Discovering Latent Network Structure in Point Process Data (Supplementary Material)

Scott W. Linderman SLINDERMAN@SEAS.HARVARD.EDU
Harvard University, Cambridge, MA 02138 USA

Ryan P. Adams RPA@SEAS.HARVARD.EDU
Harvard University, Cambridge, MA 02138 USA

A. Inference details

A.1. Derivation of conjugate prior updates

By combining the Poisson process and the Hawkes process likelihoods given in the main text, we can write the joint likelihood, with the auxiliary parent variables, as,

$$p(\{s_n, c_n, z_n\}_{n=1}^N, \mid \{\lambda_{0,k}(t)\}_{k=1}^K, \{h_{k,k'}(\Delta t)\}_{k,k'}) =$$
$$\prod_{k=1}^K \left[\exp\left\{ -\int_0^T \lambda_{0,k}(\tau) \mathrm{d}\tau \right\} \prod_{n=1}^N \lambda_{0,k}(s_n)^{\delta_{c_n,k}\delta_{z_n,0}} \right]$$
$$\times \prod_{n=1}^N \prod_{k'=1}^K \left[\exp\left\{ -\int_{s_n}^T h_{c_n,k'}(\tau - s_n) \mathrm{d}\tau \right\} \right.$$
$$\left. \prod_{n'=1}^N h_{c_n,c_{n'}}(s_{n'} - s_n)^{\delta_{c_{n'},k'}\delta_{z_{n'},n}} \right].$$

The first line corresponds to the likelihood of the background processes; the second and third correspond to the likelihood of the induced processes triggered by each spike.

To derive the updates for weights, recall from Equation 2 of the main text that $W_{k,k'}$ only appears in the impulse responses for which $c_n = k$ and $c_{n'} = k'$. so we have,

$$p(W_{k,k'} \mid \{s_n, c_n, z_n\}_{n=1}^N, \ldots)$$
$$\propto \prod_{n=1}^N \left[\exp\left\{ -\int_{s_n}^T h_{k,k'}(\tau - s_n) \mathrm{d}\tau \right\} \right.$$
$$\left. \prod_{n'=1}^N h_{k,k'}(s_{n'} - s_n)^{\delta_{c_{n'},k'}\delta_{z_{n'},n}} \right]^{\delta_{c_n,k}} \times p(W_{k,k'})$$
$$= \prod_{n=1}^N \left[\exp\left\{ -\int_{s_n}^T A_{k,k'} W_{k,k'} g_{k,k'}(\tau - s_n) \mathrm{d}\tau \right\} \right.$$
$$\left. \prod_{n'=1}^N (A_{k,k'} W_{k,k'} g_{k,k'}(s_{n'} - s_n))^{\delta_{c_{n'},k'}\delta_{z_{n'},n}} \right]^{\delta_{c_n,k}}$$
$$\times p(W_{k,k'}).$$

If $A_{k,k'} = 1$ and we ignore spikes after $T - \Delta t_{\mathsf{max}}$, this is proportional to

$$\exp\{-W_{k,k'} N_k\} W_{k,k'}^{N_{k,k'}} p(W_{k,k'}),$$

where

$$N_k = \sum_{n=1}^N \delta_{c_n,k}, \text{ and } N_{k,k'} = \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k}\delta_{c_{n'},k'}\delta_{z_{n'},n}.$$

When $p(W_{k,k'})$ is a gamma distribution, the conditional distribution is also gamma. If $A_{k,k'} = 0$, the conditional distribution reduces to the prior, as expected.

The derivations of conjugate updates for constant background rates follows the same pattern.

We use a logistic normal distribution for the impulse responses.

$$g_{k,k'}(\Delta t \mid \mu, \tau) = \frac{1}{Z} \exp\left\{ \frac{-\tau}{2} \left(\sigma^{-1}\left(\frac{\Delta t}{\Delta t_{\mathsf{max}}} \right) - \mu \right)^2 \right\}$$
$$\sigma^{-1}(x) = \ln(x/(1-x))$$
$$Z = \frac{\Delta t(\Delta t_{\mathsf{max}} - \Delta t)}{\Delta t_{\mathsf{max}}} \left(\frac{\tau}{2\pi} \right)^{-\frac{1}{2}}.$$

The normal-gamma prior $\mu, \tau \sim \mathcal{NG}(\mu, \tau | \mu_\mu^0, \kappa_\mu^0, \alpha_\tau^0, \beta_\tau^0)$ on the parameters of the logistic normal distribution is conjugate with the likelihood. To see this, note that the only place $g_{k,k'}(\Delta t)$ appears in the likelihood is in

$$\prod_{n=1}^N \left[\exp\left\{ -\int_{s_n}^T A_{k,k'} W_{k,k'} g_{k,k'}(\tau - s_n) \mathrm{d}\tau \right\} \right.$$
$$\left. \prod_{n'=1}^N (A_{k,k'} W_{k,k'} g_{k,k'}(s_{n'} - s_n))^{\delta_{c_{n'},k'}\delta_{z_{n'},n}} \right]^{\delta_{c_n,k}}.$$

Since the impulse responses are probability density functions, $g_{k,k'}$ integrates to one when $s_n < T - \Delta t_{\mathsf{max}}$.

