

SCALABLE BAYESIAN INFERENCE FOR EXCITATORY POINT PROCESS NETWORKS

BY SCOTT W. LINDERMAN AND RYAN P. ADAMS

Harvard University

Networks capture our intuition about relationships in the world. They describe the friendships between Facebook users, interactions in financial markets, and synapses connecting neurons in the brain. These networks are richly structured with cliques of friends, sectors of stocks, and a smorgasbord of cell types that govern how neurons connect. Some networks, like social network friendships, can be directly observed, but in many cases we only have an indirect view of the network through the actions of its constituents and an understanding of how the network mediates that activity. In this work, we focus on the problem of latent network discovery in the case where the observable activity takes the form of a mutually-excitatory point process, also known as a Hawkes process. We build on previous work that has taken a Bayesian approach to this problem, specifying prior distributions over the latent network structure and a likelihood of observed activity given this network. We extend this work by proposing a discrete-time formulation and developing a computationally efficient stochastic variational inference (SVI) algorithm that allows us to scale the approach to long sequences of observations. We demonstrate our algorithm on the calcium imaging data used in the Chalearn neural connectomics challenge.

1. Introduction. Networks are abstractions of the relationships and connections between real-world objects, such as people, stocks, or neurons. These connections reflect relationships like “Wilson and Brady are friends,” or “When neuron A fires it excites neuron B.” Sometimes the networks themselves are observed data, as in the case of social network friendships, but often our view of the network is indirect. We are left to infer the latent connections between objects based on our observations of their behavior. In our neural example, recording techniques can often provide a measure of the neurons’ activity but cannot resolve the individual synaptic connections between neurons. Given our knowledge of how synapses work, however, we might infer that if one neuron consistently fires shortly after another then there is likely an excitatory connection between them. This is one example of the *latent network discovery* problem that this work addresses.

We focus on the case where our observations come in the form of a series of discrete events, like a sequence of Twitter messages or the firing pattern of a population of neurons. These events do not happen independently; rather, events induce subsequent events according to an excitatory network of interactions. A connection from one object to another indicates that events by the first object increase the probability of subsequent events by the second. We model these observations with a multivariate Hawkes process, a type of point process tailored to excitatory networks of interaction.

Building on the previous work, we combine the Hawkes process observation model with a prior distribution over networks in a unified Bayesian model (Simma and Jordan, 2010; Blundell et al., 2012; Perry and Wolfe, 2013; DuBois et al., 2013; Linderman and Adams, 2014; Guo et al., 2015). Most real-world networks are not simply random, but have highly structured patterns of interaction. For example, a friendship can often be traced back to some commonality between two people, such as belonging to the same club or attending the same school. A simple model for social networks might assign each person to a group, and then connect people according to whether or not they are

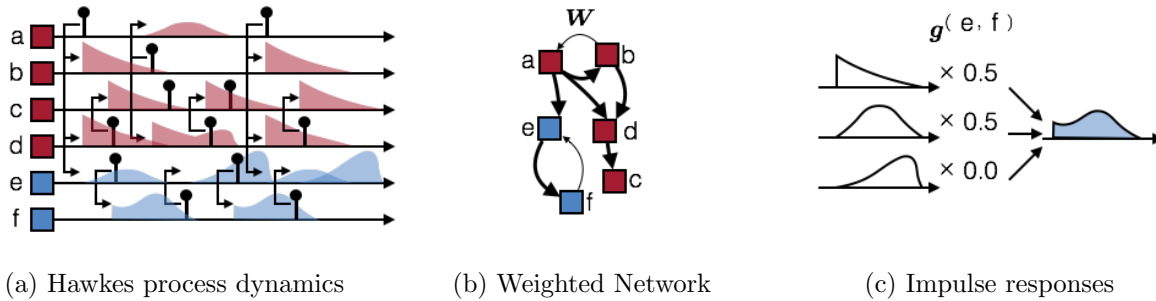


Fig 1: Components of the Network Hawkes process model. (a) Events induce weighted impulse responses on downstream processes, spawning more events according to the Hawkes process dynamics. (b) Underlying these dynamics is a weighted, directed network. Here, the network is a stochastic block model with two clusters of processes (red and blue). Processes primarily connect to others of the same type, though there is a small probability of connection from red to blue. (c) We model the impulse responses as a convex combination of normalized basis functions.

in the same group. This is known as a stochastic block model (SBM) (Nowicki and Snijders, 2001), and is one example of a random network model that may serve as prior probability distribution over networks in a Bayesian framework.

We improve upon previous work by providing a discrete-time analogue of the Hawkes process that is considerably more efficient on datasets with high rates of activity, and by devising an efficient stochastic variational inference (SVI) algorithm that can scale to long sequences of observations. Most previous work has relied upon Markov chain Monte Carlo (MCMC) methods for inference, which must consider the entire observation sequence when evaluating the likelihood of a state, and which can be prone to poor convergence. SVI provides an alternative method of inference that can work with mini-batches (small subsets) of observations per iteration, and has been shown to yield dramatic improvements in a variety of large-scale machine learning problems. Code for the models and algorithms developed in this paper is available at <https://github.com/slinderman/pyhawkes>.

2. Related Work. Hawkes processes (Hawkes, 1971) are multivariate point processes that relate excitatory interaction networks to sets of discrete events. For example, suppose we have a Hawkes process with two constituent processes. An event on the first process could add an *impulse response* to the future rate of the second process. More generally, a multivariate Hawkes process consists of K individual point processes with rates $\{\lambda_k(t)\}_{k=1}^K$ that depend upon the events that have occurred up to time t . This dependence on preceding events distinguishes the Hawkes process from the Poisson process. Given the history of events up to time t , however, Hawkes processes have conditionally Poisson dynamics.

Hawkes processes have additive, excitatory interactions. Each event adds a nonnegative impulse response to the future rate of connected processes. The rate of the k -th process is

$$(1) \quad \lambda_k(t) = \lambda_k^{(0)} + \sum_{k'=1}^K \sum_{n=1}^{N_{k'}} \mathbb{I}[s_{k',n} < t] \cdot h_{k' \rightarrow k}(t - s_{k',n}),$$

where $\{s_{k,n}\}_{n=1}^{N_k} \in [0, T]^{N_k}$ is the set of event times for events on process k , N_k is the total number of events on process k , $\lambda_k^{(0)}$ is the “background rate” of the k -th process, and $h_{k' \rightarrow k}(\Delta t)$ is an

impulse response function describing the amplitude of influence that events on process k' have on the rate of process k at time lag Δt .

The Hawkes process is closely related to the generalized linear model (GLM) with Poisson observations, which is widely used in computational neuroscience (Paninski, 2004; Pillow et al., 2008). In fact, the Hawkes process is a special case of the Poisson GLM in which the link function is linear and the impulse responses are non-negative. As in the Poisson GLM, the negative log likelihood of the Hawkes process is convex, enabling efficient maximum likelihood and maximum *a posteriori* estimation. The advantage of the Hawkes process is that its linear form allows for elegant fully-Bayesian inference — a task which is non-trivial in the Poisson GLM due to the lack of conjugacy. With these Bayesian inference algorithms, we can estimate and reason using the posterior uncertainty of the model.

One of the earliest applications of Hawkes processes in machine learning was the work of Simma and Jordan (2010), which developed an expectation-maximization algorithm based upon the auxiliary variable parent formulation. Observing that $\lambda_k(t)$ is a sum of impulse responses, Simma and Jordan (2010) invoked the Poisson superposition theorem to motivate a set of auxiliary “parent” variables, $z_{k,n}$, which denote the origin of the n -th event on process k , either the background rate or an impulse response from a previous event. Conditioned upon these auxiliary variables, the likelihood factorizes over impulse responses.

