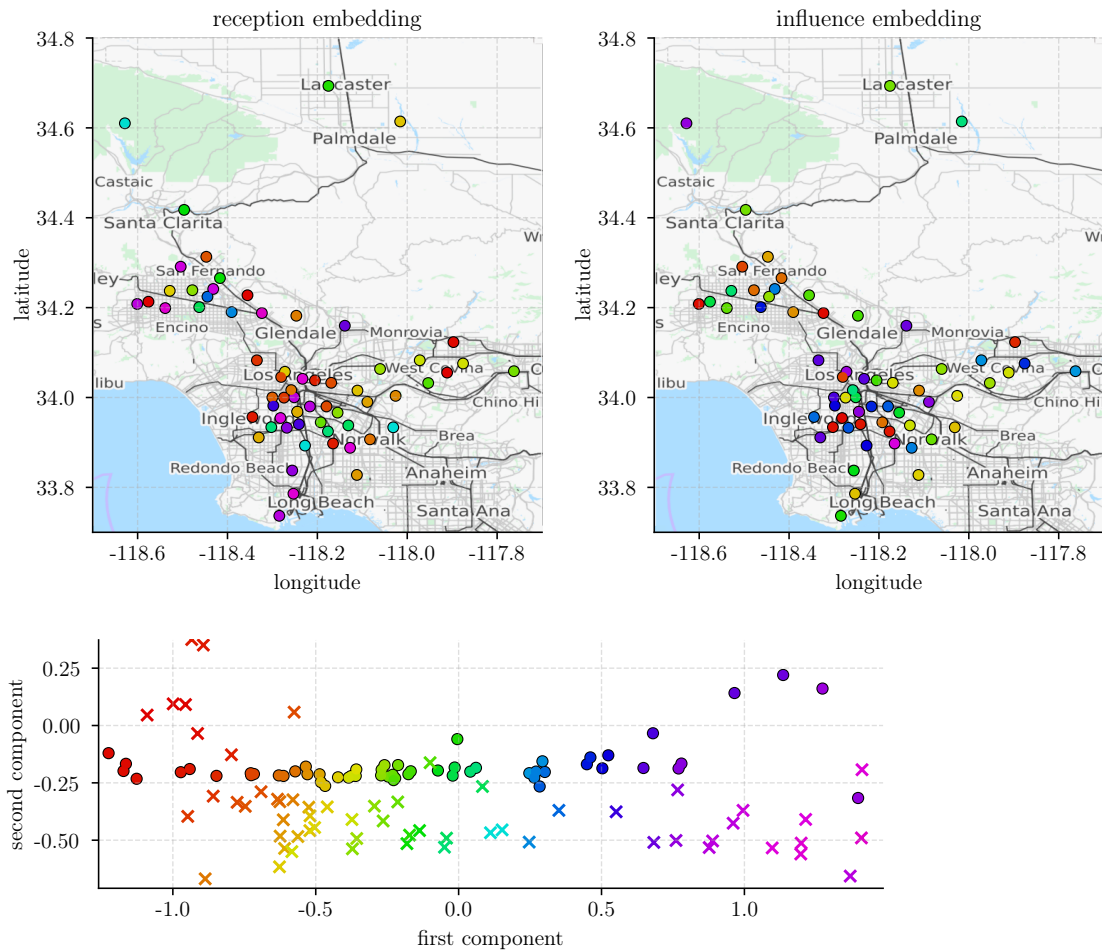


Figure 7: **Above** are the geographic locations of the 64 cities in Los Angeles county that were impacted with the most confirmed cases of COVID-19. Communities are colored by locations in the reception (left) and influence (right) embedding in the highest-scoring HHG-DM model: from violet to red along the principal axis in the embedding space. Image source: OpenStreetMap. **Below** are plotted the two principal components of the influence (dots) and reception (exes) embeddings accompanied by the same colorings as above.

Options market. The intertwined market activity of options with underlying stocks TSLA, AAPL, and IBM during the unremarkable consecutive trading days of Sept. 15 & 18, 2017 led to 24 distinct event types. Roughly 100,000 total trades per day were sampled on 120% and 80% fuzzy moneyness levels—as portrayed in Figure 10—at the expiration dates 01/18/2019 & 04/20/2018 for both puts and calls. The historical data was procured from AlgoSeek with the generosity of Prof. Roger G. Ghanem. The last 45 minutes of trading comprised each day’s test set. See Table 4.

An auxiliary accuracy metric in Table 4 is derived from categorical cross-entropy of the predicted event type at the time of an actual occurrence, rendered by the expression

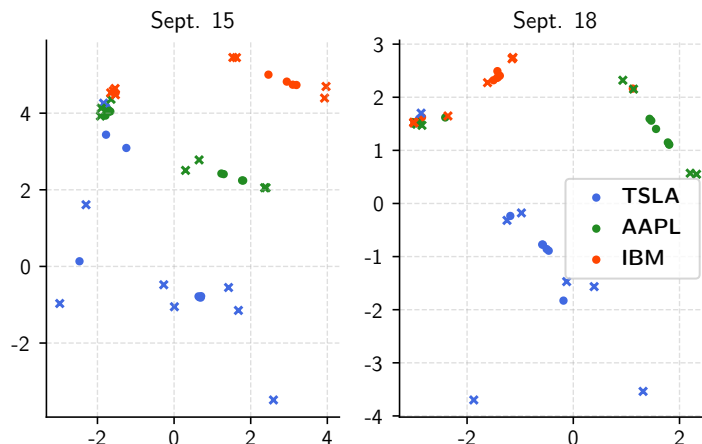


Figure 8: **Embeddings** on market data. Eight trade partitions per stock; TSLA is blue, AAPL green, and IBM orange. Exes receive; dots influence. Embedding scales are normalized to reflect a unit kernel bandwidth.