When $\Delta t_{\max} \ll T$, we can safely ignore spikes that occur at the very end of the dataset. Thus we are left with a likelihood that is proportional to a product of logistic normal densities. Since the logistic function is invertible, we may work with logit-transformed intervals instead. Then the likelihood is a product of normal densities, which is conjugate with our normal gamma prior. We can then derive the following conditional distribution:

$$\mu_{k,k'}, \tau_{k,k'} \mid \{s_n, c_n, z_n\}_{n=1}^N, \mu_\mu^0, \kappa_\mu^0, \alpha_\tau^0, \beta_\tau^0 \sim \mathcal{N}(\mu_{k,k'} | \mu_\mu, (\kappa_\mu \tau_{k,k'})^{-1}) \times \text{Gamma}(\tau_{k,k'} \mid \alpha_\tau, \beta_\tau)$$

where

$$\begin{aligned}
\mu_\mu &= \frac{\kappa_\mu^0 \mu_\mu^0 + m\bar{x}}{\kappa_\mu^0 + m}, \\
\kappa_\mu &= \kappa_\mu^0 + m, \\
\alpha_\tau &= \alpha_\tau^0 + \frac{m}{2}, \\
x_{n,n'} &= \ln(s_{n'} - s_n) - \ln(t_{max} - (s_{n'} - s_n)), \\
m &= \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k} \delta_{c_{n'},k'} \delta_{z_{n'},n}, \\
\bar{x} = \bar{x} &= \frac{1}{m} \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k} \delta_{c_{n'},k'} \delta_{z_{n'},n} x_{n,n'} \\
\beta_\tau &= \beta_\tau^0 + \frac{1}{2} \sum_{n=1}^N \sum_{n'=1}^N \delta_{c_n,k} \delta_{c_{n'},k'} \delta_{z_{n'},n} (x_{n,n'} - \bar{x})^2 \\
&\quad + \frac{\kappa_\mu^0 m (\bar{x} - \mu_\mu^0)^2}{2(\kappa_\mu^0 + m)}.
\end{aligned}$$

A.2. Log Gaussian Cox Process background rates

In the Trades on the S&P100 and the Gangs of Chicago datasets, it was crucial to model the background fluctuations that were shared among all processes. However, if the background rate is allowed to vary at time scales shorter than $\Delta t_{\max}$ then it may obscure interactions between processes. To prevent this, we sample the Log Gaussian Cox Process (LGCP) at a sparse grid of $M+1$ equally spaced points and linearly interpolate to evaluate the background rate at the exact time of each event. We have,

$$\boldsymbol{y} = \left\{ \hat{y}\left(\frac{mT}{M}\right) \right\}_{m=0}^M \sim \mathcal{GP}(\mathbf{0}, K(t,t')).$$

Then,

$$\left\{ \hat{\lambda}_{0,k}\left(\frac{mT}{M}\right) \right\}_{m=0}^M = \mu_k + \alpha_k \exp\left\{ \hat{y}\left(\frac{mT}{M}\right) \right\},$$

and $\lambda_{0,k}(s_n)$ is linearly interpolated between the rate at surrounding grid points.

The equally spaced grid allows us to calculate the integral using the trapezoid quadrature rule. We use Elliptical Slice Sampling (Murray et al., 2010) to sample the conditional distribution of the vector $\boldsymbol{y}$.

Kernel parameters are set empirically or with prior knowledge. For example, the period of the kernel is set to one day for the S&P100 dataset and one year for the Gangs of Chicago dataset since these are well-known trends. The scale and offset parameters have log Normal priors set such that the maximum and minimum homogeneous event counts in the training data are within two standard deviations of the expected value under the LGCP background rate. That is, the background rate should be able to explain all of the data without any observations if there is no evidence for interactions.

A.3. Priors on hyperparameters

When possible, we sample the parameters of the prior distributions. For example, in the Erdős-Renyi graph model we place a $\text{Beta}(1,1)$ prior on the sparsity ρ. For the latent distance model, we place a log normal prior on the characteristic length scale τ and sample it using Hamiltonian Monte Carlo.