Subsequent works leveraged this intuition to extend Hawkes processes with interpretable prior distributions over the network of impulse responses. Blundell et al. (2012) introduced an infinite relational model prior over the network of interactions as well as a Gibbs sampling algorithm for fully Bayesian inference. DuBois et al. (2013) explored the use of infinite relational models as a prior in conjunction with a point process observation model and a Gibbs sampling inference algorithm. Recently, Linderman and Adams (2014) developed a general framework for combining random “spike-and-slab” network models with Hawkes processes that uses a continuous time formulation and an auxiliary Gibbs sampling inference algorithm. Guo et al. (2015) have developed a similar model that applies Hawkes processes to language modeling problems and incorporates features of the discrete events.

We extend the work of Linderman and Adams (2014) by addressing two shortcomings: the limited scalability of the continuous time formulation which introduces auxiliary variables for each event in the dataset, and the batch nature of their Gibbs sampling algorithm. We address the former by deriving a discrete time version of their model which vastly outperforms the continuous time version on datasets with high rates of activity. To overcome the batch nature of the Gibbs algorithm, we make an approximation to the spike-and-slab network model that renders the model fully conjugate, thereby enabling efficient stochastic variational inference (Hoffman et al., 2013) on mini-batches of data.

3. The Discrete Time Network Hawkes Model. The fundamental limitation of the previously developed continuous time models is that the number of values that the auxiliary variable $z_{k,n}$ can take grows with the number of events which occurred before time $s_{k,n}$. For datasets with high rates of activity, this can quickly become the limiting factor of the inference algorithm. At the same time, it is often reasonable to assume that events do not interact on time scales faster than some Δt . This motivates a discrete time formulation in which we bin events in bins of width Δt and ignore potential interactions between events in the same bin. Then the rate becomes,

$$(2) \quad \lambda_{t,k} = \lambda_k^{(0)} + \sum_{k'=1}^K \sum_{t'=1}^{t-1} s_{t',k'} \cdot h_{k' \rightarrow k}[t - t'],$$

where $\mathbf{s} \in \mathbb{N}^{T \times K}$ is the matrix of event counts and $h_{k' \rightarrow k}[t - t']$ is an impulse response function describing the amplitude of influence that events on process k' have on the rate of process k at discrete time lag $t - t'$. As we will show, under this formulation the auxiliary variables only assume a fixed set of values independent of the rate.

In order to incorporate the network model as a prior distribution for the Hawkes process, we follow the approach of [Linderman and Adams \(2014\)](#) and decompose the impulse response function into the product of a scalar weight that specifies the strength of the interaction and a probability mass function that specifies the time course of interaction:

$$h_{k \rightarrow k'}[d] = W_{k \rightarrow k'} G_{k \rightarrow k'}[d] \equiv W_{k \rightarrow k'} \sum_{b=1}^B g_b^{(k, k')} \phi_b[d],$$

for $d \in \{1, \dots, D\}$. Here, $\mathbf{W} \in \mathbb{R}_+^{K \times K}$ is a non-negative weight matrix drawn from a spike-and-slab prior,

$$A_{k \rightarrow k'} \sim \text{Bern}(A_{k \rightarrow k'} | p_{k \rightarrow k'}), \quad W_{k \rightarrow k'} \sim \begin{cases} \text{Gam}(W_{k \rightarrow k'} | \kappa, v_{k \rightarrow k'}) & \text{if } A_{k \rightarrow k'} = 1, \\ \delta_0(W_{k \rightarrow k'}) & \text{if } A_{k \rightarrow k'} = 0. \end{cases}$$

The network model provides $p_{k \rightarrow k'}$, the probability of a directed connection from node k to k' , and $v_{k \rightarrow k'}$, the inverse scale of the gamma distribution over the corresponding connection weight. We assume that κ , the shape of the weight prior, is fixed for simplicity. The matrix $\mathbf{A} = \{\{A_{k \rightarrow k'}\}\}$ is a binary, directed adjacency matrix indicating the presence or absence of a connection for each pair of nodes, and the matrix $\mathbf{W} = \{\{W_{k \rightarrow k'}\}\}$ is a non-negative weight matrix denoting the strength of each connection.

The vector $G_{k \rightarrow k'}$ is a normalized probability mass function. We model $G_{k \rightarrow k'}[d]$ as a convex combination of basis functions, ϕ_b , which are normalized such that $\sum_{d=1}^D \phi_b[d] \Delta t \equiv 1$, and require $\sum_{b=1}^B g_b^{(k, k')} \equiv 1$ under our model. This constraint is implemented via a Dirichlet prior, $\mathbf{g}^{(k, k')} \sim \text{Dir}(\boldsymbol{\gamma})$.

Intuitively, the weight $W_{k \rightarrow k'}$ specifies how many child events on process k' will be caused by a single event on process k . Then, $G_{k \rightarrow k'}[d]$ specifies the probability that the child event will occur at lag $d\Delta t$. In fact, this procedure is the basis of a recursive algorithm for generating samples from the discrete time Hawkes process.

Figure 1 illustrates the basic components of the model. In this example, we have a stochastic block model with two types of processes (red and blue) that preferentially interact with processes of the same type. Events induce weighted impulse responses on downstream processes according to an underlying latent network (Fig. 1b). The impulse responses, $G_{k \rightarrow k'}[d]$, are modeled as a convex combination of basis functions (Fig 1c). The impulse response is weighted according to the strength of the connection, $W_{k \rightarrow k'}$ before being added to the rate of downstream processes.

With fixed basis functions, we can write the instantaneous discrete time in a simplified form,

$$(3) \quad \lambda_{t, k} = \lambda_k^{(0)} + \sum_{k'=1}^K \sum_{b=1}^B W_{k' \rightarrow k} g_b^{(k', k)} \sum_{t'=1}^{t-1} s_{t', k'} \phi_b[t - t']$$

$$(4) \quad = \lambda_k^{(0)} + \sum_{k'=1}^K \sum_{b=1}^B W_{k' \rightarrow k} g_b^{(k', k)} \widehat{s}_{t, k', b},$$

where $\widehat{s}_{t, k', b} \equiv (\mathbf{s}_{:, k'} * \phi_b)[t]$ can be precomputed. Here, the instantaneous rate reduces to a sum of a weighted inputs, which suggests a Bayesian inference algorithm via data augmentation.

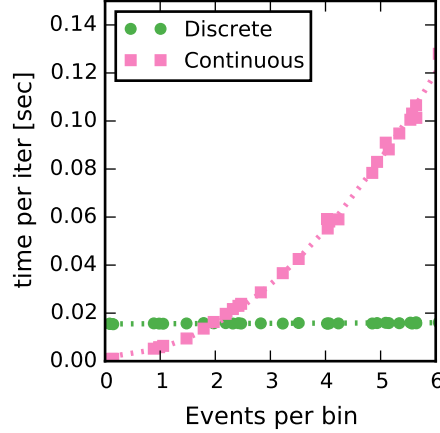


Fig 2: Comparison of run time per Gibbs sweep for the discrete and continuous network Hawkes formulations. Best fit lines added.

4. Inference with Gibbs Sampling. As in other preceding work (Simma and Jordan, 2010), we begin by introducing auxiliary parent variables for each entry $s_{t,k}$. By the superposition theorem for Poisson processes, each event can be attributed to either the background rate or one of the impulse responses.

