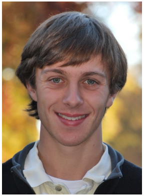

THIS PRESENTATION IS ABOUT THE QUESTION OF WHETHER ANIMAL ALTRUISM EXISTS.
I SPENT MANY HOURS DOING RESEARCH FOR THIS PRESENTATION, KNOWING FULL WELL THAT IT WOULD HAVE NO NOTICEABLE BENEFIT FOR MY CAREER.
I WAS NOT PAID TO DO IT. NO POTENTIAL MATES WILL BE IMPRESSED BY IT. I WILL GAIN NO SOCIAL STATUS FOR HAVING DONE IT. IT IS PURELY A GIFT TO THE SCIENTIFIC COMMUNITY.
THUS, THE EXISTENCE OF THIS PRESENTATION ABOUT ALTRUISM PROVES THE EXISTENCE OF ANIMAL ALTRUISM. THANKYOU.
WHAT A GREAT PRESENTATION!
STOP THAT!
smbc-comics.com

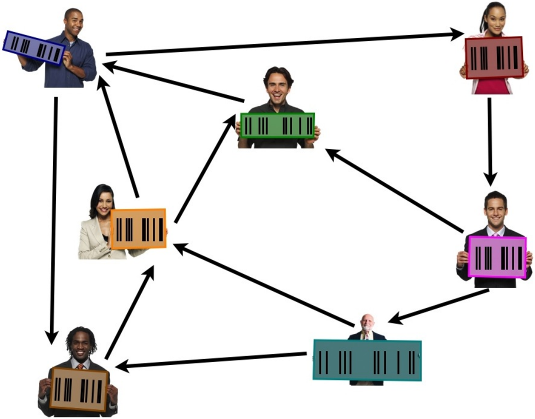
Tracking Influences within Dynamic Networks

Rebecca Willett, University of Wisconsin-Madison



Joint work
with Eric Hall

Cascading chains of interactions

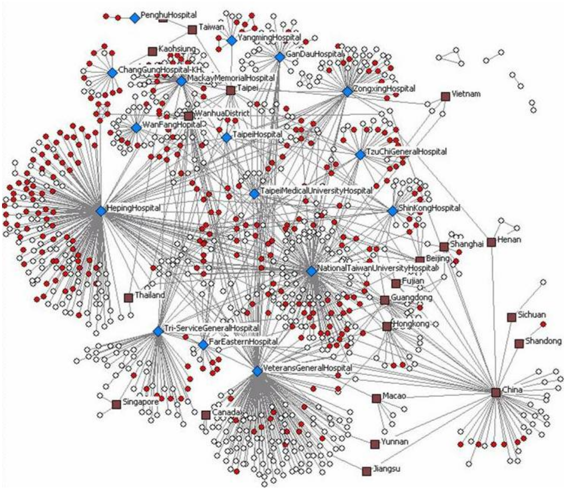


- Internet memes quickly propagate[a]
- Gang violence begets retaliations[b]
- Nation-state conflicts are accompanied by proxy wars[c]

[a] K. Zhou, H. Zha, and L. Song, 2013
[b] A. Stomakhin, M. B. Short, and A. Bertozzi, 2011
[c] C. Blundell, K. A. Heller, and J. M. Beck, 2012

Can we infer the underlying network of influences from observations of individual events?

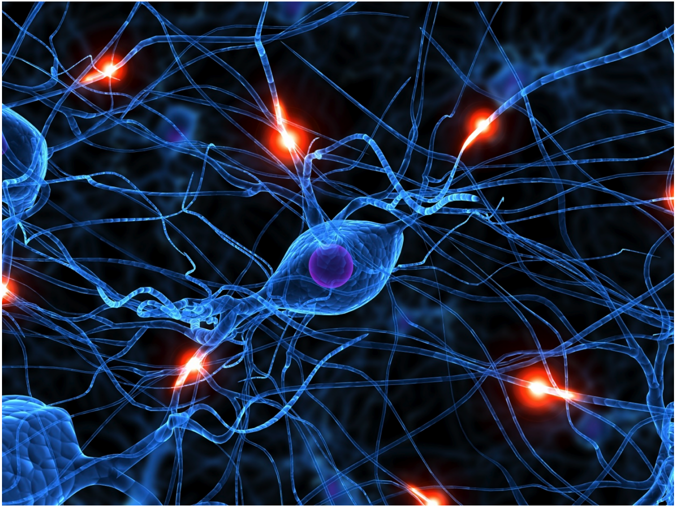
Epidemiology



Can we predict the spread of infectious disease?[1]

[1] http://ai.arizona.edu/research/bioportal/

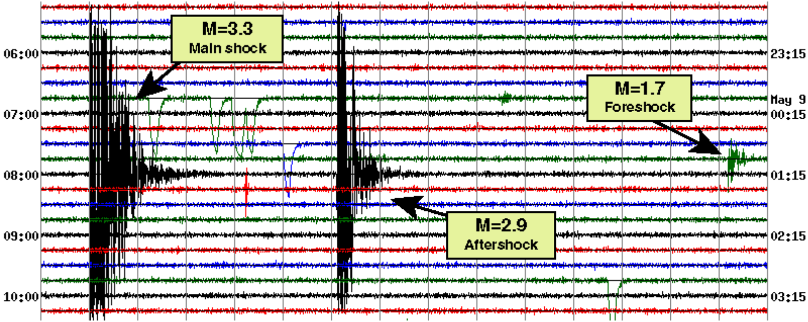
Functional neural network connectivity



We record neurons firing in response to different stimuli.

