Modeling the Influence Structure of a Network with Hawkes Processes

Steven Morse [1,2,3,4], Phil Chodrow [2,3], Marta González [2,3], Natasha Markuzon [4]

[1]Department of Mathematical Sciences, U.S. Military Academy
[2]Operations Research Center, MIT
[3]Human Mobility and Networks Lab, MIT
[4]Draper Laboratory, Cambridge

January 12, 2018

Agenda

- Introduction
  Problem statement, approach and contributions, data
- Methods: Hawkes Processes
  Model definition, Bayesian EM parameter estimation
- Results
  Univariate case, multivariate case
- Conclusion and future research

Problem Statement

Central question

How can we *identify* and *model* the structure of influence in a communication network?

- We are interested in who influences who, not necessarily just who contacts who, as measured through observation of interpersonal communication.
- For example, if A talks to B, does that increase the probability that B will talk to C?
- How can we identify and model these relationships from large-scale communication data?
- Applications: diffusion, influence maximization, social learning, centrality, ...

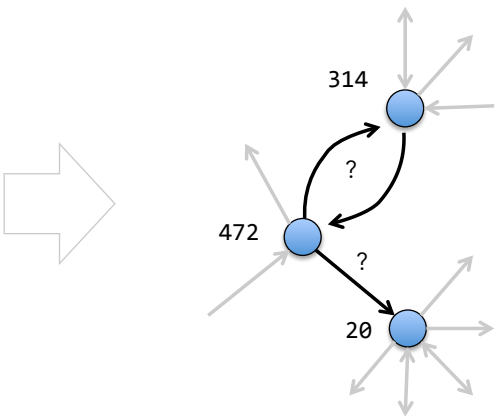
The Challenge

Large-scale communication metadata gives us little/no knowledge of content (e.g. mobile phone data) — so how do we determine which events represent "meaningful" interactions?

Approaches we might try…

- Edge-weighting
- Observing post-emergency
- Frequent subgraphs
- **Learning probabilistic structure**

Caller	*Callee*	*Time*
472	314	3970084
30	407	3970085
772	896	3970088
703	705	3970088
	...	
314	472	3982191
27	33	3982392
472	20	3983092
	...	
314	472	3411193
425	282	3411194
101	99	3411296
233	881	3413297



Background

- Based on previous findings, we would like a model which can capture temporal clustering, information cascades, and an interpretable influence structure.
- We adopt a multidimensional stochastic process called the *Hawkes process*, a flexible model with all these traits.
- The Hawkes process also has applications in a wide variety of fields:
  - Stock price fluctuation (Hawkes [1])
  - Earthquake activity (Veen & Schoenberg [8])
  - Gang violence (Stomakhin, Bertozzi et al. [6])
  - Neuron impulses in the brain (Linderman & Adams [3])
  - Social networks (Zipkin et al. [11])
  - Trend detection (Pinto et al. [5])
  - Product adoption (Valera, Gomez-Rodriguez [7])

  which gives us a rich literature of techniques to draw on.

Definitions I

- The Hawkes process (HP) is a point process defined by a conditional intensity with a background rate and additive, decaying impulses from previous events. It can be extended to multiple dimensions (or streams).

Hawkes process

A sequence of events $\tau = \{(t_i, u_i)\}_{i=1}^{n}$, consisting of a time t_i and dimension u_i, with $t_i \in \mathbb{R}^+$ and $u_i \in \mathcal{U} = \{1, 2, ..., U\}$, is a *Hawkes process* if the conditional intensity function has the form

$$\lambda_u(t; \Theta) = \mu_u + \sum_{i:t_i<t} h_{uu_i}(t - t_i; \theta_{uu_i})$$

where $\Theta = (\mu, \theta)$ are the model parameters, and $H = [h_{ij}]$, $h_*(t) : \mathbb{R}^+ \to \mathbb{R}^+$ is the matrix of *triggering kernels* varying with u and u_i.

Definitions II

- The triggering kernel controls how much previous events affect the probability of future events occurring, and should decay with time. We separate the kernel into an influence term and a decay term.

Exponential triggering kernel

Decompose $H = [h_{ij}]$ into an *influence matrix* $A = [\alpha_{ij}]$ and *exponential triggering kernel* $G(t) = [g_{ij}(t)]$, such that $H = A \odot G$ and

$$h_{uu'}(t) = \alpha_{uu'} g(t), \quad g(t) = \omega e^{-\omega t}$$

where we have defined a global (hyper)parameter ω, $\forall u, u'$.

Univariate example

- Consider the univariate case, when $U = 1$. We get a sense for the temporal clustering inherent to the process by comparing it to a Poisson process with the same background rate. Recall:

$$\lambda(t) = \mu + \sum_{t_i < t} \alpha\omega e^{-\omega(t-t_i)}$$

- Depicted below are the arrivals for a Hawkes (blue dots) and Poisson (yellow dots) process over a period $t \in [0, 500]$. The time-varying intensity of the Hawkes process is shown above, where we note the spikes accompanying each new arrival, and resulting "sawtooth" patterns.