Let $z_{t,k'}^{(k,b)} \in \{0, \dots, s_{t,k'}\}$ denote how many of the events that occurred in the t -th time bin on the k' -th process are attributed to the b -th basis function of the k -th process. Similarly, let $z_{t,k'}^{(0)}$ denote the number of events attributed to the background process. We combine these auxiliary variables into vectors, $\mathbf{z}_{t,k'} \triangleq [z_{t,k'}^{(0)}, z_{t,k'}^{(1,1)}, \dots, z_{t,k'}^{(K,B)}]$.

Due to the Poisson superposition principle, these parent variables are conditionally multinomial distributed. For time t and process k' , we resample

$$\mathbf{z}_{t,k'} \sim \text{Mult}(s_{t,k'}, \mathbf{u}_{t,k'}) \quad u_{t,k'}^{(0)} = \frac{\lambda_{k'}^{(0)}}{\lambda_{t,k'}}, \quad u_{t,k'}^{(k,b)} = \frac{W_{k \rightarrow k'} g_b^{(k,k')} \hat{s}_{t,k,b}}{\lambda_{t,k'}}.$$

Given this attribution, the likelihood factorizes into a product of Poisson distributions,

$$p(\mathbf{z} | \boldsymbol{\lambda}) = \left[\prod_{t=1}^T \prod_{k'=1}^K \text{Pois}(z_{t,k'}^{(0)} | \lambda_{k'}^{(0)} \Delta t) \right] \left[\prod_{t=1}^T \prod_{k=1}^K \prod_{k'=1}^K \text{Pois}(z_{t,k'}^{(k,b)} | W_{k \rightarrow k'} g_b^{(k,k')} \hat{s}_{t,k,b} \Delta t) \right].$$

Gibbs sampling the background rates. We use conjugate priors for the constant background rates, weights, and impulse responses. For the constant background rates we have, $\lambda_{k'}^{(0)} \sim \text{Gam}(\alpha_\lambda, \beta_\lambda)$, which results in the conditional distribution

$$\lambda_{k'}^{(0)} | \{z_{t,k}^{(0)}\} \sim \text{Gam}(\alpha_\lambda^{(k)}, \beta_\lambda^{(k)}), \quad \alpha_\lambda^{(k)} = \alpha_\lambda + \sum_{t=1}^T z_{t,k}^{(0)}, \quad \beta_\lambda^{(k)} = \beta_\lambda + T \Delta t.$$

Gibbs sampling impulse responses. The likelihood of the impulse responses, $\mathbf{g}^{(k,k')}$ is proportional to a Dirichlet distribution. Combined with a Dirichlet($\boldsymbol{\gamma}$) prior this yields

$$\mathbf{g}^{(k,k')} | \{z_{t,k}^{(k',b)}\}, \boldsymbol{\gamma} \sim \text{Dirichlet}(\boldsymbol{\gamma}^{(k,k')}), \quad \gamma_b^{(k,k')} = \gamma_b + \sum_{t=1}^T z_{t,k}^{(k,b)}.$$

Gibbs sampling the weighted adjacency matrix. Given the adjacency matrix \mathbf{A} and the parents, the weights follow a gamma distribution,

$$W_{k \rightarrow k'} | A_{k \rightarrow k'} = 1 \sim \text{Gam}(\kappa^{(k,k')}, v^{(k,k')}), \quad \kappa^{(k,k')} = \kappa + \sum_{t=1}^T \sum_{b=1}^B z_{t,k'}^{(k,b)}, \quad v^{(k,k')} = v_{k \rightarrow k'} + \sum_{t=1}^T s_{t,k}.$$

Following [Linderman and Adams \(2014\)](#), to resample \mathbf{A} , we iterate over each entry and sample from the marginal distribution after integrating out the parents. We assume the parameters of the network prior can be sampled efficiently — a reasonable assumption for many random network models.

The continuous time representation introduces a latent “parent” variable for each event in the dataset, and the parent can be any one of the events that occurred in the preceding window of influence. Call the number of potential parents M . The discrete time representation has a multinomial random variable for each time bin that contains at least one event, and the support of this multinomial is always a fixed size, $KB + 1$. When the rate of events is high, $KB + 1 \ll M$, allowing for dramatic improvements in efficiency in the discrete case.

Figure 2 shows the time per full Gibbs sweep as a function of the number of events per discrete time bin for the discrete and continuous formulations. The discrete formulation incurs a constant penalty whereas the continuous formulation quickly grows with the event rate. For low rates, the continuous formulation can be advantageous, but the discrete model is vastly superior in many realistic settings. For example, [Linderman and Adams \(2014\)](#) worked with trades on the S&P100, which occur tens or hundreds of times per second for each stock. Since the complexity of their algorithm grows with the number of events, they down-sampled the data to consider only the times when a stock price significantly changed.

5. Stochastic Variational Inference. The discrete time formulation offers advantageous complexity compared to the continuous analogue, but it still must resample the entire set of parents each iteration in order to maintain the invariance of the posterior distribution. In many cases, a mini-batch of time bins can provide substantial information about the global parameters of the model, and rapid progress can be made by iterating quickly over subsets of the data. This motivates our derivation of a stochastic variational inference (SVI) algorithm for the network Hawkes process.

Variational methods optimize a lower bound on the marginal likelihood by minimizing the KL-divergence between a tractable approximating distribution and the true posterior. Since the data-local variables (e.g., the parent identities) are conditionally independent given the global parameters (\mathbf{W} , \mathbf{g} , etc.), our variational approach will easily extend to the stochastic setting in which we compute unbiased estimates of the gradient of the variational objective using mini-batches of data.

The primary impediment to deriving a variational approximation is the non-conjugacy of the spike-and-slab prior on the weights. To overcome this, we approximate the spike-and-slab prior with a mixture of gamma distributions, as has previously explored by [Grabska-Barwinska et al. \(2013\)](#):

$$p(A_{k \rightarrow k'}) = \text{Bern}(A_{k \rightarrow k'} | p_{k \rightarrow k'}),$$

$$p(W_{k \rightarrow k'} | A_{k \rightarrow k'}) = \begin{cases} \text{Gam}(W_{k \rightarrow k'} | \kappa, v_{k \rightarrow k'}) & \text{if } A_{k \rightarrow k'} = 1, \\ \text{Gam}(W_{k \rightarrow k'} | \kappa_0, \nu_0) & \text{if } A_{k \rightarrow k'} = 0. \end{cases}$$

As $\kappa_0 \rightarrow 0$ and $\nu_0 \rightarrow \infty$, the second mixture component converges to a delta function at zero and recovers the true spike and slab model. As we relax this prior, the weights will be nonnegative when $A = 0$, but they will be small relative to the weights when $A = 1$. Importantly, with this

prior the model is rendered conjugate and amenable to a matching variational factor for each pair $(A_{k \rightarrow k'}, W_{k \rightarrow k'})$. Following [Lázaro-Gredilla and Titsias \(2011\)](#), let,

$$(5) \quad q(A_{k \rightarrow k'}) = \text{Bern}(A_{k \rightarrow k'} | \tilde{p}_{k \rightarrow k'}),$$

$$(6) \quad q(W_{k \rightarrow k'} | A_{k \rightarrow k'}) = \begin{cases} \text{Gam}(W_{k \rightarrow k'} | \tilde{\kappa}_1^{(k,k')}, \tilde{v}_1^{(k,k')}) & \text{if } A_{k \rightarrow k'} = 1, \\ \text{Gam}(W_{k \rightarrow k'} | \tilde{\kappa}_0^{(k,k')}, \tilde{v}_0^{(k,k')}) & \text{if } A_{k \rightarrow k'} = 0. \end{cases}$$

The rest of the variational approximation is fully factorized. Since the model is now conjugate, factors are easily derived. We provide a complete derivation in Appendix B and state the final forms here.