Can we track the dynamic functional network?

Seismology



We record seismic events and shocks.[2]

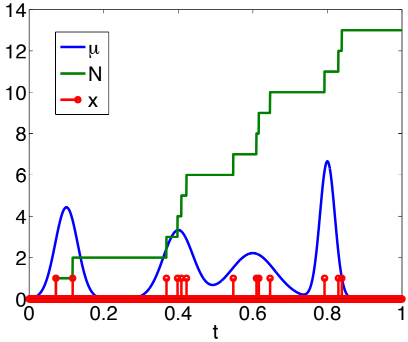
Can we infer patterns of earthquake interactions?

[2] http://earthquake.usgs.gov/monitoring/helicorders/examples/Fore_main_after.php

Point process likelihood

- For each node $k \in [\![p]\!]$, we have a point process with $N_{k,\tau}$ = the number of events up to an including time τ.
- Let $\mu_k(\tau)$ denote a time-varying rate function, so that the likelihood of node k participating in an event between times t_1 and t_2 is controlled by

$$\int_{t_1}^{t_2} \mu_k(\tau) d\tau$$



Point processes likelihood

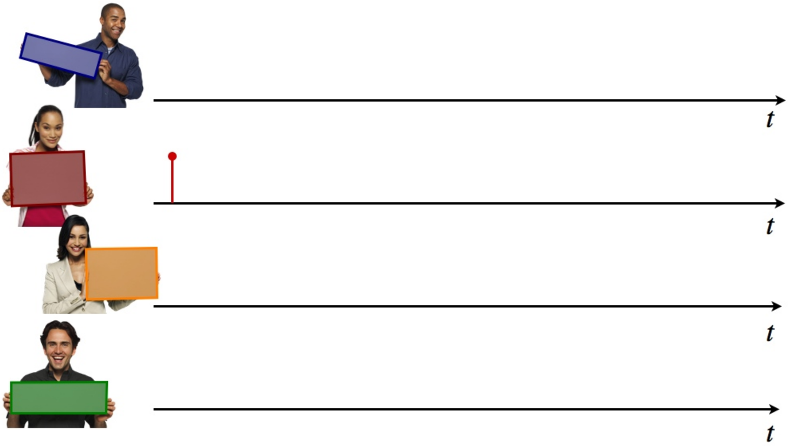
Neglecting terms independent of μ, we have

$$
\begin{aligned}
-\log(N^T|\mu) &= \sum_{k=1}^{p} \int_0^T \log \mu_k(\tau) dN_{k,\tau} - \mu_k(\tau) d\tau \\
&\approx \sum_{t=1}^{T/\delta} \langle \delta\mu_t, \mathbb{1} \rangle - \langle x_t, \log \delta\mu_t \rangle.
\end{aligned}
$$

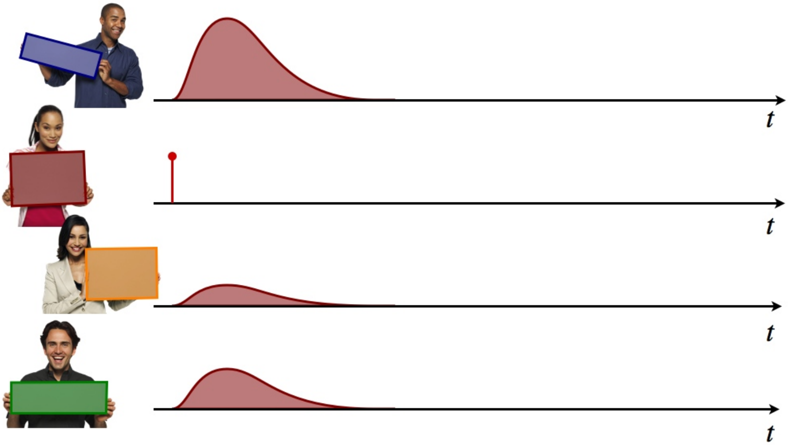
where $x_{t,k}$ is the count of events for node k in the time window $(\delta(t-1), \delta t]$.

We now need a model for μ that captures the underlying network structure...