For all of the results in this paper, we fixed the prior on the interaction kernel, $g(\Delta t)$ to a weak Normal-Gamma distribution with parameters $\mu_\mu^0 = -1.0$, $\kappa_\mu^0 = 10$, $\alpha_\tau^0 = 10$, and $\beta_\tau^0 = 1$.

Scale of gamma prior on weights. For real data, we place an uninformative prior on the weight distribution. The gamma distribution is parameterized by a shape α_W^0 and an inverse scale or rate β_W^0. The shape parameter α_W^0 is chosen by hand (typically we use $\alpha_W^0 = 2$), but the inverse scale parameter β_W^0 is sampled. We may not know a proper scale a priori, however we can use a scale-invariant Jeffrey's prior to infer this parameter as well. Jeffrey's prior is proportional to the square root of the Fisher information, which for the gamma distribution is

$$p(\beta_W^0) \propto \sqrt{I(\beta_W^0)} = \frac{\sqrt{\alpha_W^0}}{\beta_W^0}.$$

Hence the posterior is

$$\begin{aligned}
&p(\beta_W^0 \mid \{\{W_{k,k'}\}\}) \propto \\
&\frac{\sqrt{\alpha_W^0}}{\beta_W^0} \prod_{k=1}^K \prod_{k'=1}^K \frac{(\beta_W^0)^{\alpha_W^0}}{\Gamma(\alpha_W^0)} W_{k,k'}^{\alpha_W^0 - 1} e^{-\beta_W^0 W_{k,k'}} \\
&\propto (\beta_W^0)^{K^2 \alpha_W^0 - 1} \exp\left\{ -\beta_W^0 \sum_{k=1}^K \sum_{k'=1}^K W_{k,k'} \right\}.
\end{aligned}$$

This is a gamma distribution with parameters,

$$\beta_W^0 \sim \text{Gamma}(K^2\alpha_W^0, \sum_{k=1}^{K}\sum_{k'=1}^{K} W_{k,k'}).$$

B. Synthetic test details

We generated $T = 1000$s of events for each synthetic network. The average number of spikes was 25,732 $\pm$ 9,425. Network 6, the only network for which the GLM outperformed the network Hawkes model in the event-prediction test, was an outlier with 44,973 events. For event prediction, we trained on the first 900 seconds and tested on the last 100 seconds of the data. We ran our Markov chain for 2500 iterations and computed the posterior probabilities of $\boldsymbol{A}$ and $\boldsymbol{W}$ using the last 500 samples.

A simple alternative to the Hawkes model is to look at cross-correlation between the event times. First, the event times are binned into an array $\hat{\boldsymbol{s}}_k$ of length M. Let $(\hat{\boldsymbol{s}}_k \star \hat{\boldsymbol{s}}_{k'})[m]$ be the cross-correlation between $\hat{\boldsymbol{s}}_k$ and $\hat{\boldsymbol{s}}_{k'}$ at discrete time lag m. Then, $W_{k,k'} = \sum_{m=0}^{\Delta t_{\text{max}} M/T} (\hat{\boldsymbol{s}}_k \star \hat{\boldsymbol{s}}_{k'})[m]$ provides a simple measure of directed, excitatory interaction that can be thresholded to perform link prediction.

Additionally, we compare the network Hawkes process to the generalized linear model for point processes, a popular model in computational neuroscience (Paninski, 2004). Here, the event counts are modeled as $\hat{s}_{k,m} \sim \text{Poisson}(\lambda_{k,m})$. The mean depends on external covariates and other events according to

$$\lambda_{k,m} = \exp\left\{\boldsymbol{\alpha}_k^T \boldsymbol{y}_m + \sum_{k'=1}^{K}\sum_{b=1}^{B} \beta_{k,k',b}(g_b * \hat{s}_{k'})[m]\right\},$$

where $\boldsymbol{y}_m$ is an external covariate at time m, $\{g_b(\Delta m)\}_{b=1}^B$ are a set of basis functions that model impulse responses, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are parameters to be inferred. Under this formulation the log-likelihood of the events is concave function of the parameters and is easily maximized. Unlike the Hawkes process, however, this model allows for inhibitory interactions.

For link prediction, $\sum_b \beta_{k,k',b}$ provides a measure of directed excitatory interaction that can be used to compute an ROC curve. In our comparisons, we used $\boldsymbol{y}_m \equiv 1$ to allow for time-homogeneous background activity and set $\{g_b(\Delta m)\}$ to the top $B = 6$ principal components of a set of logistic normal impulse responses randomly sampled from the Hawkes prior.