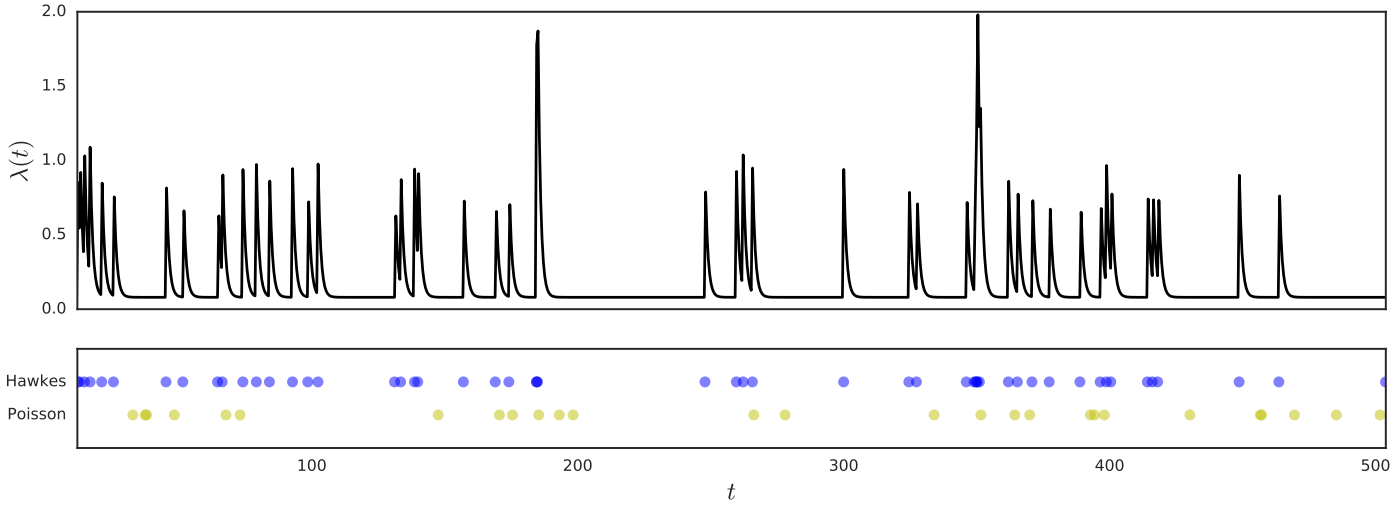
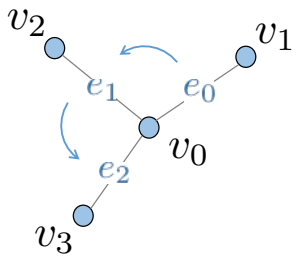


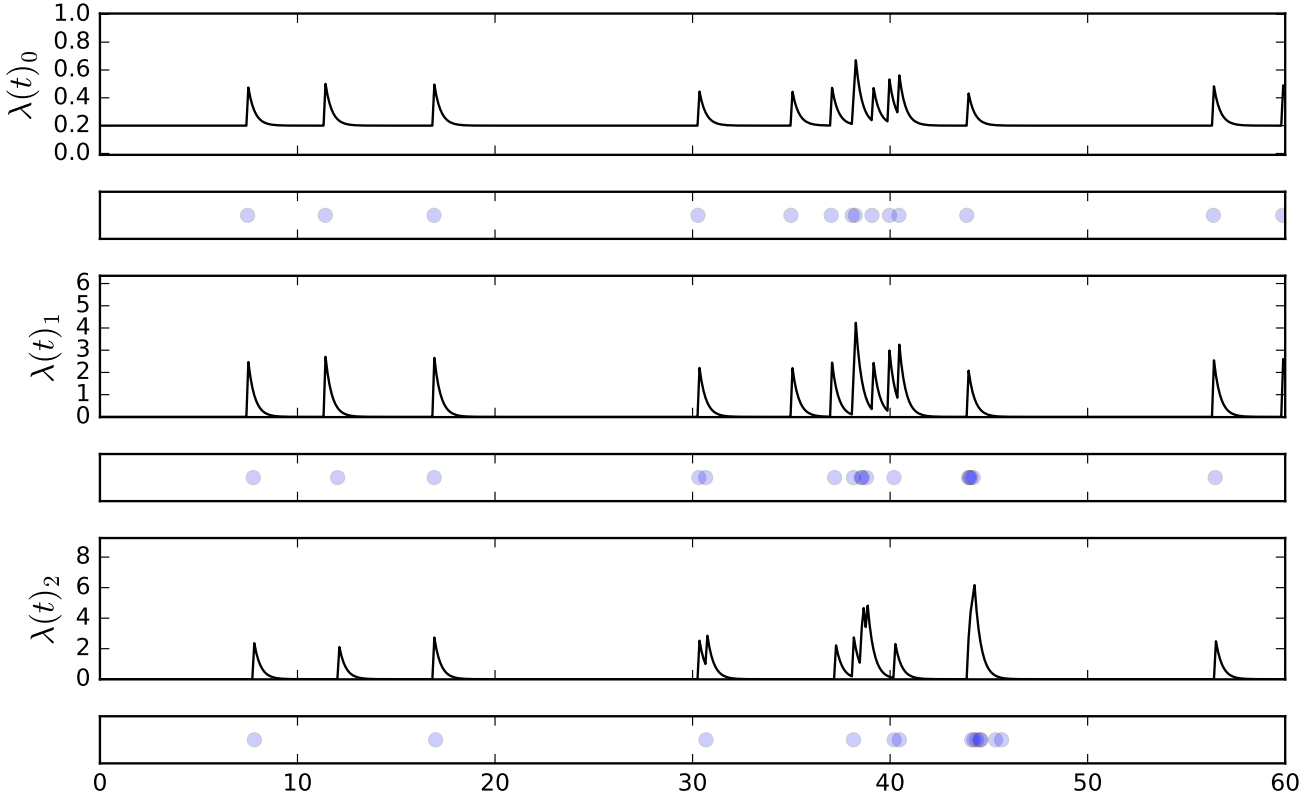
Figure 1: Poisson process has rate $\mu = 0.1$. Hawkes has $\mu = 0.1$, $\alpha = 0.5$, $\omega = 1$.