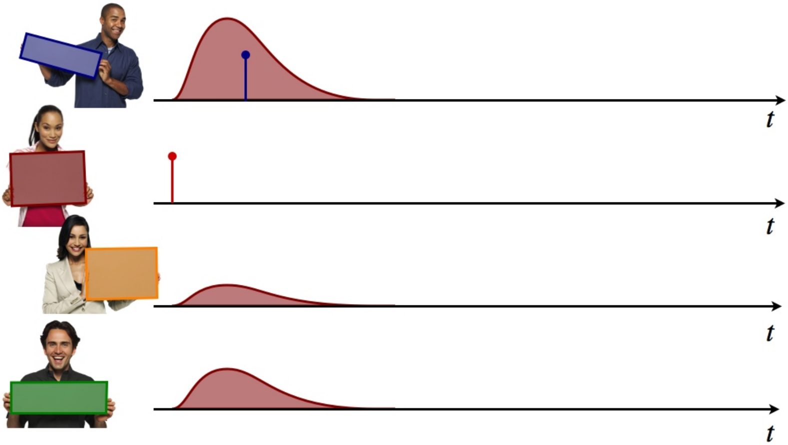
Multivariate Hawkes Processes



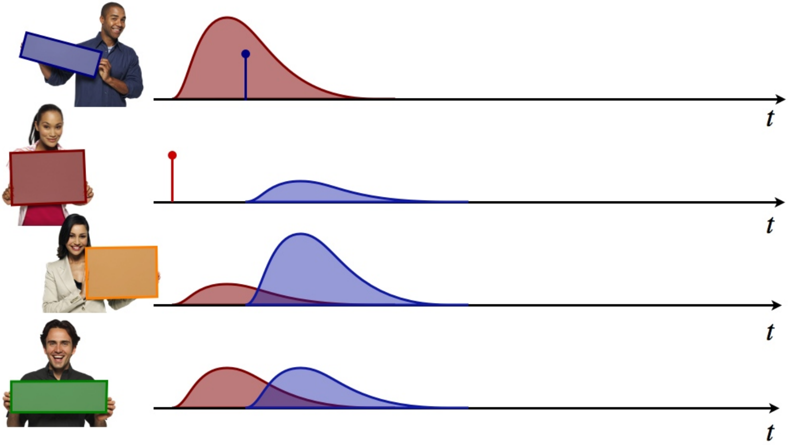
Multivariate Hawkes Processes



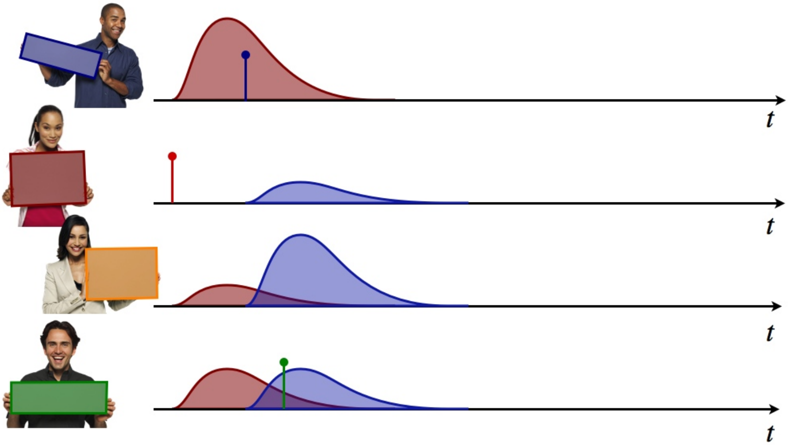
Multivariate Hawkes Processes



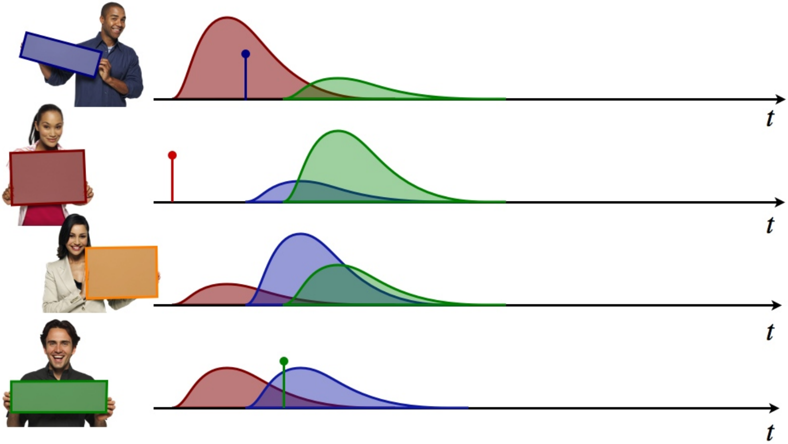
Multivariate Hawkes Processes



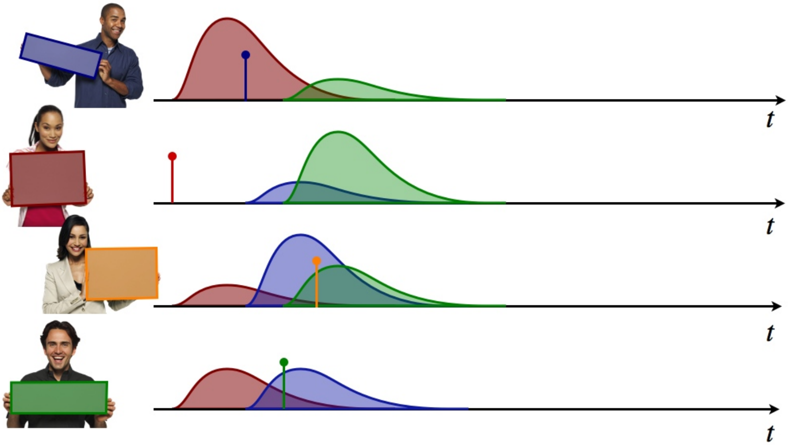
Multivariate Hawkes Processes



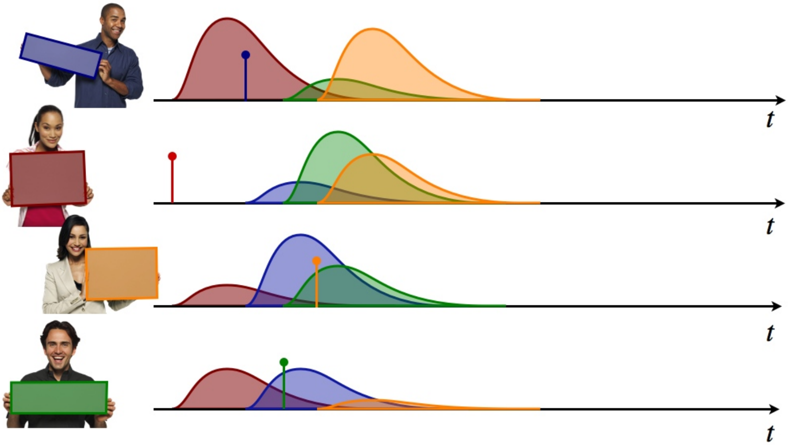
Multivariate Hawkes Processes



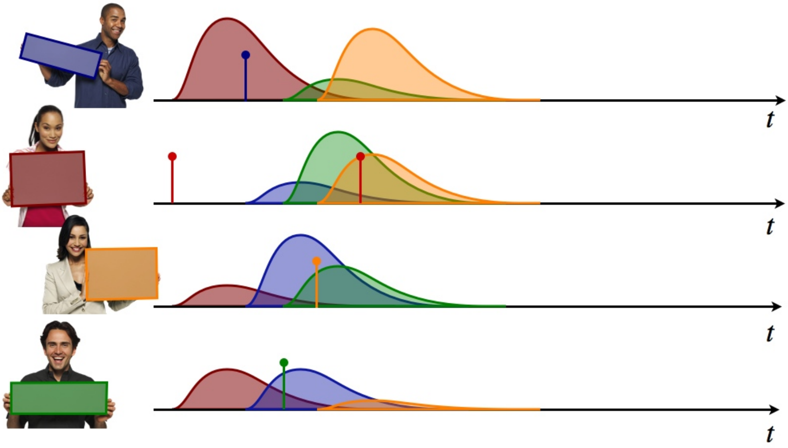
Multivariate Hawkes Processes



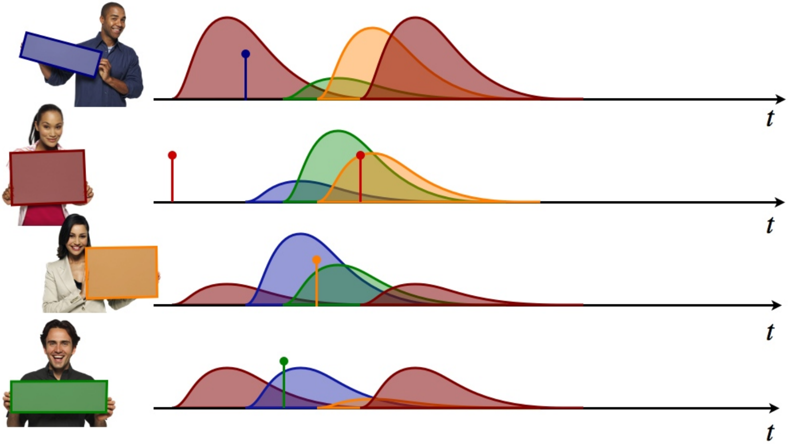