Variational updates for parent variables, $q(\mathbf{z}_t)$. For the parent variables, the variational updates are

$$(7) \quad q(\mathbf{z}_{t,k'}) = \text{Mult}(\mathbf{z}_{t,k'} | s_{t,k'}, \tilde{\mathbf{u}}_{t,k'})$$

$$(8) \quad \tilde{u}_{t,k'}^{(0)} \propto \exp \left\{ \mathbb{E}_\lambda [\ln \lambda_k^{(0)}] \right\}$$

$$(9) \quad \tilde{u}_{t,k'}^{(k,b)} \propto \hat{s}_{t,k,b} \exp \left\{ \mathbb{E}_g [\ln g_b^{(k,k')}] + \mathbb{E}_W [\ln W_{k,k'}] \right\}.$$

Variational updates for background rates, $q(\lambda_k^{(0)})$. The variational form parameters of the gamma distribution over background rates are

$$q(\lambda_k^{(0)}) = \text{Gam}(\lambda_k^{(0)} | \tilde{\alpha}_\lambda^{(k)}, \tilde{\beta}_\lambda^{(k)}), \quad \tilde{\alpha}_\lambda^{(k)} = \alpha_\lambda + \sum_{t=1}^T \mathbb{E}_z [z_{t,k}^{(0)}], \quad \tilde{\beta}_\lambda^{(k)} = \beta_\lambda + T \Delta t.$$

Variational approximation for impulse response parameters, $q(\mathbf{g}^{(k,k')})$. With the conjugate prior formulation the variational parameter updates for the Dirichlet distributed impulse response parameters are

$$q(\mathbf{g}^{(k,k')}) = \text{Dir}(\mathbf{g}^{(k,k')} | \tilde{\gamma}^{(k,k')}), \quad \tilde{\gamma}_b^{(k,k')} = \gamma_b + \sum_{t=1}^T \mathbb{E}_z [z_{t,k'}^{(k,b)}].$$

Variational approximation for the weighted adjacency matrix. The primary motivation for adopting a weakly sparse mixture of gamma distributions is to derive an efficient variational inference algorithm. The mixture-of-gammas prior is conjugate with the Poisson observations, and hence the variational distribution is also a mixture of gammas:

$$q(W_{k \rightarrow k'} | A_{k \rightarrow k'} = 1) = \text{Gam}(W_{k \rightarrow k'} | \tilde{\kappa}_1^{(k,k')}, \tilde{v}_1^{(k,k')})$$

$$\tilde{\kappa}_1^{(k,k')} = \kappa + \sum_{t=1}^T \sum_{b=1}^B \mathbb{E} [z_{t,k'}^{(k,b)}] \quad \tilde{v}_1^{(k,k')} = \mathbb{E}[v_{k \rightarrow k'}] + \sum_{t=1}^T s_{t,k},$$

and likewise for the ‘‘spike’’ factor,

$$q(W_{k \rightarrow k'} | A_{k \rightarrow k'} = 0) = \text{Gam}(W_{k \rightarrow k'} | \tilde{\kappa}_0^{(k,k')}, \tilde{v}_0^{(k,k')})$$

$$\tilde{\kappa}_0^{(k,k')} = \kappa_0 + \sum_{t=1}^T \sum_{b=1}^B \mathbb{E} [z_{t,k'}^{(k,b)}] \quad \tilde{v}_0^{(k,k')} = \nu_0 + \sum_{t=1}^T s_{t,k}.$$

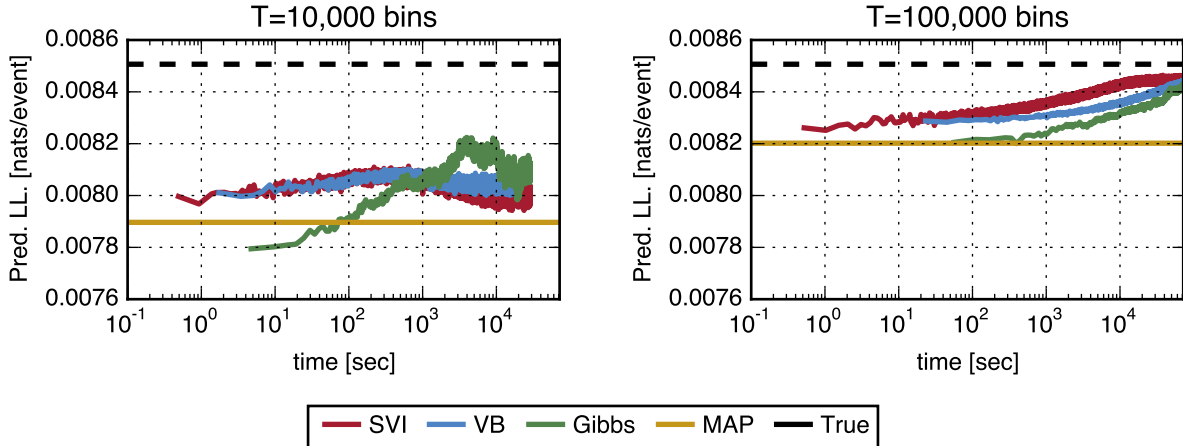


Fig 3: Predictive log likelihood versus wall clock time for three Bayesian inference algorithms on a dataset of $K = 50$ processes and $T = 10^4$ and $T = 10^5$ time bins on the left and right, respectively.

This leaves us with $q(A_{k \rightarrow k'})$, which is Bernoulli distributed with parameter $\tilde{p}_{k \rightarrow k'}$. Collecting all the terms that include $A_{k \rightarrow k'}$ and lack $W_{k \rightarrow k'}$ yields

$$\frac{\tilde{p}_{k \rightarrow k'}}{1 - \tilde{p}_{k \rightarrow k'}} = \frac{\exp\{\mathbb{E}[\ln p_{k \rightarrow k'}]\}}{\exp\{\mathbb{E}[\ln(1 - p_{k \rightarrow k'})]\}} \times \frac{(\exp\{\mathbb{E}[\ln v_{k \rightarrow k'}]\})^\kappa}{\Gamma(\kappa)} \times \frac{\Gamma(\tilde{\kappa}_1^{(k,k')})}{(\tilde{v}_1^{(k,k')})^{\tilde{\kappa}_1^{(k,k')}}} \times \frac{\Gamma(\kappa_0)}{(\nu_0)^{\kappa_0}} \times \frac{(\tilde{v}_0^{(k,k')})^{\tilde{\kappa}_0^{(k,k')}}}{\Gamma(\tilde{\kappa}_0^{(k,k')})}.$$

As with Gibbs sampling, we assume a variational approximation for the network model can be derived, and provide access to the necessary expectations, $\mathbb{E}[\ln p_{k \rightarrow k'}]$, $\mathbb{E}[\ln(1 - p_{k \rightarrow k'})]$, $\mathbb{E}[v_{k \rightarrow k'}]$, and $\mathbb{E}[\ln v_{k \rightarrow k'}]$.

As aforementioned, the time bins are conditionally independent given the network weights and the adjacency matrix — a common pattern exploited by stochastic variational inference (SVI) algorithms (Hoffman et al., 2013). These methods optimize the variational objective using stochastic gradient methods that work with mini-batches of data. Often, a mini-batch of data can provide valuable information about the global parameters, in our case the network and background rates. Quickly iterating over these global parameters allows us to reach good modes of the posterior distribution in a fraction of the time that batch variational Bayes and Gibbs sampling require, since those methods must process the entire dataset before making an update. SVI does require some tuning, however. In particular, we must set a mini-batch size and a step size schedule. In this work, we fix the mini-batch size to $T_{\text{mb}} = 1024$ and set the step size at iteration i to $(i + 1)^{-0.5}$. These parameters may be tuned with general purpose hyperparameter optimization techniques (Snoek et al., 2012).