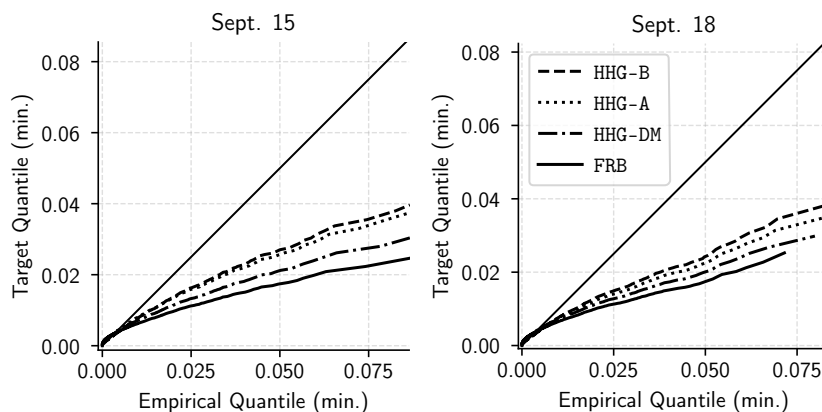


Figure 9: **Quantile-quantile plots** on inter-arrival times of all background market trades in the test partition, versus their supposed distributions from the sum of the model’s background rates.

$\exp\{\mathbb{E}_i[\log \lambda(k_i, t_i) - \log \sum_l \lambda(l, t_i)]\}$, in other words the geometric average of prediction accuracies. Between events, intensities decrease monotonically and with a homogeneous rate, so the ratio of intensities should stay roughly the same between occurrences—barring relatively outsized background rates. The “naive” score is computed from simply the mean empirical rate of each event type. We visualized the two-dimensional color-coded embeddings in Figure 8 and dsitributional fits for background events in Figure 9.

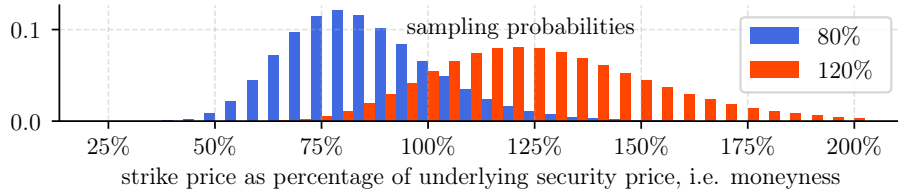


Figure 10: Trades at discrete strike prices were resampled according to quantized log-Gaussian profiles with reference to moneyness at any given point in time. Standard deviations, in logarithmic space, were half the separation between the two densities’ centers. They were kept “loose” for the sake of seamless translation even under abrupt fluctuations in the underlying stock price.

Dataset	Model	Train	Test	Half Life (min.)	Categorical Accuracy	
Sept. 15	HHG-B ($\varepsilon_1 = 10^1$)	3.13	2.60	6.50×10^{-4}	0.123 vs. naive	0.099
	HHG-A ($\varepsilon = 20$)	3.05	2.57	4.09×10^{-4}	0.119	0.099
	HHG-DM	2.85	2.34	2.26×10^{-4}	0.103	0.099
	FRB	2.54	2.11	1.15×10^{-3}	0.103	0.099
Sept. 18	HHG-B ($\varepsilon_1 = 10^{-1}$)	3.27	2.82	5.41×10^{-4}	0.135 vs. naive	0.108
	HHG-A ($\varepsilon = 1$)	3.13	2.72	5.88×10^{-4}	0.123	0.108
	HHG-DM	2.88	2.48	2.32×10^{-4}	0.103	0.108
	FRB	2.75	2.43	8.78×10^{-4}	0.117	0.108

Table 4: Market fits with associated half lives. In HHG-A, the best in-sample ε was picked out of a handful of candidates. A similar grid search was enacted on powers of 10 for ε_1 in HHG-B. Outcome of 1,000 epochs depicted. Categorical accuracy is a prediction score $\in [0, 1]$ for the next event type, at the point of occurrence of the next event.

4. Discussion

We expose some empirical details for estimating models and then analyze significant results.

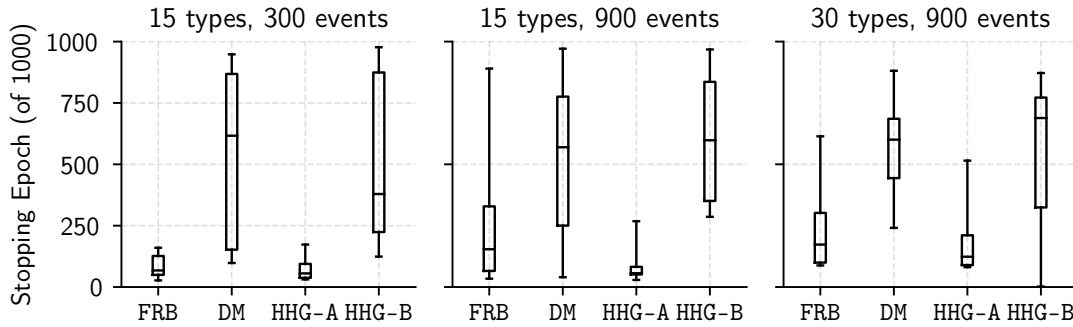


Figure 11: Indices of the stopping epochs for the synthetic experiments. Closer to the end (1,000 epochs) suggests convergence, hence stability of the algorithm.