Multivariate Hawkes Processes



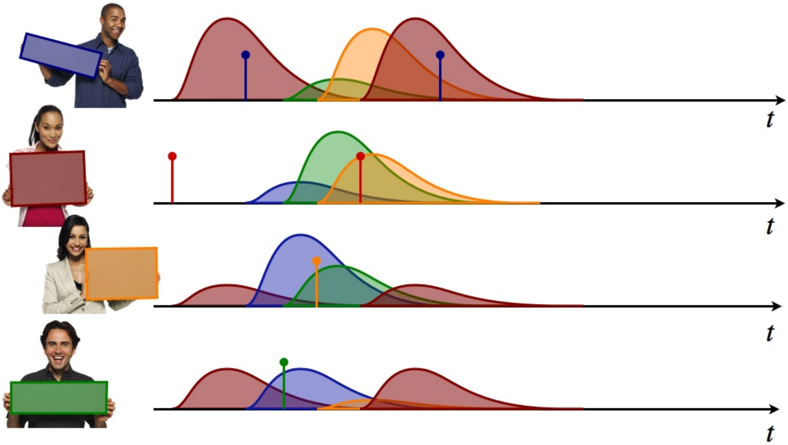
Multivariate Hawkes Processes



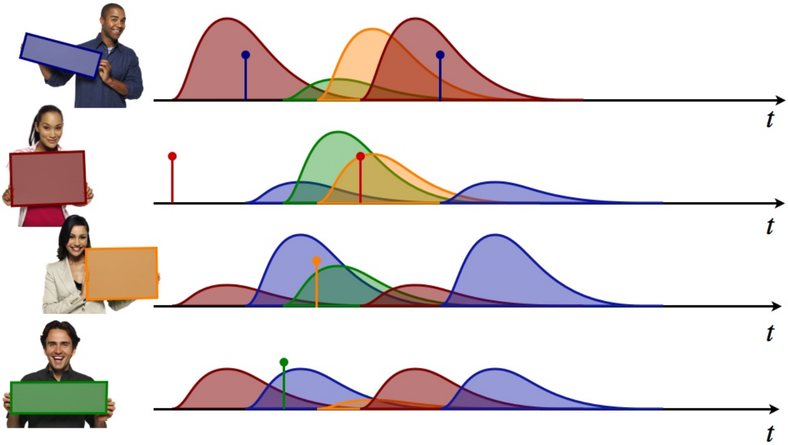
Multivariate Hawkes Processes



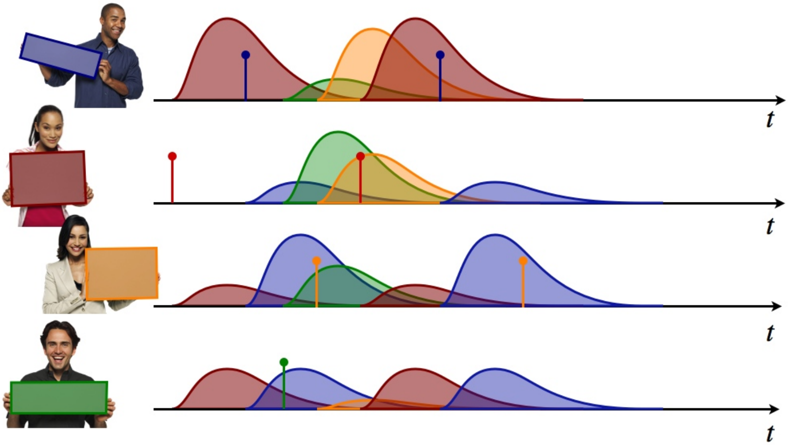
Multivariate Hawkes Processes



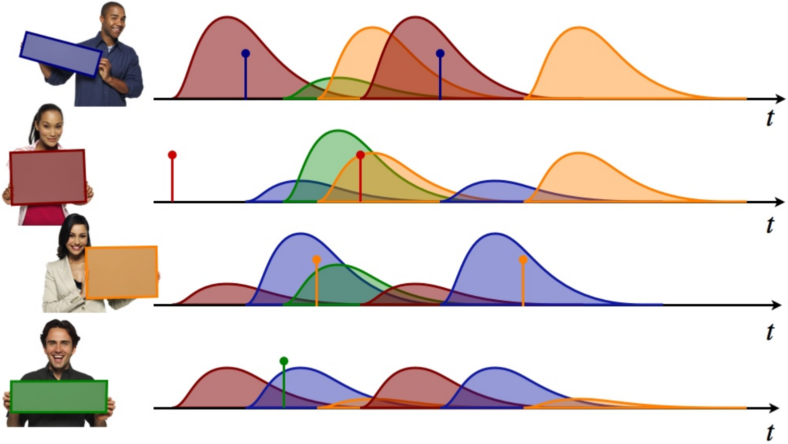
Multivariate Hawkes Processes



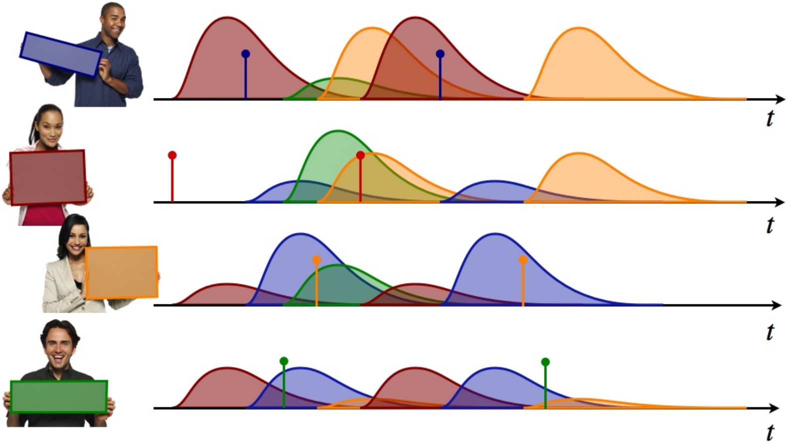
Multivariate Hawkes Processes



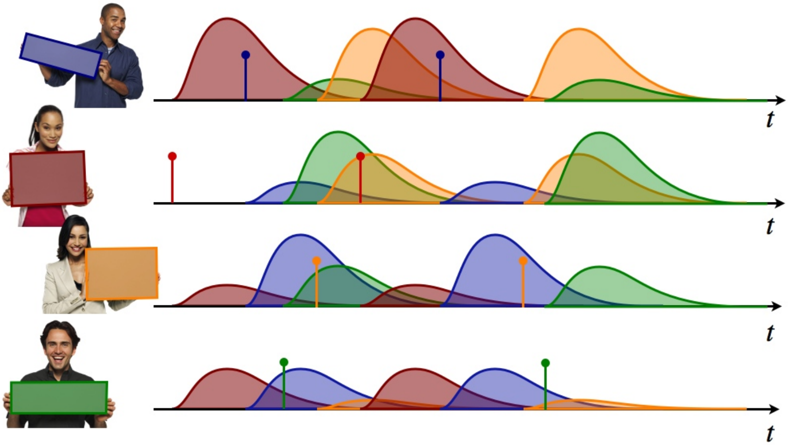
Multivariate Hawkes Processes



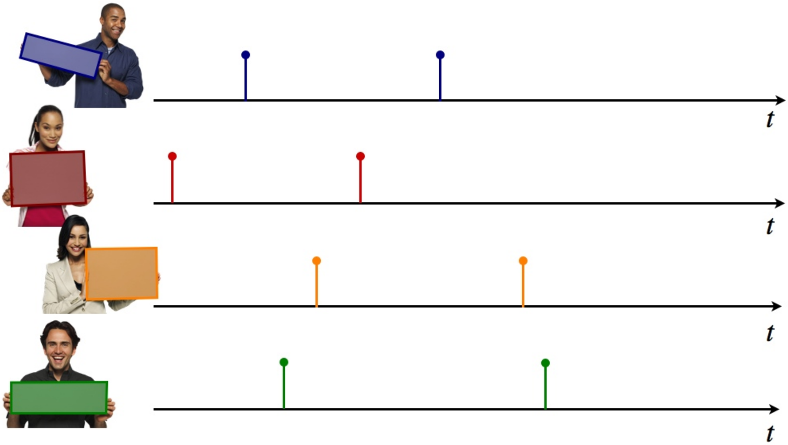
Multivariate Hawkes Processes



Multivariate Hawkes Processes



Multivariate Hawkes Processes



Multivariate Hawkes Processes

The multivariate Hawkes processes[3] considered here are essentially an **autoregressive point process**, where each rate function $\mu_k(\tau)$ depends on the history of past events, N^τ:

$$\mu_k(\tau) = \bar{\mu}_k + \sum_{n=1}^{N_\tau} h_{k,k_n}(\tau - \tau_n)$$

The p^2 functions $h_{k_1,k_2}(\tau) = W_{k_1,k_2} h(\tau)$ describe how events associated with node k_1 will impact the likelihood of events associated with node k_2.