6. Synthetic Results. We assess the performance of the proposed inference algorithms on a synthetic dataset generated by a strongly sparse, discrete time Hawkes process with $K = 50$ processes. The network is an Erdős-Renyi graph with uniform connection probability of $p = 0.08$, and the weights are independently and identically distributed with a $\text{Gam}(3, 15)$ prior. We simulate $T = 10^5$ time bins in steps of size $\Delta t = 1$. The processes have a mean background rate of 1.0 event per time bin and, due to the network interactions, the average total rate of the processes is 16.7 ± 12.0 events per bin. Referring to Figure 2, this is a regime that favors the discrete model. Then we trained the strongly sparse discrete time model using Gibbs sampling, and the weakly

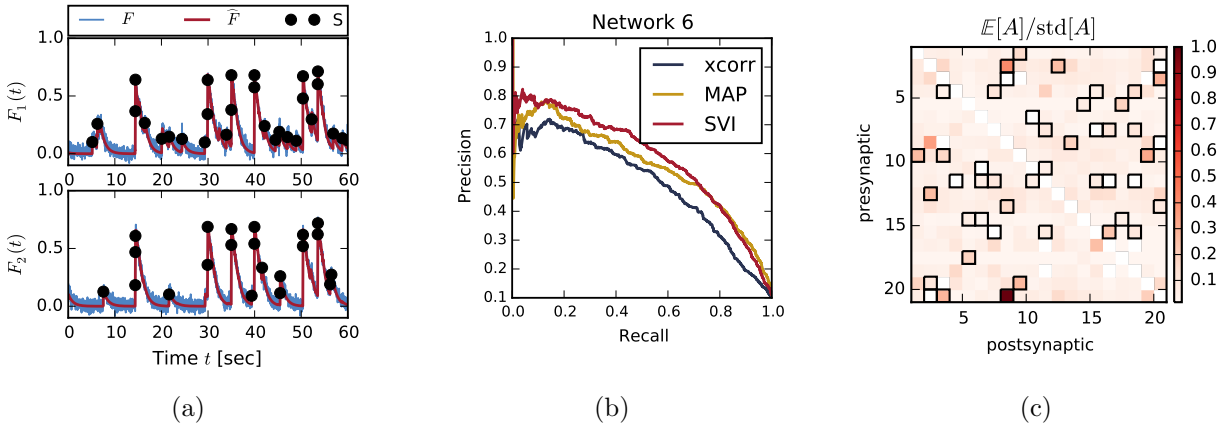


Fig 4: Application of the network Hawkes model to a connectomics challenge. The data are in the form of a calcium fluorescence trace, which we preprocess to extract neural spike times (a). We measure performance on a link prediction task using a precision-recall curve and find that the posterior estimates of SVI provide the best estimates on some networks (b). In addition to an estimate of the connection probability and weight, SVI provides an estimate of the posterior uncertainty. (c) shows $\mathbb{E}_q[\mathbf{A}]/\text{std}_q[\mathbf{A}]$ for the first 20 neurons. True connections are outlined in black.

sparse discrete time model using either batch variational Bayesian inference (VB) or stochastic variational inference (SVI)

We evaluated the algorithms in terms of their predictive log likelihood on a held-out dataset of length $T_{\text{test}} = 10^3$. First, we trained the models on only the first $T = 10^4$ time bins of data. Figure 3 (left) shows the predictive log likelihood as a function of wall-clock time, measured in units of nats per event improvement over a homogeneous Poisson process baseline. We initialized the Gibbs sampling and variational inference algorithms by rounding the cross-validated MAP estimate as described in the appendix. For this relatively short dataset, SVI and batch VB converge at comparable rates, achieving competitive predictive log likelihood in a matter of minutes. Gibbs sampling for the strongly sparse model converges at a considerably slower rate, but eventually outperforms the variational results from the weakly sparse model. The MAP estimate, even with cross validated regularization, underperforms the other competing algorithms.

This trend is amplified when we consider the entire training set of size $T = 10^5$. Figure 3 (right) illustrates the power of SVI in handling these large time datasets. Considerable information about the global parameters (e.g., the network) can be gained from just a mini-batch of time points. Hence, we can make rapid improvements in predictive log likelihood very quickly. By contrast, each step of the Gibbs and batch VB algorithms is approximately 10 times slower, and even after computing sufficient statistics over the entire dataset, the algorithm is only able to make limited progress per iteration.

7. Connectomics Results. We tested these inference algorithms on the data from the Chalearn neural connectomics challenge¹ (Stetter et al., 2012). The data consist of calcium fluorescence traces, \mathbf{F} , from six networks of $K = 100$ neurons each. We use ten minutes of data at 50Hz sampling frequency to yield $T = 3 \times 10^6$ entries in \mathbf{S} . In this case, the networks are purely excitatory, and each action potential, or spike, increases the probability of the downstream neurons firing as a result. This matches the underlying intuition of the Hawkes process model, making it a

¹<http://connectomics.chalearn.org>

Algorithm	Network 1		Network 2		Network 3		Network 4		Network 5	
	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC	ROC	PRC
xcorr	0.596	0.139	0.591	0.133	0.701	0.198	0.745	0.296	0.798	0.359
MAP	0.607	0.174	0.619	0.143	0.698	0.178	0.790	0.334	0.859	0.408
SVI	0.649	0.184	0.605	0.141	0.673	0.176	0.774	0.342	0.844	0.410

TABLE 1

Comparison of inference algorithms on link prediction for five networks from the Chalearn connectomics challenge. Performance is measured by area under the ROC curve and area under the precision recall curve (PRC). In four of the five networks a Hawkes process model provides the best results.

natural choice.

In order to apply the Hawkes model, we first convert the fluorescence traces into a spike count matrix using OOPSI, a Bayesian inference algorithm based on a model of calcium fluorescence (Vogelstein et al., 2010). The output is a filtered fluorescence trace, $\hat{\mathbf{F}}$, and a probability of spike for each time bin. We threshold this at probability 0.7 to get a $T \times K$ binary spike matrix, \mathbf{s} . This preprocessing is shown in Figure 4a.

Figure 4b shows the precision-recall curve we used to evaluate the algorithms’ performance on network recovery. As a baseline, we compare against simple thresholding of the cross correlation matrix. On Network 6, SVI offers the best network inference. Table 5 shows the results on the other five networks using the same model parameters. On 4/5 of these networks, the Bayesian methods offer the best performance.

Figure 4c illustrates one of the main advantages of the fully Bayesian inference algorithm – calibrated estimates of posterior uncertainty. Here we show the SVI algorithm’s estimate of the posterior mean of \mathbf{A} normalized by the posterior standard deviation for a subset of 20 neurons from Network 6. We also outline the true connections to show that the most confident predictions are more likely to correspond to true connections. Such estimates of the posterior uncertainty are not available with standard heuristic methods or point estimates.

8. Conclusion. We presented a scalable stochastic variational inference algorithm for the problem of Bayesian network discovery with Hawkes process observations. Building on previous modeling work, we leveraged a weak sparsity model to obtain a fully conjugate model. We focused on scaling to long duration recordings (large T). Scaling to large networks (large K) is nontrivial due to dependencies among weights, but in future work we hope to explore approximate algorithms to tackle this important problem.

Acknowledgments. We thank the members of the Harvard Intelligent Probabilistic Systems (HIPS) group, especially Matthew Johnson, for many helpful conversations. S.W.L. is supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. R.P.A. is partially supported by NSF IIS-1421780.

References.