4.1 Convergence

Of our three proposals for hidden-embedding point process estimators HHG-B, HHG-A, and HHG-DM, we concede that the iterator’s stability tends to decline in that order of enumeration.

Observe, in Figure 11, the epochs of the best models found. HHG-DM and HHG-B converge best in the synthetic experiments. Interestingly, that tendency did not transfer onto real datasets. That phenomenon demonstrates the value of theoretical guarantees for convergence. HHG-DM appears to learn in the same way that a random search with some upward drift does, contrary to our initial expectations. Every now and then the HHG-DM sampler would stumble upon a model with high in-sample likelihood that also garnered the best fit out of sample.

On the other hand, HHG-A and even more so HHG-B typically achieved their ideal fits at the end of their learning curves, even though HHG-A underwhelmed in the synthetic cases. In the Korean dataset, for instance, both HHG-A and HHG-B reached their maximum likelihoods at the last epoch. That was not the case for HHG-DM.

4.2 Time-Scale Drift

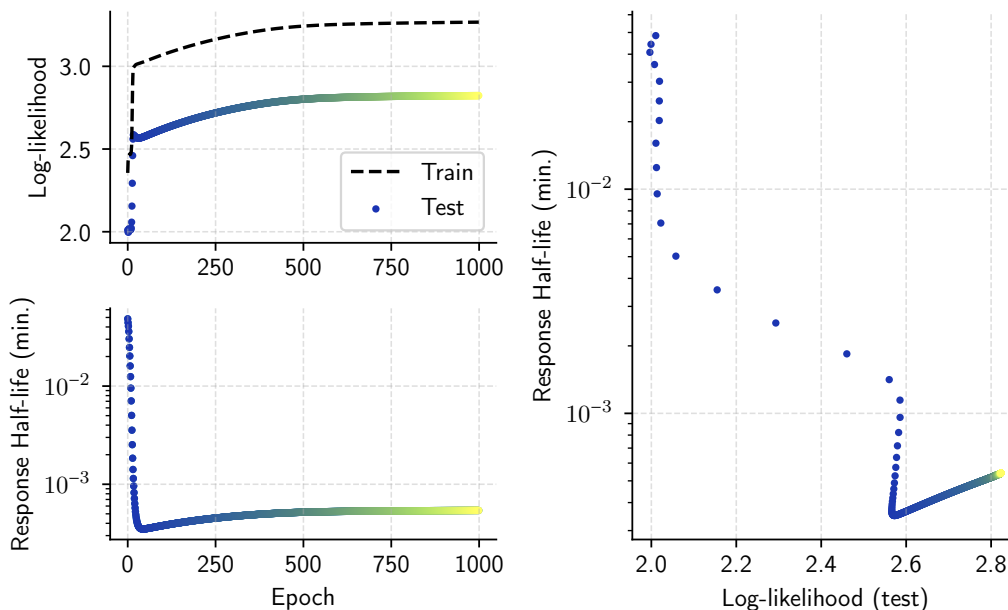


Figure 12: Illustrative scenario (Sept. 18 options) of time-scale drift in the model’s response function alongside a growing test-set likelihood.

As a model would converge to its optimal estimate, certain overarching parameters like the response time scale (homogeneous across event-type pairs) would drift. See e.g. Figure 12. It proved quite difficult to steer the optimizer away from its “destined” time scale.

Disentangling time scales. Parsing out longer behaviors out of high-frequency trades is severely difficult, and the gamma prior failed to facilitate it. Whenever we implemented a prior strong enough to influence the learning algorithm, it would overpower the course of training and degenerate the model. Should one desire to investigate longer-term patterns through

individual trades, it would be imperative to study higher-order interactions (Marmarelis & Berger, 2005) in the future.

4.3 Predictive Ability

We display the epoch with the best *training* score in all of our experiments. Most notable in Table 1 of synthetic results are the following: **HHG-A/B** yield a higher test $\log L$ than **FRB** in all three rows; **HHG-DM**'s results are a little more mixed. The ability to reveal chains of causality between excited events, as measured by the Hellinger divergence from the ground-truth branching structure, is better in **HHG** than in **FRB**. On average, the embeddings mostly mirror the ground-truth space more accurately than GloVe does, but **HHG-DM** here proved most consistent with significant improvement throughout.

We conjecture that the stiffness of a (normalized) spectral decomposition in **HHG-DM** provides some benefits and some drawbacks: it facilitates topological discovery—hence its correlation to the ground-truth layout—yet clashes with the real point-process objective. The direct optimizers of gradient-based **HHG-A** and curvature-informed **HHG-B** appear increasingly more resilient in general.

The excessive overfitting of the full-rank Hawkes process (**FRB**) is illuminated in Figure 2, as well as the more granular Figure 4.

As is the case for convergence in the synthetic experiments, **HHG-B** and **HHG-DM** stand out as most accurately recovering the influence matrix in Figure 3.

Focusing on the first two columns of the aforementioned figures, (15, 900) & (30, 900), we examine the relative performance of **HHG-B** and **FRB** as a function of increasing number of event types (the original process dimensionality). The full-rank baseline, **FRB**, exhibits inferior ability in recovering influences than **HHG-B** when the event types increase from 15 to 30, as witnessed in Figure 3. A similar effect occurs to a lesser extent in Figure 4. The main driving force behind the failure of **FRB** appears to manifest in the over-determination of possible interactions between event types.