[3] Hawkes (1971)

Network model

The functions $h_{k_1,k_2}(\tau)$ depend on the (unknown) underlying network connectivity. We assume

$$h_{k_1,k_2}(\tau) = W_{k_1,k_2} h(\tau),$$

where the matrix W represents excitatory influences between nodes.

Our goal is to learn and track Hawkes processes *efficiently* and *robustly* from streaming observations.

Online learning

Let θ be a parameter defining our Hawkes process. For instance, θ might be the weighted adjacency matrix W.

Sequence of events: set initial "prediction" $\widehat{\theta}_1$. At time t:

1. Observe **datum** x_t indicating which nodes participated in events at time t.
2. Incur **loss** $\ell_t(\widehat{\theta}_t) \propto -\log p(x_t|\widehat{\theta}_t)$
3. Make a **prediction** $\widehat{\theta}_{t+1}$, which determines the likelihood of nodes participating in an event at time $t+1$

How do we make these predictions? How do we evaluate the efficacy of different prediction strategies?

Regret

Definition: The **regret** of $\widehat{\boldsymbol{\theta}}_T = (\widehat{\theta}_1, \ldots, \widehat{\theta}_T)$ with respect to a comparator $\boldsymbol{\theta}_T = (\theta_1, \ldots, \theta_T)$ is

$$R_T(\boldsymbol{\theta}_T) \triangleq \sum_{t=1}^{T} \ell_t(\widehat{\theta}_t) - \sum_{t=1}^{T} \ell_t(\theta_t).$$

Goal: Generate losses comparable to what a batch algorithm might achieve; *i.e.*, **sublinear regret**:

$$\frac{1}{T} R_T(\boldsymbol{\theta}_T) \to 0 \text{ as } T \to \infty$$

Mirror descent[4]

$$\widehat{\theta}_{t+1} = \arg\min_{\theta} \eta_t \left\langle \nabla \ell_t(\widehat{\theta}_t), \theta \right\rangle + D(\theta, \widehat{\theta}_t)$$

- $\nabla \ell_t$ is an arbitrary subgradient of ℓ_t
- η_t is the step size
- Special case where $D(\theta, \theta') = \|\theta - \theta'\|^2$:

$$\widehat{\theta}_{t+1} \equiv \widehat{\theta}_t - \frac{1}{\eta_t} \nabla \ell_t(\widehat{\theta}_t)$$

[4] Nemirovski & Yudin 1983; Beck & Teboulle 2003; Zinkevich 2003

Tracking W directly

With the Hawkes process, we have a negative log likelihood

$$-\log(N^T|\mu) \approx \sum_{t=1}^{T/\delta} \langle \delta\mu_t, \mathbb{1} \rangle - \langle x_t, \log \delta\mu_t \rangle$$

where

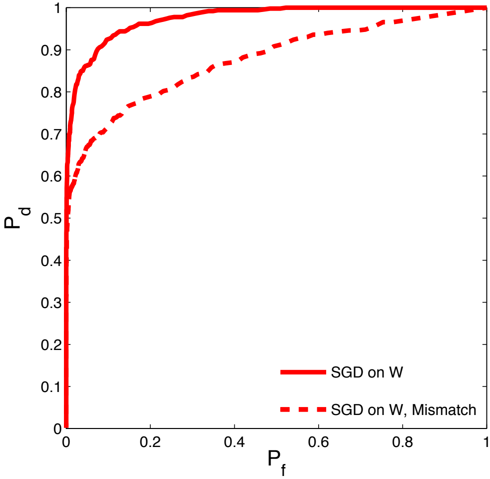
$$\mu_{t,k}(W) = \bar{\mu}_k + \sum_{n=1}^{N_\tau} W_{k,k_n} h(\delta t - \tau_n).$$

Thus we could define the loss function

$$\ell_t(W) = \langle \delta\mu_t(W), \mathbb{1} \rangle - \langle x_t, \log \delta\mu_t(W) \rangle$$

and perform mirror descent directly on W over a convex feasible space $\mathcal{W}$ (*e.g.*, ℓ_1 or nuclear norm ball).

Detection of strong network edges



Unfortunately, this only works well when the influence functions $h_{k,m}(\tau)$ are known *exactly*.

Just tracking W is fragile to model mismatch. Can we instead track W and μ *simultaneously* for increased robustness?

We have

$$-\log(N^T|\mu) \approx \sum_{t=1}^{T/\delta} \langle \delta\mu_t, \mathbb{1} \rangle - \langle x_t, \log \delta\mu_t \rangle.$$

Define the loss to be

$$\ell_t(\mu) \triangleq \langle \delta\mu, \mathbb{1} \rangle - \langle x_t, \log \delta\mu \rangle$$

Static regret bounds

Theorem[5]**:** Assume $\boldsymbol{\mu}_T$ is static, so that $\mu \triangleq \mu_1 = \mu_2 = \ldots = \mu_T$. If $\eta_t \propto 1/\sqrt{T}$, then

$$R_T(\boldsymbol{\mu}_T) \triangleq \sum_{t=1}^{T} \ell_t(\widehat{\mu}_t) - \sum_{t=1}^{T} \ell_t(\mu_t) = O\left(\sqrt{T}\right).$$

What's missing?