We used an L1 penalty to promote sparsity in the parameters of the GLM, and chosen the penalty using cross validation on the last 100 seconds of the training data.

Model	Relative prediction improvement
Network Hawkes	100%
Standard Hawkes	59.2$\pm$14.2%
GLM	71.6$\pm$9.2%

Figure 1: Relative improvement in predictive log likelihood over a homogeneous Poisson process baseline. Relative to the network Hawkes, the standard Hawkes and the GLM yield significantly less predictive power.

Figure 4 of the main text shows the predictive log likelihoods for the Hawkes model with the correct Erdös-Renyi prior, the standard Hawkes model with a complete graph of interactions, and a GLM. On all but network 6, the network Hawkes model outperforms the competing models in terms of predictive log likelihood. Table 1 shows the average predictive performance across sample nextworks. The standard Hawkes and the GLM provide only 59.2% and 71.6%, respectively, of this predictive power.

C. Trades on the S&P100 model details

We study the trades on the S&P 100 index collected at 1s intervals during the week of Sep. 28 through Oct. 2, 2009. We group both positive and negative changes in price into the same process in order to measure overall activity. Another alternative would be to generate an "uptick" and a "downtick" process for each stock. We ignored trades outside regular trading hours because they tend to be outliers with widely varying prices. Since we are interested in short term interactions, we chose $\Delta t_{\text{max}} = 60$s. This also limits the number of potential event parents. If we were interested in interactions over longer durations, we would have to threshold the price changes at a higher level. We precluded self-excitation for this dataset since upticks are often followed by downticks and vice-versa. We are seeking to explain these brief price jumps using the activity of other stocks.

We run our Markov chain for 2000 iterations and compute predictive log likelihoods and the eigenvalues of the expected interaction matrix, $\mathbb{E}[\boldsymbol{A} \odot \boldsymbol{W}]$, using the last 400 iterations of the chain. The posterior sample illustrated in the main text is the last sample of the chain.

Trading volume varies substantially over the course of the day, with peaks at the opening and closing of the market. This daily variation is incorporated into the background rate via a Log Gaussian Cox Process with a periodic kernel. We set the period to one day. Figure 2 shows the posterior distribution over the background rate.

Though it is not discussed in the main text, we also considered stochastic block model (SBM) priors as well (Hoff, 2008), in hopes of recovering latent sector affiliations based

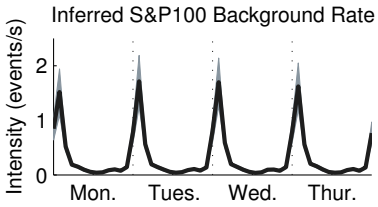

Figure 2: Posterior distribution over shared background rates for the S&P100. Shading indicates two standard deviations from the mean.

on patterns of interaction between sectors. For example, stocks in the financial sector may have 90% probability of interacting with one another, and 30% probability of interacting with stocks in the energy sector. Rather than trying to interpret this from the embedding of a latent distance model, we can capture this belief explicitly with a stochastic block model prior on connectivity. We suppose there are J sectors, and the probability of belonging to a given sector is $\boldsymbol{\alpha} \in [0,1]^J \sim \text{Dirichlet}(\boldsymbol{\alpha}_0)$. The latent sector assignments are represented by the vector $\boldsymbol{b} \in [1,J]^K$, where $b_k \sim \text{Cat}(\boldsymbol{\alpha})$. The probability of a directed interaction is $\Pr(A_{k,k'} = 1) = B_{b_k,b_{k'}}$, where $\boldsymbol{B}$ is a $J \times J$ matrix of Bernoulli probabilities. We place a beta prior on the entries of $\boldsymbol{B}$.

Our experiments with the SBM prior yield comparable predictive performance to the latent distance prior, as shown in Figure 3. The inferred clusters (not shown) are correlated with the clusters identified by Bloomberg.com, but more analysis is needed. It would also be interesting to study the difference in inferred interactions under the various graph models; this is left for future work.

Financial Model	Pred. log lkhd. (bits/spike)
Indep. LGCP	0.594
Std. Hawkes	0.912
Net. Hawkes (Erdős-Renyi)	0.903
Net. Hawkes (Latent Distance)	0.888
Net. Hawkes (SBM)	0.894

Figure 3: Comparison of financial models on a event prediction task, relative to a homogeneous Poisson process baseline.