Included additionally are the results of **model misspecification** for **HHG-B** in the (30, 900) case within the final rows of Table 1. Standing against the two-dimensional ground truth, models with $m = 1, 3, 4, 5$ showcase the general sensitivity to over-compression ($m = 1$) yet **FRB**-like performance following under-compression ($m = 3, 4, 5$). A surprising outcome is the divergence and GloVe-correlation improvement of the under-compressed models. All three of those models exhibit consistent lowered divergence metrics. The experiments with $m = 4, 5$ yielded astounding correlation improvements, albeit of high variance. Sometimes—although not always, as is evident in the Ebola investigation below—a few extra degrees of freedom garner the helpful kind of flexibility.

Benchmark on the Ebola dataset. Table 2's results depict superiority in our models to the SotA. It is puzzling that our simple full-rank baseline EM estimator **FRB** outperforms the more sophisticated formulations, when MLE-SGLP resembles a regularized version thereof. Perhaps its shrinkage priors do not suit this particular dataset, or all $\sim 17,000$ training events sufficiently fleshed out the possible interactions without need for regularization. Sparsity per se did not bear out as an appropriate constraint. Our hidden-embedding model **HHG-B** was the only such estimator that outperformed the spatiotemporal model built on geographic

“ground truths.” This finding suggests that geographic information is recovered in that embedding, at least.

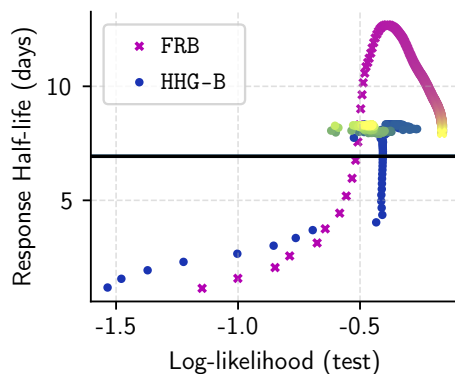


Figure 13: Per-epoch evolution of model estimates for the Ebola dataset. Black line indicates the decay-rate parameter set a priori for the other SotA models in Table 2.

It was difficult to transparently compare our methodologies against the SotA from Salehi et al., 2019. Part of the separation of outcomes could have been due to our framing that compels the estimator to learn a response decay rate on its own. The other cited models had those parameters selected via grid search, and remain static during the course of training. As witnessed in Figure 13, our proposed models roughly converged to similar decay rates as those chosen in Salehi et al., 2019, but the paths followed were rather variable.

The half lives of all models are largely similar, aligning with the time scales of the particular Ebola strain under consideration that had an incubation period of about 8–12 days (Kerkhove, Bento, Mills, Ferguson, & Donnelly, 2015).

Intrinsic dimensionality of the Ebola process. One would expect that a model like HHG-B would replicate FRB as the degrees of freedom equilibrate, i.e. $m \rightarrow n$. However, that was not found to be the case. It appears that our fixed-point EM algorithms vastly complicate learning, for little payoff, in comparison to FRB when m increases. We deemed it sufficient to demonstrate that the likelihood is not monotonic in embedding dimensionality m . To make our case, listed are likelihoods of the most stable HHG-B as a function of m : $\text{HHG-B}(2) = -1.7$, $\text{HHG-B}(3) = -0.46$, $\text{HHG-B}(4) = -0.80$. It therefore appears that the optimal dimensionality lies at $m = 3$, hinting at an intrinsic Euclidean geometry of the process.

Transmission network implicit in COVID-19 infection rates. The manifold structure of contagion networks has long been studied (Taylor, Klimm, Harrington, Kramár, Mischaikow, Porter, & Mucha, 2015), although estimation thereof remains elusive. Our point-process models identified embedding candidates based on the multivariate statistics of the infections. For one, we observe that the two urban hubs Seoul and Busan are not as far apart in the hidden influence space (Figure 6) as they are geographically. They still cluster in their respective neighborhoods—Busan to the right, Seoul to the left. The generally rural regions in between seem to spread across this topology. Regarding Los Angeles (Figure 7,) we first note that there is visible geographic locality preserved in the colorings. The reception

embedding exhibits contiguous streaks of red/orange, of yellow/green, and of blue/violet. The influence embedding has communities of green/yellow scattered about.

There are clear and acute differences between the embeddings and geography, albeit some geographic structure is recovered in the embedding. We expected divergences, as previously stated, since physical communities and even mobility networks do not always correspond to the underlying transmission process. All these factors may serve to inform an epidemiological model as proxies, in agglomeration, to the true contact rates between individuals. We provide an additional factor, derived from the measured infection rates themselves. It is distinct from the other datasets that estimate interaction between communities. We also demonstrated its predictive ability.

On our results pertaining to the pandemic, we decided against including the auxiliary goodness-of-fit tests on inter-arrival times and classification accuracies. Our reasoning is that the COVID-19 infection process is clearly not a Hawkes process; it is unstable, multiphasic, and multifaceted. Our models provide a coarse approximation of its dynamics. Other quantities more specific to a Hawkes process may be misleading. On the other hand, the likelihood is salient on any point process and embeddings should yield qualitative meaning. Our other point-process applications demonstrate the validity of the presented models.

Analysis of market embeddings. Events belonging to each stock ticker tend to attract influencing points of the same color in Figure 8, with deviations probably due to relative perceptions by the market. Efficient estimators are necessary in order to discern a lack of stationarity; in this case, transferring our Sept. 15 HHG-A model onto the Sept. 18 test set yielded an average $\log L$ improvement from 2.57 to 2.82. Vice versa, transferring from Sept. 18 to Sept. 15 led to a decrease from 2.72 to 2.45. Thus point-process behavior largely persists, but not entirely. Further, aggregate $\xi(l)$'s confirm our broad intuition that out-of-money trades move markets the most (Yang, Kutan, & Ryu, 2018).