- Comparing against a static model is weak; how do we do relative to a dynamic comparator?
- What about unknown underlying networks reflecting interactions between data and the θ_ts?

[5] Nemirovski & Yudin 1983; Beck & Teboulle 2003; Zinkevich 2003

Tracking regret against time-varying reference models

Theorem:[6] If $\eta_t = 1/\sqrt{t}$, then

$$R_T(\boldsymbol{\mu}_T) = O\left(\sqrt{T}(V_T(\boldsymbol{\mu}_T) + 1)\right),$$

where

$$V_T(\boldsymbol{\mu}_T) \triangleq \sum_{t=1}^{T-1} \|\mu_{t+1} - \mu_t\|$$

measures the temporal variation in $\boldsymbol{\mu}_T$.

In other words, the algorithm can track a dynamically changing environment, provided the **changes are sufficiently infrequent and/or smooth (restrictive!)**

[6] Herbster & Warmuth 2001, Cesa-Bianchi & Lugosi 2006, Cesa-Bianchi *et al.* 2012

A dynamical model perspective of Hawkes processes

Recall the Hawkes model

$$\mu_k(\tau) = \bar{\mu}_k + \sum_{n=1}^{N_t} W_{k,k_n} h(\tau - \tau_n)$$

and let

$$h(\tau) = e^{-r\tau} u(\tau).$$

This suggests the dynamical models

$$\mu_{t+1} \approx \Phi_t(\mu_t, W) \triangleq (1 - e^{-r\delta})\bar{\mu} + e^{-r\delta}\mu_t + e^{-r\delta} W x_t$$

How do we incorporate dynamics into mirror descent?

Dynamic Mirror Descent (DMD)

Our approach: Let Φ_t be a series of predetermined dynamical models; set

$$\widetilde{\mu}_{t+1} = \arg\min_{\mu} \eta_t \langle \nabla \ell_t(\Phi_t(\widetilde{\mu}_t)), \mu \rangle + D(\mu \| \Phi_t(\widetilde{\mu}_t))$$

$$\widehat{\mu}_{t+1} = \Phi_{t+1}(\widetilde{\mu}_{t+1})$$

Theorem: Assume each Φ_t is contractive, so that

$$D(\Phi_t(\mu) \| \Phi_t(\mu')) \le D(\mu \| \mu') \;\; \forall \mu, \mu'.$$

Then if $\eta_t \propto \frac{1}{\sqrt{t}}$ we have $R_T(\boldsymbol{\mu}_T) = O(\sqrt{T}[1 + V_\Phi(\boldsymbol{\mu}_T)])$ where

$$V_\Phi(\boldsymbol{\mu}_T) \triangleq \sum_{t=1}^{T} \|\mu_{t+1} - \Phi_{t+1}(\mu_t)\|$$

measures **the deviation of the comparator from the dynamic models (Φ_ts)**.

Contractivity

Contractivity condition:

$$D(\Phi_t(\theta) \| \Phi_t(\theta')) - D(\theta \| \theta') \leq 0 \qquad \forall \theta, \theta', t$$

Q: How much does this condition restrict the class of W we may track?

A: W must ensure $\widetilde{\Phi}_t(\mu) \geq 0$ for all $\mu \succeq 0$ – e.g. W models excitation, not inhibition.

In particular, the dynamics are contractive whenever $\widetilde{\Phi}_t(\mu)$ has the form

$$\widetilde{\Phi}_t(\mu) = A_t \mu + W_t b_t + c_t$$

for arbitrary nonnegative W_t, b_t and c_t as long as the eigenvalues of A are bounded by one.

In our setup, $A_t = e^{-rt} I$, so we simply need $r > 0$.

Tracking W indirectly

In our setting the dynamical model Φ_t is a function of W.

- W is unknown – and it may be changing over time
- The space of possible Ws is huge

Fortunately, there is still a way to track W.

Lemma: For any $W, W' \in \mathbb{R}^{p\times p}$, let $\widehat{\mu}_1^{(W)} = \widehat{\mu}_1^{(W')}$. Then

$$\widehat{\mu}_t^{(W)} = \widehat{\mu}_t^{(W')} + (W - W')K_t$$

where

$$K_t = (1 - \eta_{t-1})A_{t-1}K_{t-1} + X_{t-1}.$$

This lemma suggests we may compute $\widehat{\mu}_t^{(W)}$ for **any** W, and from there easily calculate the $\widehat{\mu}_t^{(W)}$ we **would have computed** had we used a different W from the beginning.

Tracking W indirectly

This lemma suggests we may compute $\widehat{\mu}_t^{(W)}$ for **any** W, and from there easily calculate the $\widehat{\mu}_t^{(W)}$ we **would have computed** had we used a different W from the beginning.