Multivariate example



$$A = \begin{pmatrix} 0.5 & 0 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \end{pmatrix}$$

$$\mu = \begin{pmatrix} 0.2 & 0 & 0 \end{pmatrix}$$



Branching process and stability

- We can interpret the Hawkes process as a branching process where each event (t_i, u_i) in the sequence is either a *parent*, or *child* of a previous event.
- Under this interpretation, one can show that the entries of $A = [\alpha_{ij}]$ give the *expected children* of parent j into dimension i.
- Stability. For $U = 1$ therefore, if $\alpha > 1$, each event produces more than one child in expectation, and the process is called unstable or nonstationary (it "blows up"). So we generally require $\alpha < 1$.
- For $U > 1$ this has the analogy to constraining the largest eigenvalue of A. We require

$$\rho(A) \stackrel{\text{def}}{=} \max_i \ |\lambda_i(A)| < 1$$

to ensure stability, where $\lambda_i(A)$ here represents the eigenvalues of A.

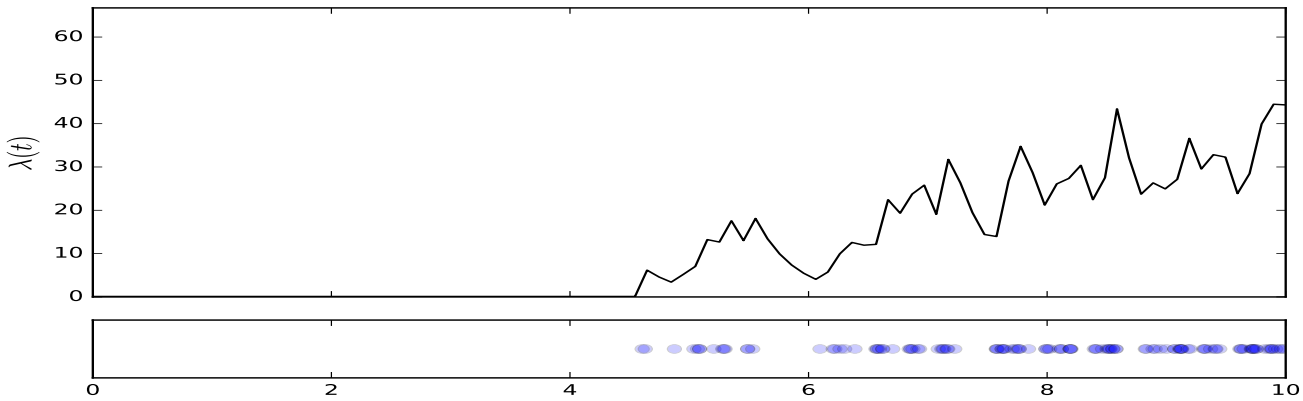


Figure 2: Example unstable process ($U = 1$), with $\mu = 0.1$, $\alpha = 1.1$, $\omega = 3$.

Parameter estimation: Challenges in MLE

- The log-likelihood $\mathcal{L}(\{(t_i, u_i)\}, \Theta)$ has a closed form, and we might like to do MLE as

$$\min_{\Theta} -\mathcal{L}(\tau, \Theta) + \mathcal{R}(\Theta)$$

  with regularization terms $\mathcal{R}(\cdot)$.
- In practice, this function tends to be very "flat" near the optimum, leading to near-zero gradients, degenerate Hessians, and slow or no convergence.
- With strict regularization on A (for example L_1 in [9], L_2 in [7], or L_* in [10]), we can apply sophisticated optimization machinery (for example ADMM with an MM-step in [10]) and achieve a sparse parameter estimate.

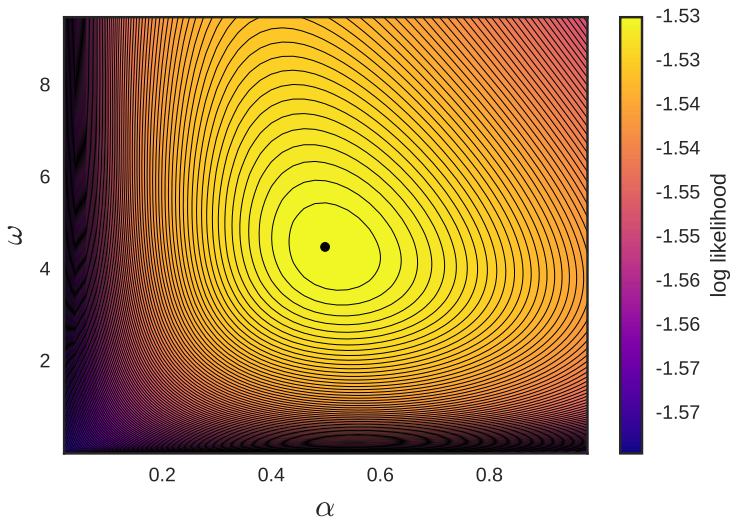


Figure 3: Log-likelihood with $U = 1$, μ fixed. Optimum shown as black dot.