Quantile plots in Figure 9 indicate that the embedding models—all with the HHG- suffix—outperformed the baseline in statistically filtering out white background events according to their estimated probabilities p_{bj} . A constant-intensity Poisson process would witness arrival times distributed exponentially (Bacry & Muzy, 2016); therefore, any deviation from that in a model's sampled background events suggests the presence of uncaptured interactions.

We observed that the categorical accuracy measured in Table 4, assessing classification ability towards the next event to come by ignoring temporal granularities, is highest for HHG-B in the two trading days studied. It beats the naive score by at least 2.4% each time.

4.4 Modality Prescription

Given the most consistently smooth learning curves under HHG-B during both synthetic experimentation and empirical investigations, we recommend the implementation of HHG-B in future endeavors. The final test-set likelihoods of HHG-B were always remarkably close to the top, if not the actually the highest. There was no empirical dataset scrutinized for which HHG-B failed significantly in comparison to HHG-A, HHG-B, or even FRB.

5. Related Work

The fundamental notion driving the doubly stochastic process, first attributed to Cox (1955), manifests in a variety of ways including the Latent Point Process Allocation (Lloyd, Gunter, Osborne, Roberts, & Nickson, 2016) model. In fact, note the meteoric diversity of Bayesian methods permeating serially dependent point processes (e.g. Apostolopoulou, Linderman, Miller, & Dubrawski, 2019; Detommaso, Hoitzing, Cui, & Alamir, 2019). Zhang et al. (2019) sample the branching structure in order to infer the Gaussian process (GP) that constitutes the influence function. GPs, usually accompanied by inducing points, sometimes directly modulate the intensity (Liu & Hauskrecht, 2019; Lloyd et al., 2016; Aglietti, Bonilla, Damoulas, & Cripps, 2019; Ding, Khan, Sato, & Sugiyama, 2018; Flaxman, Chirico, Pereira, & Loeffler, 2019). Linderman and Adams (Linderman & Adams, 2015) took an approach that estimated a matrix very similar to our $[\varphi_{ijr}]$, relying on discretized binning and variational approximation. Salehi et al. (2019) exploited a reparametrization trick akin to those in variational autoencoders in order to efficiently estimate the tensor of basis-function coefficients.

Recent progress has been made in factorizing interactions with a direct focus on scalability (Nickel & Le, 2020), improving on prior work in low-rank processes (Lemonnier et al., 2017). Block models on observed interaction pairs also exist (Junuthula et al., 2019). While these all achieve compact Hawkes processes, our methodology distinguishes itself by learning a Euclidean embedding with the semantics of a metric space, not a low-rank projection.

Deep nonparametric (Poisson or even intensity-free) point process models³ made massive strides in the past few years. However, not all domains permit such expressive characterization. Our baseline estimator resembles most closely the one described in the work of Zhou et al. (2013b), whereas the spatiotemporal aspect is inspired from the likes of Schoenberg et al. (see Veen & Schoenberg, 2008; Mohler, 2014; Yuan et al., 2019,). Variational substitutes in the EM algorithm have also been explored (Cho, Galstyan, Brantingham, & Tita, 2014). A concurrent study to ours by Zhu et al. (2020) parametrizes a heterogeneous kernel in real Euclidean space by deep neural networks.

6. Conclusions

We demonstrated the viability of estimating embeddings for events in an interpretable metric space by tying it to a self-exciting point process. The three proposed expectation-maximization algorithms extract parsimonious serial dependencies to different degrees of reliability, partly depending on record length and dimensionality. The optimizer of second order HHG-B converges most consistently to an accurate model. In comparison to learning the full triggering matrix (with FRB,) our embedding models achieve a higher likelihood out of sample. The estimated time scale for Ebola contagion settles close to its natural rate. Our framework paves the way for generalization, extension to more elaborate models, and consequent potential for actionable insights.

3. (notably Mei & Eisner, 2017; Lu, Wang, Shi, Yu, & Ye, 2019; Loaiza-Ganem, Perkins, Schroeder, Churchland, & Cunningham, 2019; Okawa, Iwata, Kurashima, Tanaka, Toda, & Ueda, 2019; Omi, 2019; Sharma, Ghosh, & Fiterau, 2019; Shang & Sun, 2019; He et al., 2015; Du, Dai, Trivedi, Upadhyay, Gomez-Rodriguez, & Song, 2016; Jia & Benson, 2019; Mehra, Deng, Ahmed, Chang, He, Durand, Brubaker, & Mori, 2019,)

Appendix A

Proof of Lemma 1

We seek to demonstrate that $\log L_c \leq \log L$ always. Recall the likelihood function and its complete-data relative:

$$\begin{aligned} \log L &= \sum_{j=1}^N \log \lambda(k_j, t_j) - \Lambda, \\ \log L_c &= \sum_{j=1}^N \left(\sum_{i=1}^N \sum_{r=1}^R p_{ijr} \log h_r(k_j, k_i, t_j - t_i) + p_{bj} \log \mu(k_j) \right) - \Lambda, \quad \text{where } p_{ijr} = \frac{h_r(k_j, k_i, t_j - t_i)}{\lambda(k_j, t_j)}. \end{aligned}$$