- Aleksandr Simma and Michael I Jordan. Modeling events with cascades of Poisson processes. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.
- Charles Blundell, Katherine Heller, and Jeffrey Beck. Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems*, 2012.
- Patrick O Perry and Patrick J Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.
- Christopher DuBois, Carter Butts, and Padhraic Smyth. Stochastic blockmodeling of relational event dynamics. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.
- Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1413–1421, 2014.

- Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. The Bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2015.
- Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.
- Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83, 1971.
- Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.
- Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.
- Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
- Agnieszka Grabska-Barwinska, Jeff Beck, Alexandre Pouget, and Peter Latham. Demixing odors – Fast inference in olfaction. In *Advances in Neural Information Processing Systems*, pages 1968–1976, 2013.
- Miguel Lázaro-Gredilla and Michalis K Titsias. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2011.
- Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.
- Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS computational biology*, 8(8):e1002653, 2012.
- Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi, Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of neurophysiology*, 104(6):3691–3704, 2010.

APPENDIX A: DERIVATION OF THE GIBBS SAMPLING ALGORITHM

The updates for the weights and the background rates are straightforward extensions of those presented in (Linderman and Adams, 2014). The only nontrivial derivation of the Gibbs sampling is the update for the impulse response parameters. The likelihood of the impulse responses, $\mathbf{g}^{(k,k')}$ is proportional to a Dirichlet distribution. This can be seen by observing that the Hawkes process can be sampled by recursing over events. For each event we sample the number of offspring from $\text{Pois}(W_{k \rightarrow k'})$, and then we sample the offsets of those offspring from $G_{k \rightarrow k'}[d]$. Since G is modeled as a convex combination of basis functions, the choice of basis function is essentially a

draw from a categorical distribution. We can also derive this directly from the likelihood:

$$\begin{aligned}
p(\{\{z_{t,k'}^{(k,b)}\}_{t=1}^T\}_{b=1}^B \mid \mathbf{g}^{(k,k')}, \mathbf{W}) &\propto \prod_{t=1}^T \prod_{b=1}^B \text{Poisson}\left(z_{t,k'}^{(k,b)} \mid W_{k,k'} g_b^{(k,k')} \widehat{s}_{t,k,b} \Delta t\right) \\
&\propto \prod_{t=1}^T \prod_{b=1}^B \left(W_{k,k'} g_b^{(k,k')} \widehat{s}_{t,k,b} \Delta t\right)^{z_{t,k'}^{(k,b)}} \times \exp\left\{-W_{k,k'} g_b^{(k,k')} \widehat{s}_{t,k,b} \Delta t\right\} \\
&\propto \prod_{b=1}^B \left(g_b^{(k,k')}\right)^{\sum_{t=1}^T z_{t,k'}^{(k,b)}} \exp\left\{-W_{k,k'} g_b^{(k,k')} \sum_{t=1}^T \widehat{s}_{t,k,b}\right\} \\
&\propto \prod_{b=1}^B \left(g_b^{(k,k')}\right)^{\sum_{t=1}^T z_{t,k'}^{(k,b)}} \exp\left\{-W_{k,k'} g_b^{(k,k')} N_k\right\} \\
&\propto \left[\prod_{b=1}^B \left(g_b^{(k,k')}\right)^{\sum_{t=1}^T z_{t,k'}^{(k,b)}}\right] \exp\left\{-W_{k,k'} N_k \sum_{b=1}^B g_b^{(k,k')}\right\} \\
&\propto \left[\prod_{b=1}^B \left(g_b^{(k,k')}\right)^{\sum_{t=1}^T z_{t,k'}^{(k,b)}}\right] \exp\left\{-W_{k,k'} N_k\right\} \\
&\propto \text{Dirichlet}\left(\mathbf{g}^{(k,k')} \mid \left[\sum_{t=1}^T z_{t,k'}^{(k,1)}, \dots, \sum_{t=1}^T z_{t,k'}^{(k,B)}\right]\right).
\end{aligned}$$

Combined with a Dirichlet(γ) prior this yields,

$$\begin{aligned}
\mathbf{g}^{(k,k')} \mid \{z_{t,k}^{(k',b)}\}, \gamma &\sim \text{Dirichlet}(\gamma'), \\
\gamma'_b &= \gamma_b + \sum_{t=1}^T z_{t,k'}^{(k,b)}.
\end{aligned}$$

APPENDIX B: DERIVATION OF THE VARIATIONAL INFERENCE ALGORITHM

Our goal is to approximate the posterior distribution, $p(\mathbf{z}, \mathbf{A}, \mathbf{W}, \boldsymbol{\lambda}^{(0)}, \mathbf{g}, \boldsymbol{\theta}_{\text{net}} \mid \mathbf{s})$, with a variational distribution, $q(\mathbf{z}, \mathbf{A}, \mathbf{W}, \boldsymbol{\lambda}^{(0)}, \mathbf{g}, \boldsymbol{\theta}_{\text{net}})$, by minimizing the KL-divergence between q and p . As before, we have introduced auxiliary parent variables, \mathbf{z} , to decouple the events attributed to each potential parent process. We then restrict q to take a factorized form,

$$(10) \quad q(\mathbf{z}, \mathbf{A}, \mathbf{W}, \boldsymbol{\lambda}^{(0)}, \mathbf{g}, \boldsymbol{\theta}_{\text{net}}) = \prod_{t=1}^T q(\mathbf{z}_t) \prod_{k=1}^K q(\lambda_k^{(0)}) \prod_{k=1}^K \prod_{k'=1}^K q(A_{k,k'}, W_{k,k'}) q(\mathbf{g}^{(k,k')}) q(\boldsymbol{\theta}_{\text{net}})$$

This is essentially the same as a typical factorized approximation except that we have combined $A_{k,k'}$ and $W_{k,k'}$ into a single factor, as in [Lázaro-Gredilla and Titsias \(2011\)](#). This allows a multimodal spike-and-slab approximating distribution.

With this factorized approximation each individual factor must satisfy a set of *mean field* consistency equations in which the log of each factor is equal (up to a constant) to the expectation of the log posterior under the remaining variational factors.

Variational approximation for parent variables, $q(\mathbf{z}_t)$. For the parent variables, the consistency equations imply,

$$\begin{aligned} \ln q(z_{t,k'}^{(0)}) &= \mathbb{E}_{\boldsymbol{\lambda}^0} \left[\ln p(z_{t,k'}^{(0)}, \mathbf{A}, \mathbf{W}, \boldsymbol{\lambda}^{(0)} \mid \mathbf{s}) \right] + \text{const.} \\ &= \mathbb{E}_{\boldsymbol{\lambda}^0} \left[\ln p(z_{t,k'}^{(0)} \mid \lambda_k^{(0)}) \right] + \text{const.} \\ &= -\ln z_{t,k'}^{(0)}! - \mathbb{E}_{\boldsymbol{\lambda}^0} \left[\lambda_k^{(0)} \right] + z_{t,k'}^{(0)} \mathbb{E}_{\boldsymbol{\lambda}^0} \left[\ln \lambda_k^{(0)} \right] + \text{const.}, \end{aligned}$$

and

$$\begin{aligned} \ln q(z_{t,k'}^{(k,b)}) &= \mathbb{E}_{\mathbf{A}, \mathbf{W}, \mathbf{g}} \left[\ln p(z_{t,k'}^{(k,b)}, \mathbf{A}, \mathbf{W}, \boldsymbol{\lambda}^{(0)}, \mathbf{g} \mid \mathbf{s}) \right] + \text{const.} \\ &= \mathbb{E}_{\mathbf{A}, \mathbf{W}, \mathbf{g}} \left[\ln p(z_{t,k'}^{(k,b)} \mid \mathbf{A}, \mathbf{W}, \mathbf{g} \mid \mathbf{s}) \right] + \text{const.} \\ &= -\ln z_{t,k'}^{(k,b)}! - \mathbb{E}_{\mathbf{W}, \mathbf{A}, \mathbf{g}} \left[W_{k,k'} g_b^{(k,k')} \right] \hat{s}_{t,k,b} + z_{t,k'}^{(k,b)} \mathbb{E}_{\mathbf{W}, \mathbf{A}, \mathbf{g}} \left[\ln \left\{ W_{k,k'} g_b^{(k,k')} \hat{s}_{t,k,b} \right\} \right] + \text{const.} \end{aligned}$$