Parameter estimation: Bayesian MAP EM

- Instead, we propose a novel method using Expectation-Maximization (EM). In particular, we will use a Bayesian maximum aposteriori (MAP) EM approach that maximizes the complete data log posterior

$$\log p(\Theta|\tau, Q) \propto \log p(\tau, Q|\Theta) + \log p(\Theta)$$

  where $\tau = \{(t_i, u_i)\}$ and $Q = [q_{ij}]$ is the latent branching matrix such that $q_{ij} = 1$ if j is the parent event of i.
- Since Q is unknown, the EM approach finds its expected value $P = [p_{ij}]$ based on the current parameter estimate, maximizes the posterior with this expected value for new estimates, and iterates until convergence.
- Formally, in the E-step we compute

$$P^{(k+1)} = \mathbb{E}[Q|\tau, \Theta^{(k)}]$$

  and in the M-step we compute

$$\Theta^{(k+1)} = \text{argmax} \left\{ \mathbb{E}[\log p(\tau, Q; \Theta) \mid Q = P^{k+1}] + \mathbb{E}[\log p(\Theta; V)] \right\}$$

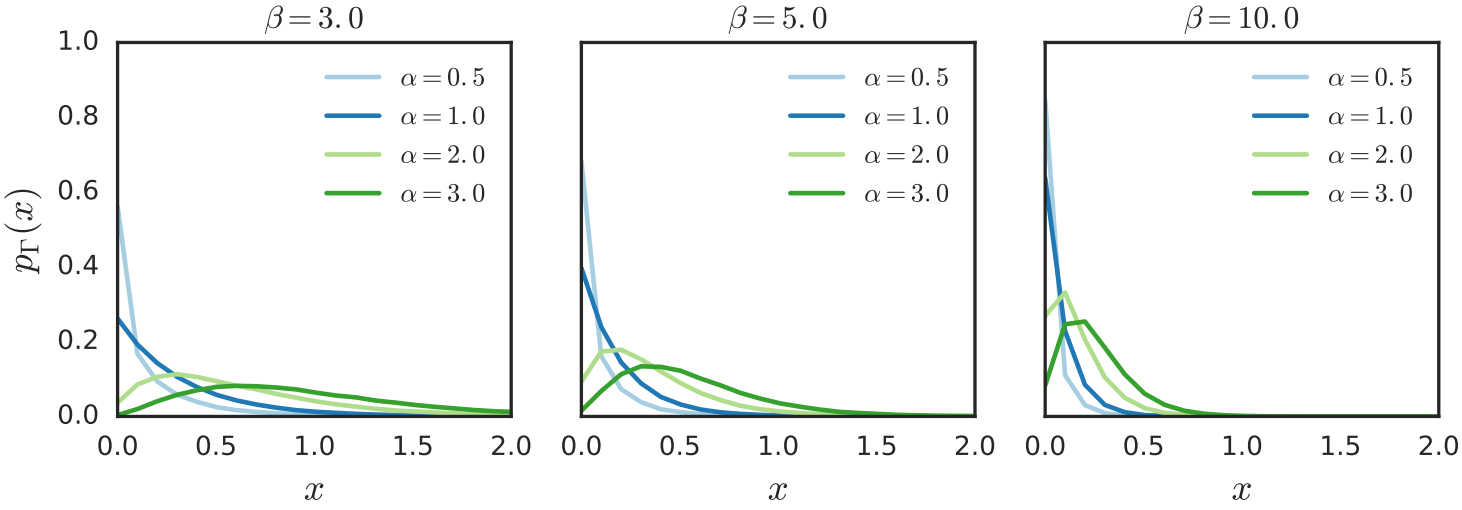
Parameter estimation: Prior

- We place a Gamma prior on the influence matrix entries $A = [\alpha_{ij}]$. This has nonnegative support and is conjugate with the exponential terms in the complete data log likelihood.

$$\text{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

- We assume that each relationship (i, j) is iid, so we have the prior

$$p(A) = \prod_{i,j} \text{Gamma}(\alpha_{ij};\ s_{ij}, t_{ij}) \quad \Rightarrow \quad \log p(A) = \sum_{i,j} \left[(s_{ij} - 1) \log \alpha_{ij} - t_{ij} \alpha_{ij} \right] + C$$



Parameter estimation: Summary

E-step

Compute $P^{(k+1)} = \mathbb{E}[Q|\tau, \Theta^k]$ as

$$p_{ii}^{(k+1)} = \frac{\mu_{u_i}^{(k)}}{\mu_i^{(k)} + \sum_{j=1}^{i-1} \alpha_{u_i u_j}^{(k)} g(t_i - t_j)}, \qquad p_{ij}^{(k+1)} = \frac{\alpha_{u_i u_j}^{(k)} g(t_i - t_j)}{\mu_i^{(k)} + \sum_{j=1}^{i-1} \alpha_{u_i u_j}^{(k)} g(t_i - t_j)}$$

M-step

Compute $\Theta^{(k+1)} = (\mu^{(k+1)}, A^{(k+1)})$, with $G(t) = \int_0^t g(s)\mathrm{d}s$, as

$$\mu_u^{(k+1)} = \frac{\sum_{i:u_i=u} p_{ii}^{(k)}}{T}$$

$$\alpha_{uu'}^{(k+1)} = \frac{\sum_{i:u_i=u} \sum_{j:u_j=u', j<i} p_{ij}^{(k)} + s_{uu'} - 1}{\sum_{i=1}^{N} \sum_{j:\ u_j=u', j<i} G(T - t_j) + t_{uu'}}$$