The intensity may be written as $\lambda(k, t) = \sum_{i=1}^N \sum_{r=1}^R h_r(k, k_i, t - t_i) + \mu(k)$. Compare each pairing of a receiving event's (indexed by j) additive contribution to the two objectives; on one hand, where $\log L$ contains $\log \lambda(k_j, t_j)$, $\log L_c$ reckons with

$$\begin{aligned} &\sum_{i=1}^N \sum_{r=1}^R p_{ijr} \log h_r(k_j, k_i, t_j - t_i) + p_{bj} \log \mu(k_j) \\ &= \frac{\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) \log h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \log \mu(k_j)}{\lambda(k_j, t_j)} \\ &= \frac{\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) \log h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \log \mu(k_j)}{\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) + \mu(k_j)}. \end{aligned}$$

To minimize distractions, we focus on the structure of what is found inside the logarithms and what is found on the outside. Accordingly, the above expression takes the abstract form $\frac{\sum_q w_q \log x_q}{\sum_q w_q}$ in terms of positive (for they serve as arguments to a logarithm) weights w_q and contents x_q , over arbitrary indices q . The concave logarithm admits the following relation, by Jensen's inequality:

$$\frac{\sum_q w_q \log x_q}{\sum_q w_q} \leq \log \frac{\sum_q w_q x_q}{\sum_q w_q}.$$

Since $w_q = x_q$ corresponding to expansion terms in the intensity function,

$$\log \frac{\sum_q w_q x_q}{\sum_q w_q} = \log \left[\sum_{i=1}^N \sum_{r=1}^R h_r^2(k_j, k_i, t_j - t_i) + \mu^2(k_j) \right] - \log \left[\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \right].$$

We can likewise compare $\log \frac{\sum_q w_q x_q}{\sum_q w_q}$ to $\log \sum_q x_q$:

$$\begin{aligned} \log \frac{\sum_q w_q x_q}{\sum_q w_q} - \log \sum_q x_q &= \log \left[\sum_{i=1}^N \sum_{r=1}^R h_r^2(k_j, k_i, t_j - t_i) + \mu^2(k_j) \right] - 2 \log \left[\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \right] \\ &= \log \sum_q x_q^2 - \log \left(\sum_q x_q \right)^2 \leq 0 \quad \text{by the multinomial theorem.} \end{aligned}$$

Hence

$$\begin{aligned} \frac{\sum_q w_q \log x_q}{\sum_q w_q} &\leq \log \frac{\sum_q w_q x_q}{\sum_q w_q} \leq \log \sum_q x_q. \\ \therefore \frac{\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) \log h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \log \mu(k_j)}{\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) + \mu(k_j)} &\leq \log \left[\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \right] \\ \sum_{i=1}^N \sum_{r=1}^R p_{ijr} \log h_r(k_j, k_i, t_j - t_i) + p_{bj} \log \mu(k_j) &\leq \log \left[\sum_{i=1}^N \sum_{r=1}^R h_r(k_j, k_i, t_j - t_i) + \mu(k_j) \right]. \end{aligned}$$

Therefore we hold that $\log L_c \leq \log L$. \square

Proof of Proposition 2

We are once again concerned with the behavior of

$$j \mapsto \sum_{i=1}^N \sum_{r=1}^R p_{ijr} \log h_r(k_j, k_i, t_j - t_i) + p_{bj} \log \mu(k_j),$$

now in the case that $p'_{ijr} = \frac{h'_r(k_j, k_i, t_j - t_i)}{\lambda'(k_j, t_j)}$ and the contents of the aforementioned logarithms do not match—hence the apostrophes. Note that the above is equivalent to

$$\sum_{i=1}^N \sum_{r=1}^R p_{ijr} \log \frac{h_r(k_j, k_i, t_j - t_i)}{\lambda(k_j, t_j)} + p_{bj} \log \frac{\mu(k_j)}{\lambda(k_j, t_j)} + \log \lambda(k_j, t_j),$$

since $\sum_{i=1}^N \sum_{r=1}^R p_{ijr} + p_{bj} = 1 \forall j$. Now let us examine the effects of a mismatch, $h'_r \neq h_r \iff p'_{ijr} \neq p_{ijr}$. We have, essentially,

$$\sum_{i=1}^N \sum_{r=1}^R p'_{ijr} \log p_{ijr} + p'_{bj} \log p_{bj} + \log \lambda(k_j, t_j).$$

When $[p]$ and $[p']$ are put under the lens of discrete/categorical distributions, we may parse the expression in terms of a cross entropy

$$-H([p'], [p]) + \log \lambda(k_j, t_j).$$

Since $[p']$ is independent of $\lambda(k_j, t_j)$, our quantity reaches its maximum when the distribution $[p']$ is set to equal $[p]$ (where $H([p'], [p]) = H([p']) + D_{KL}([p'] || [p])$ is minimized). The same cannot be said about maximization with respect to $[p]$, which depends on $\log \lambda(k_j, t_j)$ as well. The latter task proves useful in learning.

Thus when we perform a single learning iteration and vary the model parameters to evaluate $\log L_c$ on some $[p]$ fixed a priori, it will still offer a lower bound to $\log L$. \square