Combined with the constraint that $z_{t,k'}^{(0)} + \sum_{k,b} z_{t,k'}^{(k,b)} = s_{t,k'}$, this is the form of a multinomial distribution with variational parameter $\tilde{\mathbf{u}}_{t,k'}$,

$$\begin{aligned} (11) \quad q(\mathbf{z}_{t,k'}) &= \text{Multinomial}(\mathbf{z}_{t,k'} \mid s_{t,k'}, \tilde{\mathbf{u}}_{t,k'}), \\ \tilde{u}_{t,k'}^{(0)} &= \frac{1}{Z} \exp \left\{ \mathbb{E}[\ln \lambda_k^{(0)}] \right\} \\ \tilde{u}_{t,k'}^{(k,b)} &= \frac{1}{Z} \hat{s}_{t,k,b} \exp \left\{ \mathbb{E}[\ln g_b^{(k,k')}] \right\} \exp \left\{ \mathbb{E}[\ln W_{k,k'}] \right\} \\ Z &= \exp \left\{ \mathbb{E}[\ln \lambda_k^{(0)}] \right\} + \sum_{k=1}^K \sum_{b=1}^B \hat{s}_{t,k,b} \exp \left\{ \mathbb{E}[\ln g_b^{(k,k')}] \right\} \exp \left\{ \mathbb{E}[\ln W_{k,k'}] \right\}. \end{aligned}$$

Variational approximation for impulse response parameters, $q(\mathbf{g}^{(k,k')})$. The consistency equations yield,

$$\begin{aligned} \ln q(\mathbf{g}^{(k,k')}) &= \mathbb{E}_{\mathbf{z}, \mathbf{W}} [\ln p(\mathbf{z}_{t,k}, \mathbf{W}, \mathbf{g} \mid \mathbf{s})] + \text{const.} \\ &= \sum_{b=1}^B \left(\gamma_b + \sum_{t=1}^T \mathbb{E}_{\mathbf{z}} \left[z_{t,k'}^{(k,b)} \right] \right) \ln g_b^{(k,k')} + \text{const.} \end{aligned}$$

With the conjugate prior formulation the variational approximation is again a Dirichlet,

$$(12) \quad q(\mathbf{g}^{(k,k')}) = \text{Dirichlet}(\tilde{\boldsymbol{\gamma}}^{(k,k')}) \quad \tilde{\gamma}_b^{(k,k')} = \gamma_b + \sum_{t=1}^T \mathbb{E}_{\mathbf{z}} \left[z_{t,k'}^{(k,b)} \right].$$

Variational approximation for constant background rates, $q(\lambda_k^{(0)})$. The variational form for $q(\lambda_k^{(0)})$ is also determined by the conjugate model. We have,

$$\begin{aligned} \ln q(\lambda_k^{(0)}) &= \mathbb{E}_{\mathbf{z}} \left[\ln p(z_{t,k}, \mathbf{A}, \mathbf{W}, \boldsymbol{\lambda}^{(0)}, \mathbf{g} \mid \mathbf{s}) \right] + \text{const.} \\ &= \sum_{t=1}^T -\lambda_k^{(0)} \Delta t + \mathbb{E}_{\mathbf{z}} \left[z_{t,k}^{(0)} \right] \ln \lambda_k^{(0)} + (\alpha_\lambda - 1) \ln \lambda_k^{(0)} - \beta_\lambda \lambda_k^{(0)} + \text{const.} \end{aligned}$$

This is the form of a gamma distribution with variational parameters,

$$(13) \quad q(\lambda_k^{(0)}) = \text{Gam}(\tilde{\alpha}_\lambda^{(k)}, \tilde{\beta}_\lambda^{(k)}) \quad \tilde{\alpha}_\lambda^{(k)} = \alpha_\lambda + \sum_{t=1}^T \mathbb{E}_{\mathbf{z}} \left[z_{t,k}^{(0)} \right] \quad \tilde{\beta}_\lambda^{(k)} = \beta_\lambda + T \Delta t.$$

Variational approximation for spike-and-slab weights, $q(A_{k,k'}, W_{k,k'})$. Returning to the variational factor for \mathbf{z} in Equation 11, we see that a problem arises with the spike-and-slab model. If our model and our variational approximation are to share a spike-and-slab formulation then the expected log weight will be,

$$\mathbb{E}_{\mathbf{W}}[\ln W_{k,k'}] = \mathbb{E}_{\mathbf{A}} \mathbb{E}_{\mathbf{W} | \mathbf{A}}[\ln W_{k,k'}] = p \mathbb{E}[\ln W_{k,k'} | A_{k,k'} = 1] + (1 - p) \ln 0 = -\infty.$$

To avoid this degeneracy, we replace the delta function with a gamma distribution, $p(W_{k,k'} | A_{k,k'} = 0) = \text{Gam}(\kappa_0, v_0)$, which converges to a delta function as $\kappa_0 \rightarrow 0$ and $v_0 \rightarrow \infty$. The prior on weights is then a mixture of two gamma distributions, and is conjugate with the Poisson observations. This in turn implies a variational distribution that is also a mixture of gammas. We derive this with the consistency equations,

$$\begin{aligned} \ln q(A_{k,k'}, W_{k,k'}) &= \mathbb{E} \left[\ln p(\mathbf{z}_{t,k'}, \mathbf{A}, \mathbf{W}, \mathbf{g}, \boldsymbol{\theta}_{\text{net}} | \mathbf{s}) \right] + \text{const.} \\ &= \sum_{t=1}^T \sum_{b=1}^B -W_{k,k'} \mathbb{E}_{\mathbf{g}} \left[g_b^{(k,k')} \right] \widehat{s}_{t,k,b} \Delta t + \mathbb{E}_{\mathbf{z}} \left[z_{t,k'}^{(k,b)} \right] \ln W_{k,k'} \\ &\quad + A_{k,k'} \left[\kappa \mathbb{E}_{\boldsymbol{\theta}_{\text{net}}}[\ln v_{k \rightarrow k'}] - \Gamma(\kappa) + (\kappa - 1) \ln W_{k,k'} - \mathbb{E}_{\boldsymbol{\theta}_{\text{net}}}[v_{k \rightarrow k'}] W_{k,k'} + \mathbb{E}_{\boldsymbol{\theta}_{\text{net}}}[\ln p_{k \rightarrow k'}] \right] \\ &\quad + (1 - A_{k,k'}) \left[\kappa_0 \ln v_0 - \Gamma(\kappa_0)(\kappa_0 - 1) \ln W_{k,k'} - v_0 W_{k,k'} + \mathbb{E}_{\boldsymbol{\theta}_{\text{net}}}[\ln(1 - p_{k \rightarrow k'})] \right] + \text{const.} \end{aligned}$$