Incorporating kernel updates (univariate case)

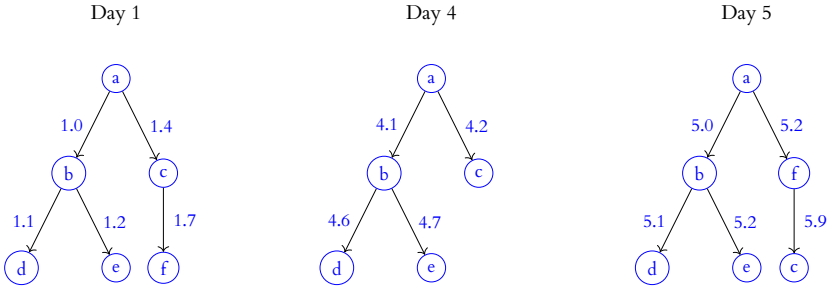
- When $U = 1$, we may prefer to estimate ω along with μ and α (instead of treating it as a hyperparameter). We will again apply a Gamma prior to ω, with hyperparameters (u, v).
- This gives an additional update equation for ω as:

$$\omega^{k+1} = \frac{\sum_{j<i} P_{ij}^{k+1} + u - 1}{\sum_{j<i} P_{ij}^{k+1}(t_i - t_j) + v} .$$

with μ and α identical to the multivariate case.

Univariate case: modeling group conversations

- Extract recurring group conversations ("persistent cascades") from data using methodology in [4].
- Are these recurrent group conversations well-modeled by this self-exciting point process?
- We first collapse the conversations into a single sequence of events. As a short example, consider the following recurrent conversation pattern:

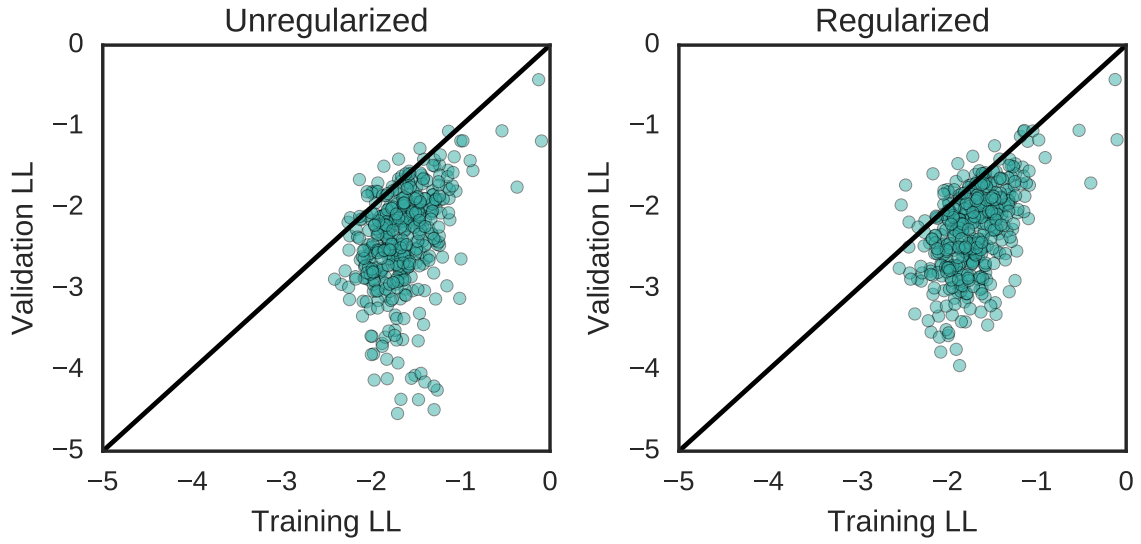


and the corresponding sequence of events:

$$\tau = \{1.0, 1.1, 1.2, 1.3, 1.4, 1.7, 4.1, ..., 5.2, 5.9\}.$$

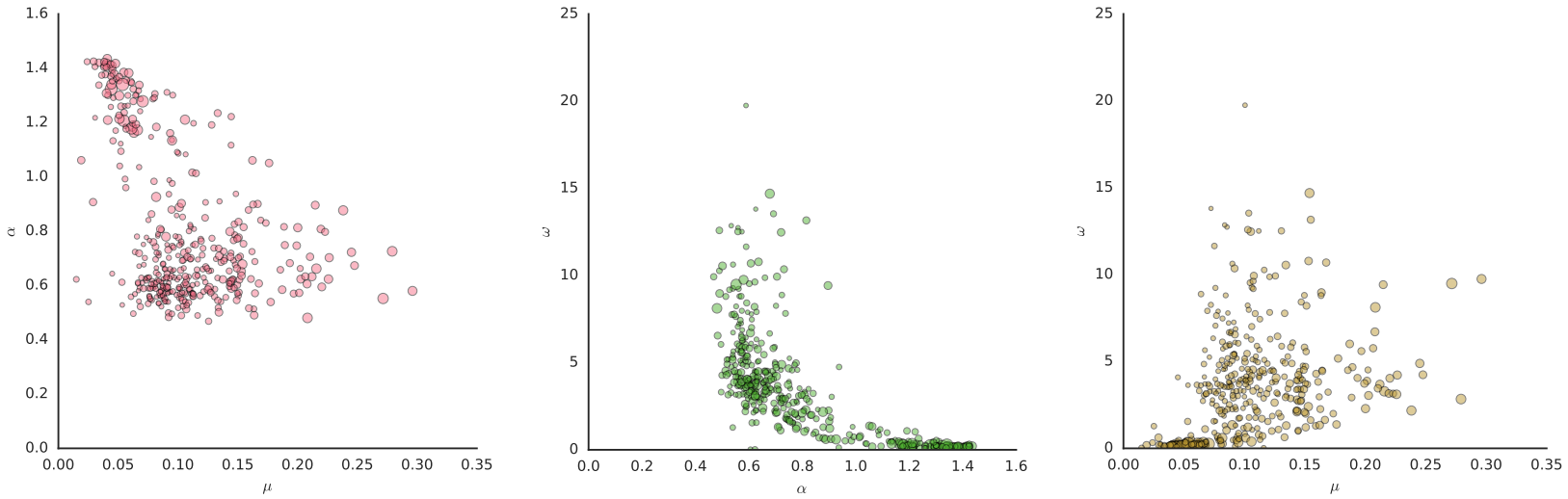
Univariate case: effect of regularization