References

- Achab, M., Bacry, E., Gaiffas, S., Mastromatteo, I., & Muzy, J.-F. (2017). Uncovering causality from multivariate hawkes integrated cumulants. *International Conference on Machine Learning*, 34.
- Aglietti, V., Bonilla, E. V., Damoulas, T., & Cripps, S. (2019). Structured variational inference in continuous cox process models. *33rd Conference on Neural Information Processing Systems*.
- Apostolopoulou, I., Linderman, S., Miller, K., & Dubrawski, A. (2019). Mutually regressive point processes. *33rd Conference on Neural Information Processing Systems*.
- Bacry, E., & Muzy, J. F. (2016). First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4), 2184–2202.
- Bertozi, A. L., Franco, E., Mohler, G., Short, M. B., & Sledge, D. (2020). The challenges of modeling and forecasting the spread of covid-19. *Proc Natl Acad Sci U S A*, 117(29), 16732–16738.
- Campbell, T., & Li, X. (2019). Universal boosting variational inference. *33rd Conference on Neural Information Processing Systems*.
- Cho, Y.-S., Galstyan, A., Brantingham, P. J., & Tita, G. (2014). Latent self-exciting point process model for spatial-temporal networks. *Discrete and Continuous Dynamical Systems Series B*, 19(5), 1335–1354.
- Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, 21, 5–30.
- Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society B*, 17(2), 129–164.
- Detommaso, G., Hoitzing, H., Cui, T., & Alamir, A. (2019). Stein variational online changepoint detection with applications to hawkes processes and neural networks. *arXiv:1901.07987v2*.
- Ding, H., Khan, M. E., Sato, I., & Sugiyama, M. (2018). Bayesian nonparametric poisson-process allocation for time-sequence modeling. *arXiv:1705.07006*.
- Drakopoulos, K., Ozdaglar, A., & Tsitsiklis, J. N. (2017). When is a network epidemic hard to eliminate?. *Mathematics of Operations Research*, 42(1), 1–14.
- Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., & Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. *KDD*.
- Etesami, J., Kiyavash, N., Zhang, K., & Singhal, K. (2016). Learning network of multivariate hawkes processes: A time series approach. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*.
- Fefferman, C., Mitter, S., & Narayanan, H. (2016). Testing the manifold hypothesis. *J. Amer. Math. Soc.*, 29(4), 983–1049.
- Flaxman, S., Chirico, M., Pereira, P., & Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods. *arXiv:1801.02858*.

- Garske, T., Cori, A., Ariyaratnam, A., Blake, I. M., Dorigatti, I., Eckmanns, T., Fraser, C., Hinsley, W., Jombart, T., Mills, H. L., Nedjati-Gilani, G., Newton, E., Nouvellet, P., Perkins, D., Riley, S., Schumacher, D., Shah, A., Kerkhove, M. D. V., Dye, C., Ferguson, N. M., & Donnelly, C. A. (2017). Heterogeneities in the case fatality ratio in the west african ebola outbreak 2013-2016. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1721).
- Gibson, G. J., Streftaris, G., & Thong, D. (2018). Comparison and assessment of epidemic models. *Statistical Science*, *33*(1), 19–33.
- Goodhart, C. (1981). *Inflation, Depression, and Economic Policy in the West*, chap. Problems of Monetary Management: The U.K. Experience, pp. 111–146. Rowman & Littlefield.
- Halpin, P. F., & Boeck, P. D. (2013). Modelling dyadic interaction with hawkes processes. *Psychometrika*, *78*(4), 793–814.
- Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90.
- He, X., Rekatsinas, T., Foulds, J., Getoor, L., & Liu, Y. (2015). Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. *International Conference on Machine Learning*, *32*(37), 871–880.
- Hunter, D. R., & Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, *58*(1), 30–37.
- Jia, J., & Benson, A. R. (2019). Neural jump stochastic differential equations. *33rd Conference on Neural Information Processing Systems*.
- Junuthula, R. R., Haghdan, M., Xu, K. S., & Devabhaktuni, V. K. (2019). The block point process model for continuous-time event-based dynamic networks. *Proceedings of the World Wide Web Conference*.
- Kalatzis, D., Eklund, D., Arvanitidis, G., & Hauberg, S. (2020). Variational autoencoders with riemannian brownian motion priors. *arXiv:2002.05227*.
- Kerkhove, M. D. V., Bento, A. I., Mills, H. L., Ferguson, N. M., & Donnelly, C. A. (2015). A review of epidemiological parameters from ebola outbreaks to inform early public health decision-making. *Scientific Data*, *2*(150019).
- Kim, J. (2020). Ds4c: Data science for covid-19 in south korea. Tech. rep., Korea Centers for Disease Control & Prevention.
- Lemonnier, R., Scaman, K., & Kalogeratos, A. (2017). Multivariate hawkes processes for large-scale inference. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Li, H., Lindenbaum, O., Cheng, X., & Cloninger, A. (2019). Diffusion variational autoencoders. *arXiv:1905.12724*.
- Lian, W., Talmon, R., Zaveri, H., Carin, L., & Coifman, R. (2015). Multivariate time-series analysis and diffusion maps. *Signal Processing*, *116*, 13–28.
- Linderman, S. W., & Adams, R. P. (2015). Scalable bayesian inference for excitatory point process networks. *arXiv:1507.03228*.