D. Gangs of Chicago model details

The first 12 years are used for training, 1993 is reserved for cross-validation, and the remaining two years are used to test the predictive power of the models. We also considered the crime dataset from `www.data.cityofchicago.org`, but this does not identify gang-related incidents.

We run our Markov chain for 700 iterations and use the last 200 iterations to compute predictive likelihoods and expectations. The posterior sample illustrated in the figure in main text is the last sample of the chain.

Since this is a spatiotemporal dataset, our intensities are functions of both spatial location and time. For simplicity we factorize the intensity into $\lambda_{k,x}(\boldsymbol{x})\lambda_{k,t}(t)$, where $\lambda_{k,t}(t)$ is a Gaussian process as described above, and $\lambda_{k,x}(\boldsymbol{x})$ is uniformly distributed over the spatial region associated with process k and is normalized such that it integrates to 1.

In the case of the latent distance model with the community process model, each community's location is fixed to its center of mass. With the cluster process model, we introduce a latent location for each cluster and use a Gaussian distribution for the prior probability that a community belongs to a cluster. This encourages spatially localized clusters.

Figure 4 shows the cross validation results used to select the number of clusters, K, in the clustered process identity model and each of the four graph models. For the empty, complete, and Erdös-Renyi graph priors, we discover $K = 15$, 4, and 4 clusters respectively. The latent distance model, with its prior for spatially localized clusters, has its best performance for $K = 5$ clusters.

The spatial GMM process ID model from Cho et al. (2013) fails on this dataset because it assigns its spatial intensity over all of $\mathbb{R}^2$, whereas the clustering model concentrates the rate on only the communities in which the data resides. Figure 5 shows the results of this spatial process ID model on the prediction task. We did not test a latent distance model with the spatial GMM, but it would likely suffer in the same way as the empty, complete, and Erdős-Renyi graph priors.

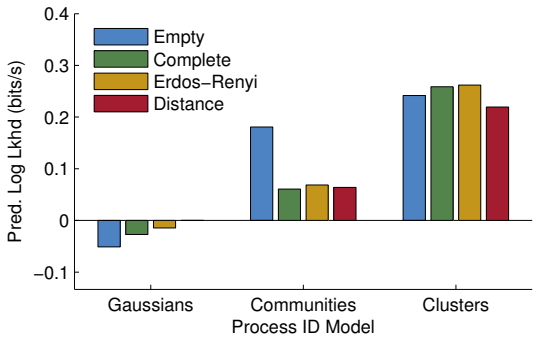

Figure 5: Comparison of predictive log likelihoods for Chicago homicides. This is the same as Figure 7a of the main text, but also includes the spatial GMM process identity model.

References

Cho, Yoon Sik, Galstyan, Aram, Brantingham, Jeff, and Tita, George. Latent point process models for spatial-temporal networks. *arXiv:1302.2671*, 2013.

Hoff, Peter D. Modeling homophily and stochastic equiv-

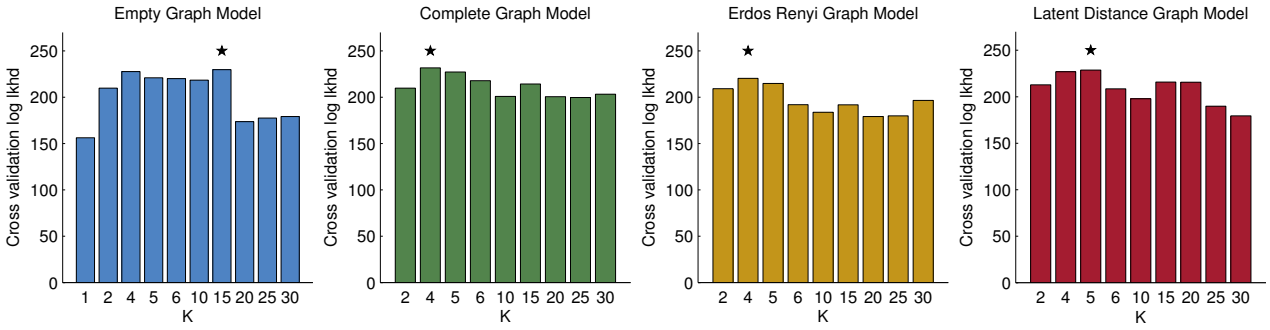

Figure 4: Cross validation results for Chicago models with K clusters for each of the four graph models.

alence in symmetric relational data. *Advances in Neural Information Processing Systems 20*, 20:1–8, 2008.

Murray, Iain, Adams, Ryan P., and MacKay, David J.C. Elliptical slice sampling. *Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS)*, 9:541–548, 2010.

Paninski, Liam. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, January 2004.