- Regularization increases out-of-sample predictive performance.
- Shown are scatterplots of the training log-likelihood (horizontal axis) and validation log-likelihood (vertical axis) for unregularized (left) parameter estimates and optimal regularized (right) estimates found through grid-search on the validation set.
- Introducing regularization leads to higher validation likelihoods and stronger correlation between training and validation scores.



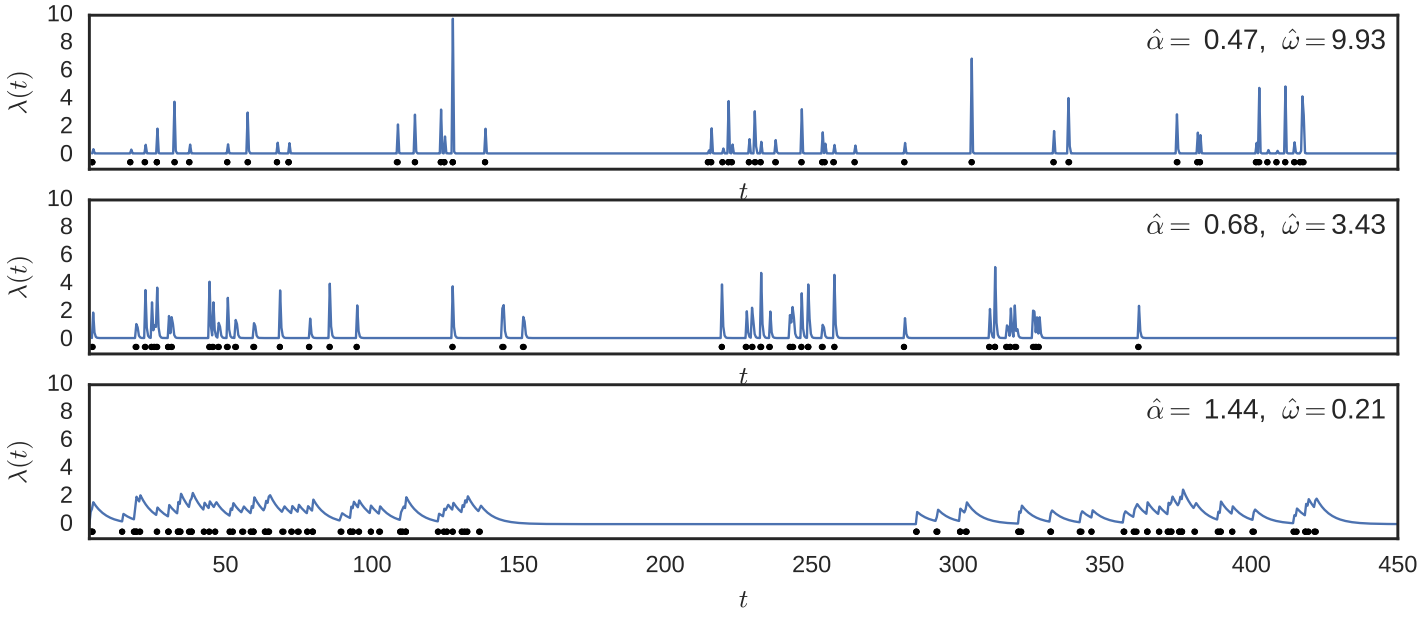
Univariate case: group conversation types

- Parameter estimates reveal two types of recurring group conversations.
- Shown are scatterplots of parameter estimates for μ, α, ω under regularization. Size of dot indicates size of the sequence.
- Two distinct clusters of recurring conversation type are evident:
  - Group 1 — Low background activity (μ) but high self-excitation (α) and slow decay (ω).
  - Group 2 — High background activity, moderate self-excitation, fast decay.



Univariate case: examples

- Depicted are the process events (black dots) and estimated intensities using estimated parameters for sequences with low (top), median (middle) and high (bottom) estimated branching ratios $\hat{\alpha}$.
- The bottom sequence corresponds to the non-stationary category of conversation, its nonstationarity reflected in the fact that the intensity is almost never at its baseline value.
- We see, as expected, the top two sequences are characterized by frequent, small bursts of activity, while the bottom sequence is characterized by long periods of dense activity.