- Liu, S., & Hauskrecht, M. (2019). Nonparametric regressive point processes based on conditional gaussian processes. *33rd Conference on Neural Information Processing Systems*.
- Lloyd, C., Gunter, T., Osborne, M., Roberts, S., & Nickson, T. (2016). Latent point process allocation. *International Conference on Artificial Intelligence and Statistics*, 19(51), 389–397.
- Loaiza-Ganem, G., Perkins, S. M., Schroeder, K. E., Churchland, M. M., & Cunningham, J. P. (2019). Deep random splines for point process intensity estimation of neural population data. *33rd Conference on Neural Information Processing Systems*.
- Lu, Y., Wang, X., Shi, C., Yu, P. S., & Ye, Y. (2019). Temporal network embedding with micro- and macro-dynamics. *Proceedings of the Conference of Information and Knowledge Management*.
- Marmarelis, V. Z., & Berger, T. W. (2005). General methodology for nonlinear modeling of neural systems with poisson point-process inputs. *Mathematical Biosciences*, 196, 1–13.
- McFadden, J. A. (1965). The entropy of a point process. *Society for Industrial and Applied Mathematics*, 13(4), 988–994.
- Mehrasa, N., Deng, R., Ahmed, M. O., Chang, B., He, J., Durand, T., Brubaker, M., & Mori, G. (2019). Point process flows. *arXiv:1910.08281*.
- Mei, H., & Eisner, J. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. *31st Conference on Neural Information Processing Systems*.
- Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 100–108.
- Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30, 491–497.
- Morariu-Patrichi, M., & Pakkanen, M. S. (2018). State-dependent hawkes processes and their application to limit order book modeling. *arXiv:1809.08060v2*.
- Nesterov, Y., & Polyak, B. (2006). Cubic regularization of newton method and its global performance. *Math. Program.*, A(108), 177–205.
- Newson, R. (2002). Parameters behind “nonparametric” statistics: Kendall’s tau, somers’ d and median differences. *The State Journal*, 2(1), 45–64.
- Nickel, M., & Le, M. (2020). Learning multivariate hawkes processes at scale. *arXiv:2002.12501*.
- Ogata, Y. (1981). On lewis’ simulation method for point processes. *IEEE Transactions on Information Theory*, IT-27(1), 23–31.
- Okawa, M., Iwata, T., Kurashima, T., Tanaka, Y., Toda, H., & Ueda, N. (2019). Deep mixture point processes: Spatio-temporal event prediction with rich contextual information. *KDD*.
- Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Delataille, S., De Nadai, M., Letouzé, E., Salah, A. A., Benjamins, R., Cattuto, C., Colizza, V., de Cordes, N., Fraiberger, S. P.,

- Koebe, T., Lehmann, S., Murillo, J., Pentland, A., Pham, P. N., Pivetta, F., Saramäki, J., Scarpino, S. V., Tizzoni, M., Verhulst, t., & Vinck, P. (2020). Mobile phone data for informing public health actions across the covid-19 pandemic life cycle. *Science Advances*.
- Omi, T. (2019). Fully neural network based model for general temporal point processes. *33rd Conference on Neural Information Processing Systems*.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Conference on Empirical Methods in Natural Language Processing*.
- Pham, K., & Chen, G. (2018). Large-scale spectral clustering using diffusion coordinates on landmark-based bipartite graphs. *Workshop on Graph-Based Methods for Natural Language Processing, 12*, 28–37.
- Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signaling in a complete neuronal population. *Nature, 454*, 995–999.
- Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33.
- Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science, 33*(3), 299–318.
- Rey, L. A. P., Menkovski, V., & Portegies, J. W. (2019). Diffusion variational autoencoders. *arXiv:1901.08991*.
- Salehi, F., Trouleau, W., Grossglauser, M., & Thiran, P. (2019). Learning hawkes processes from a handful of events. *33rd Conference on Neural Information Processing Systems*.
- Shang, J., & Sun, M. (2019). Geometric hawkes processes with graph convolutional recurrent neural networks. *Association for the Advancement of Artificial Intelligence*.
- Sharma, A., Ghosh, A., & Fiterau, M. (2019). Generative sequential stochastic model for marked point processes. *ICML Time Series Workshop*.
- Soize, C., & Ghanem, R. (2016). Data-driven probability concentration and sampling on manifold. *Journal of Computational Physics, 321*, 242–258.
- Student (1908). The probable error of a mean. *Biometrika, 6*(1).
- Swishchuk, A., & Huffman, A. (2020). General compound hawkes processes in limit order books. *Risks, 8*(28).
- Taylor, D., Klimm, F., Harrington, H. A., Kramár, M., Mischaikow, K., Porter, M. A., & Mucha, P. J. (2015). Topological data analysis of contagion maps for examining spreading processes on networks. *Nat Commun, 6*, 7723.
- Torricelli, M., Karsai, M., & Gauvin, L. (2020). weg2vec: Event embedding for temporal networks. *Scientific Reports, 10*(7164).
- Veen, A., & Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an em-type algorithm. *Journal of the American Statistical Association, 103*(482), 614–624.

- Welsh, B. (2020). Los angeles times. <https://github.com/datadesk/california-coronavirus-data>.
- Xu, H., Farajtabat, M., & Zha, H. (2016). Learning granger causality for hawkes processes. *Proceedings of the 33rd International Conference on Machine Learning*, 48.
- Yang, H., Kutan, A. M., & Ryu, D. (2018). Option moneyness and price disagreements. *Applied Economics Letters*, 25(3), 192–196.
- Yuan, B., Li, H., Bertozzi, A. L., Brantingham, P. J., & Porter, M. A. (2019). Multivariate spatiotemporal hawkes processes and network reconstruction. *SIAM J. Math. Data Sci.*, 1(2), 356–382.
- Zhang, R., Walder, C., Rizoïu, M.-A., & Xie, L. (2019). Efficient non-parametric bayesian hawkes processes. *International Joint Conference on Artificial Intelligence*, 28.
- Zhou, K., Zha, H., & Song, L. (2013a). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. *16th International Conference on Artificial Intelligence and Statistics*, 31.
- Zhou, K., Zha, H., & Song, L. (2013b). Learning triggering kernels for multi-dimensional hawkes processes. *30th International Conference on Machine Learning*.
- Zhu, S., Li, S., & Xie, Y. (2020). Interpretable generative neural spatio-temporal point processes. *arXiv:1906.05467v2*.
- Zhu, S., & Xie, Y. (2019). Crime event embedding with unsupervised feature selection. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.
- Zuo, Y., Liu, G., Lin, H., Guo, J., Hu, X., & Wu, J. (2018). Embedding temporal network via neighborhood formation. *KDD*.