As expected, $q(W_{k,k'} | A_{k,k'})$ has the form of a gamma distribution,

$$\begin{aligned} (14) \quad q(W_{k,k'} | A_{k,k'} = 1) &= \text{Gam}(W_{k,k'} | \kappa_1^{(k,k')}, v_1^{(k,k')}) \\ \kappa_1^{(k,k')} &= \kappa + \sum_{t=1}^T \sum_{b=1}^B \mathbb{E}_{\mathbf{z}} \left[z_{t,k'}^{(k,b)} \right] \\ v_1^{(k,k')} &= \mathbb{E}_{\boldsymbol{\theta}_{\text{net}}}[v_{k \rightarrow k'}] + \sum_{t=1}^T \sum_{b=1}^B \mathbb{E}_{\mathbf{g}} \left[g_b^{(k,k')} \right] \widehat{s}_{t,k,b} \Delta t \\ &= \mathbb{E}_{\boldsymbol{\theta}_{\text{net}}}[v_{k \rightarrow k'}] + N_k \sum_{b=1}^B \mathbb{E}_{\mathbf{g}} \left[g_b^{(k,k')} \right], \end{aligned}$$

and

$$\begin{aligned} (15) \quad q(W_{k,k'} | A_{k,k'} = 0) &= \text{Gam}(W_{k,k'} | \kappa_0^{(k,k')}, v_0^{(k,k')}) \\ \kappa_0^{(k,k')} &= \kappa_0 + \sum_{t=1}^T \sum_{b=1}^B \mathbb{E}_{\mathbf{z}} \left[z_{t,k'}^{(k,b)} \right] \\ v_0^{(k,k')} &= v_0 + \sum_{t=1}^T \sum_{b=1}^B \mathbb{E}_{\mathbf{g}} \left[g_b^{(k,k')} \right] \widehat{s}_{t,k,b} \Delta t \\ &= v_0 + N_k \sum_{b=1}^B \mathbb{E}_{\mathbf{g}} \left[g_b^{(k,k')} \right]. \end{aligned}$$

This leaves us with $q(A_{k,k'})$, which we take to be Bernoulli distributed with parameter $\tilde{p}_{k,k'}$. This implies,

$$A_{k,k'} \left[\ln \tilde{p}_{k \rightarrow k'} + \ln \text{Gam}(\tilde{\kappa}_1^{(k,k')}, \tilde{v}_1^{(k,k')}) \right] + (1 - A_{k,k'}) \left[\ln(1 - \tilde{p}_{k \rightarrow k'}) + \ln \text{Gam}(\tilde{\kappa}_0^{(k,k')}, \tilde{v}_0^{(k,k')}) \right] = A_{k,k'} \left[\mathbb{E}_{\theta_{\text{net}}}[\ln p_{k \rightarrow k'}] + \mathbb{E}_{\theta_{\text{net}}}[\ln \text{Gam}(\kappa, v_{k \rightarrow k'})] \right] + (1 - A_{k,k'}) \left[\mathbb{E}_{\theta_{\text{net}}}[\ln(1 - p_{k \rightarrow k'})] + \ln \text{Gam}(\kappa_0, v_0) \right].$$

Collecting all the terms that include $A_{k,k'}$ and lack $W_{k,k'}$ yields,

$$(16) \quad \frac{\tilde{p}_{k \rightarrow k'}}{1 - \tilde{p}_{k \rightarrow k'}} = \frac{\exp\{\mathbb{E}_{\theta_{\text{net}}}[\ln p_{k \rightarrow k'}]\}}{\exp\{\mathbb{E}_{\theta_{\text{net}}}[\ln(1 - p_{k \rightarrow k'})]\}} \times \frac{(\exp\{\mathbb{E}_{\theta_{\text{net}}}[\ln v_{k \rightarrow k'}]\})^\kappa}{\Gamma(\kappa)} \times \frac{\Gamma(\tilde{\kappa}_1^{(k,k')})}{(\tilde{v}_1^{(k,k')})^{\tilde{\kappa}_1^{(k,k')}}} \times \frac{\Gamma(\kappa_0)}{(v_0)^{\kappa_0}} \times \frac{(\tilde{v}_0^{(k,k')})^{\tilde{\kappa}_0^{(k,k')}}}{\Gamma(\tilde{\kappa}_0^{(k,k')})}.$$

Variational updates for the network model. In the above derivations, the only requirement on the network model is that it provide the following expectations: $\mathbb{E}[p_{k \rightarrow k'}]$, $\mathbb{E}[\ln p_{k \rightarrow k'}]$, $\mathbb{E}[\ln(1 - p_{k \rightarrow k'})]$, $\mathbb{E}[v_{k \rightarrow k'}]$, and $\mathbb{E}[\ln v_{k \rightarrow k'}]$. It may make sense to fix one of these values; for example, we may have a good empirical estimate of the weight scale, $v_{k \rightarrow k'}$, and therefore do not need to model its posterior distribution. Alternatively, we may know the overall sparsity but be uncertain of the scale. For some models, computing the necessary variational expectations and deriving the variational updates is straightforward. For example, a stochastic block model (Nowicki and Snijders, 2001) with a gamma prior on the scale is fully conjugate and admits closed form updates.

Completing the variational expectations. Now we can compute the necessary expectations for the variational parameter updates. These are all available in closed form,

$$\begin{aligned} \mathbb{E} \left[\ln \lambda_k^{(0)} \right] &= \psi(\tilde{\alpha}_\lambda^{(k)}) - \ln \tilde{\beta}_\lambda^{(k)}, \\ \mathbb{E} \left[\ln W_{k,k'} \right] &= \tilde{p}_{k \rightarrow k'} \left(\psi(\tilde{\kappa}_1^{(k,k')}) - \ln \tilde{v}_1^{(k,k')} \right) + (1 - \tilde{p}_{k \rightarrow k'}) \left(\psi(\tilde{\kappa}_0^{(k,k')}) - \ln \tilde{v}_0^{(k,k')} \right), \\ \mathbb{E} \left[\ln g_b^{(k,k')} \right] &= \psi \left(\tilde{\gamma}_b^{(k,k')} \right) - \psi \left(\sum_{b'=1}^B \tilde{\gamma}_{b'}^{(k,k')} \right), \end{aligned}$$

where $\psi(\cdot)$ denotes the digamma function. To complete the variational updates for the background rate, impulse response, and weight parameters we have,

$$\mathbb{E} \left[z_{t,k'}^{(0)} \right] = \tilde{u}_{t,k'}^{(0)} s_{t,k'}, \quad \mathbb{E} \left[z_{t,k'}^{(k,b)} \right] = \tilde{u}_{t,k'}^{(k,b)} s_{t,k'},$$

and

$$\mathbb{E} \left[g_b^{(k,k')} \right] = \frac{\tilde{\gamma}_b^{(k,k')}}{\sum_{b'=1}^B \tilde{\gamma}_{b'}^{(k,k')}}.$$

With this final expectation, the parameter updates for the variational weight distributions simplify to $\tilde{v}_0^{(k,k')} = v_0 + N_k$ and $\tilde{v}_1^{(k,k')} = \mathbb{E}[v_{k \rightarrow k'}] + N_k$.

B.1. Initialization. Variational inference algorithms can be quite sensitive to initialization of their model parameters. A poor initialization may converge to a suboptimal mode of the posterior distribution. In order to initialize our algorithm, we first fit a standard (log concave) Hawkes

process using MAP estimation on a subset of the data. We use an exponential prior on the weights, equivalent to L1 regularization, and we tune the scale of the prior with cross validation. To convert the weights inferred under the standard Hawkes model to a sample of the network Hawkes model, we keep the largest p -fraction of the weights and set the remainder to zero. Finally, we initialize the variational parameters of the network Hawkes model such that they are peaked at the sample values. For example, we set $\tilde{v}_1^{(k,k')} = 100$ and $\tilde{\kappa}_1^{(k,k')} = 100 \cdot W_{k \rightarrow k'}$.