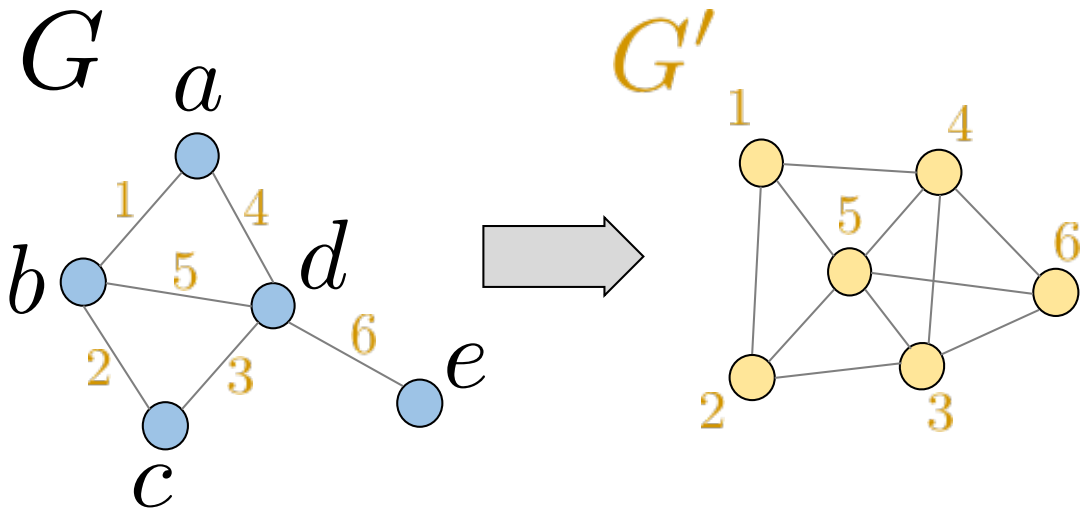
Multivariate case: Modeling the Influence Network

- We now consider $U > 1$. Typically in network applications of this model, each u corresponds to a *node* (e.g. stocks, neurons, gangs).
- However since we are measuring influence through interpersonal communication, we prefer to represent each undirected edge, or dyad in the network as a dimension.
- We call this the Dyadic Network Hawkes model. It is straightforward to move from one model to the other.
- Map each edge in G to a node in G', and let two nodes in G' be connected if the corresponding edges in G share an endpoint. G' is called the line graph of G. We can compute its adjacency matrix

$$A(G') = B(G)B(G)^T - 2I$$

with $B(G)$ the incidence matrix of G.

Multivariate case: example



$$B(G) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix} \qquad A(G') = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Multivariate case: priors and stability

- We construct the hyperparameters (s_{ij}, t_{ij}) for a parameter α_{ij} as

$$s_{ij} \stackrel{\text{def}}{=} a_{ij} s_0 + 1, \qquad t_{ij} \stackrel{\text{def}}{=} t_0$$

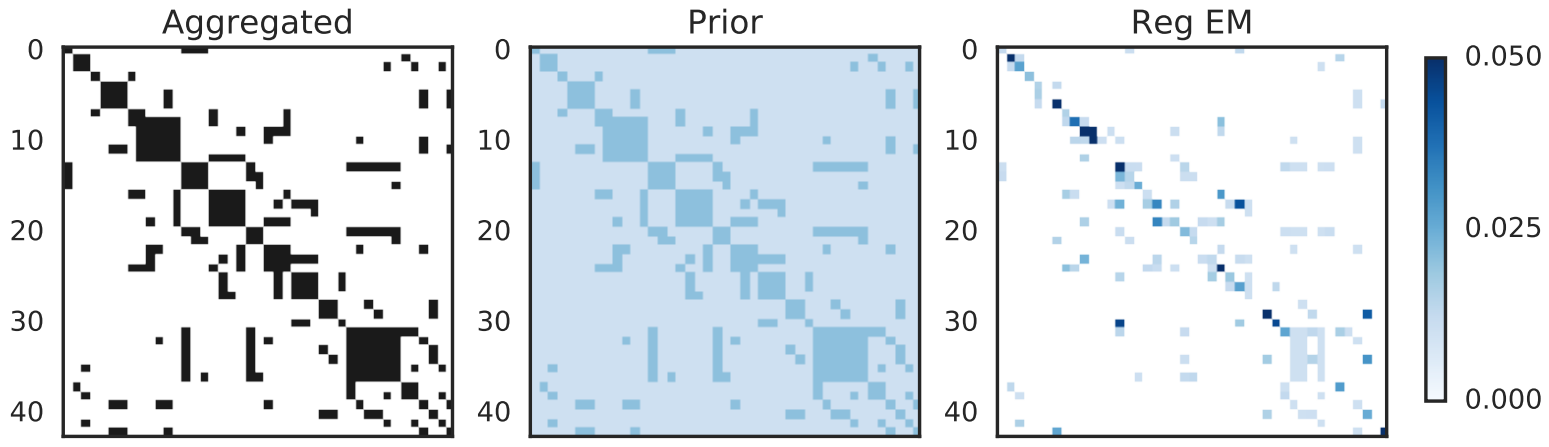
  where s_0, t_0 and a_{ij} are hyperparameters we will select with a validation set.
- In particular, we will select s_0 and ω using a standard grid-search technique, and use the line graph adjacency matrix $A(G')$ of the aggregated network in the validation data to select the a_{ij}'s.
- We set t_0 based on s_0 in order to ensure stability, a problem which is more pronounced in the multivariate case. We would like to choose a prior that places most of its mass on stable sequences. The *circular law* states that the maximum eigenvalue of a $K \times K$ stochastic matrix is distributed as $\lambda_{\max} \sim \mathcal{N}(\mu K, \sigma^2)$. So we can ensure, for example, $\mu K + 2\sigma^2 < 1$ by setting

$$t_0 > \frac{1}{2}\left(K s_0 + \sqrt{(K s_0)^2 + 8 s_0}\right)$$

  (Adapted from Linderman [2].)

Multivariate case: parameter estimation

- At **left** is the adjacency matrix of the line graph of the aggregated network based on the validation data. At **center** is the prior on the influence matrix A, with $s_0 = 5$. At **right** is the estimated parameter values for A. Color scale is the same for the middle and right plots.
- Note several critical differences where the model is detecting influential relationships over edges that do not even exist in the aggregated network. The data overwhelms the prior in these cases.



Conclusion & Future Work

- Repo: `github.com/stmorse/hawkes`

- Ties to mobility: does influence in communication patterns relate to influence in movement behavior?
- Compare predictive ability of our MHP parameter estimation approach to existing approaches
- Quantitatively test the "strength of weak ties" hypothesis by measuring the influence (α_{ij}) of community bridges
- Extension of non-Poissonian interaction to other diffusion models, e.g. does temporal clustering affect the submodularity of the influence maximization problem?

Thank you! Questions?

References I


A. G. Hawkes.
Spectra of some self-exciting and mutually exciting point processes.
Biometrika, 58(1):83–90, 1971.


S. W. Linderman and R. P. Adams.
Discovering Latent Network Structure in Point Process Data.
In *Proc. of the 31st International Conference on Machine Learning*, volume 32, 2014.


S. W. Linderman and R. P. Adams.
Scalable Bayesian Inference for Excitatory Point Process Networks.
2015.


S. Morse, M. C. González, and N. Markuzon.
Persistent Cascades : Measuring Fundamental Communication Structure in Social Networks.
In *Proc. IEEE Conference on Big Data*, pages 969–975, 2016.

References II


J. C. L. Pinto, T. Chahed, and E. Altman.
Trend detection in social networks using Hawkes processes.
Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15, pages 1441–1448, 2015.


A. Stomakhin, M. B. Short, and A. L. Bertozzi.
Reconstruction of missing data in social networks based on temporal patterns of interactions.
Inverse Problems, 27(11):115013, 2011.


I. Valera and M. Gomez-Rodriguez.
Modeling Adoption and Usage of Competing Products.
In *IEEE International Conference on Data Mining (ICDM)*, 2015.

References III


A. Veen and F. P. Schoenberg.
Estimation of Space–Time Branching Process Models in Seismology Using an EM-Type Algorithm.
Journal of the American Statistical Association, 103(482):614–624, 2008.


K. Zhou.
Extending Low-Rank Matrix Factorizations for Emerging Applications.
(December), 2013.


K. Zhou, H. Zha, and L. Song.
Learning Triggering Kernels for Multi-dimensional Hawkes Processes.
In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1301–1309, 2013.

References IV

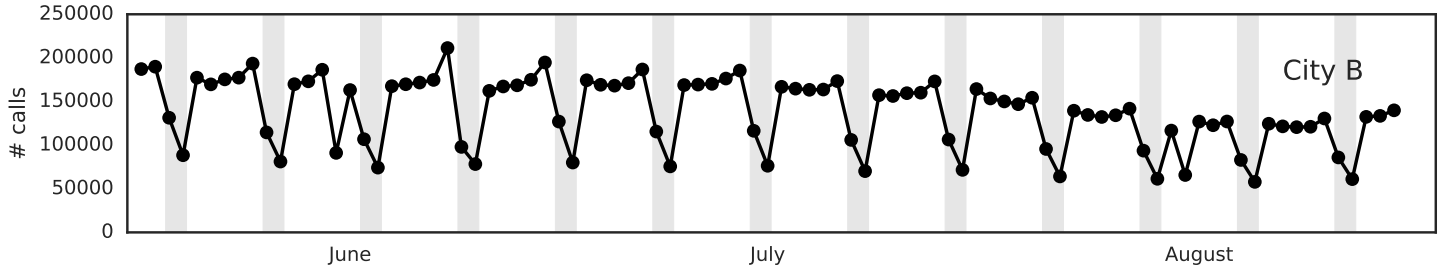

J. R. Zipkin, F. Schoenberg, K. Coronges, and A. Bertozzi.
Point-process models of social network interactions: parameter estimation and missing data recovery.
Euro Journal of Applied Mathematics, 1, 2014.

Backup

Data (summary)

- Our data consists of call detail records (CDRs) from two European cities (City "A" and "B") and one Central American city (City "C").
- Call activity follows predictable population-level patterns. For example, individuals are about half as active on weekends compared to weekdays at a population level.

City	Unique IDs ($\times 10^3$) avg. / mo.	Unique IDs ($\times 10^3$) total	Calls ($\times 10^3$) avg. /mo.	Calls ($\times 10^3$) total	# months	Degree (k), avg.	Calls /edge/mo. (w), avg.
A	331.2	648.1	6,334.6	82,350.1	13	3.88	11.59
B	258.0	523.5	4,172.2	55,747.5	13	3.62	10.52



Weekends highlighted with gray bars. The two weekday outliers correspond to national holidays.