Define the loss with respect to W

$$g_t(W) \triangleq \ell_t(\widehat{\mu}_t^{(W)}).$$

- $g_t(W)$ is convex in W
- $g_t(W)$ and its gradient are both easily computable
- We can use mirror descent on the sequence of losses g_t over any convex feasible set $\mathcal{W}$

Proposed method

Initialize $\hat{W}_1 = \mathbf{0}$, $K_1 = \mathbf{0}$, $\hat{\mu}_1 = \mathbb{1}$
For $t = 1, ..., T$

$$\ell_t(\widehat{\mu}_t) = \langle \delta \widehat{\mu}_t, \mathbb{1} \rangle - \langle \log \delta \hat{\mu}_t, x_t \rangle$$

incur loss

$$\widehat{W}_{t+1} = \mathrm{Proj}_{\mathcal{W}} \left[\widehat{W}_t - \tau_t \left(\frac{-K_t^T X_t}{\widehat{\mu}_t^0 + K_t \widehat{W}_t} + K_t^T \mathbb{1} \right) \right]$$

update network estimate

$$K_{t+1} = (1 - \eta_t) A_t K_t + X_t$$

bookkeeping

$$\widetilde{\mu}_{t+1} = (1 - \eta_t) \widehat{\mu}_t + \eta_t x_t$$

gradient descent

$$\widehat{\mu}_{t+1} = \Phi_t(\widetilde{\mu}_{t+1}, \widehat{W}_t) + (\widehat{W}_{t+1} - \widehat{W}_t) K_{t+1}$$

update prediction using current network est.

Proposed method

Initialize $\hat{W}_1 = \mathbf{0}$, $K_1 = \mathbf{0}$, $\hat{\mu}_1 = \mathbb{1}$
For $t = 1, ..., T$

$$\begin{aligned}
\ell_t(\widehat{\mu}_t) =& \langle \delta\widehat{\mu}_t, \mathbb{1}\rangle - \langle \log \delta\hat{\mu}_t, x_t\rangle \\
\widehat{W}_{t+1} =& \mathrm{Proj}_{\mathcal{W}}\left[\widehat{W}_t - \tau_t\left(\frac{-K_t^T X_t}{\widehat{\mu}_t^0 + K_t\widehat{W}_t} + K_t^T\mathbb{1}\right)\right] \\
K_{t+1} =& (1-\eta_t)A_t K_t + X_t \\
\widetilde{\mu}_{t+1} =& (1-\eta_t)\widehat{\mu}_t + \eta_t x_t \\
\widehat{\mu}_{t+1} =& \Phi_t(\widetilde{\mu}_{t+1}, \widehat{W}_t) + (\widehat{W}_{t+1} - \widehat{W}_t)K_{t+1}
\end{aligned}$$

Main result

Theorem: Let $\mathcal{W}$ be a convex set of feasible influence matrices W; this set may reflect sparsity or rank constraints.

Let $\Phi_t(\cdot, W)$ be a contractive dynamical model for all $W \in \mathcal{W}$ and $t = 1, 2, \ldots$. Let the sequence $\widehat{\boldsymbol{\mu}}_T$ be the output of our method, and let $\boldsymbol{\mu}_T$ be an arbitrary sequence. If $\eta_t = 1/\sqrt{t}$, then

$$R_T(\boldsymbol{\mu}_T) = O(\sqrt{T}[1 + \min_{W \in \mathcal{W}} V_{\Phi,W}(\boldsymbol{\mu}_T)])$$

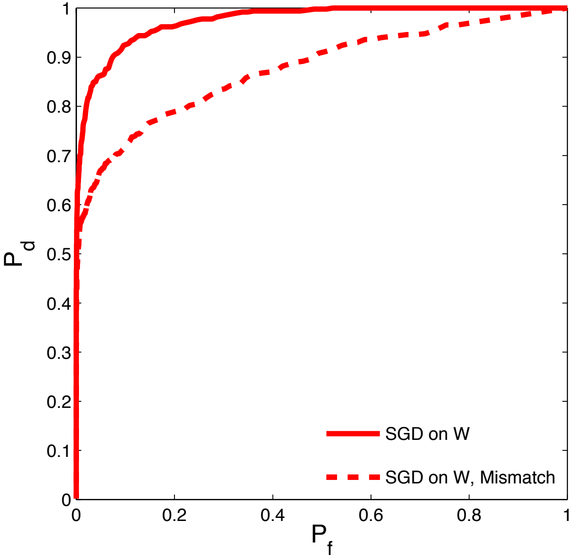
where

$$V_{\Phi,W}(\boldsymbol{\mu}_T) \triangleq \sum_{t=1}^{T} \|\mu_{t+1} - \Phi_t(\mu_t, W)\|$$

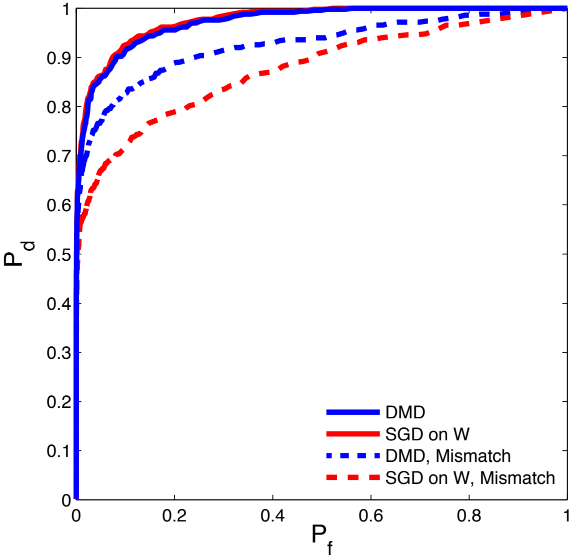
measures variations or deviations of the comparator sequence $\boldsymbol{\mu}_T$ from the sequence of dynamical models $\Phi_1, \Phi_2, \ldots, \Phi_T$.

This regret is low for very large sets of $\boldsymbol{\mu}_T$s.

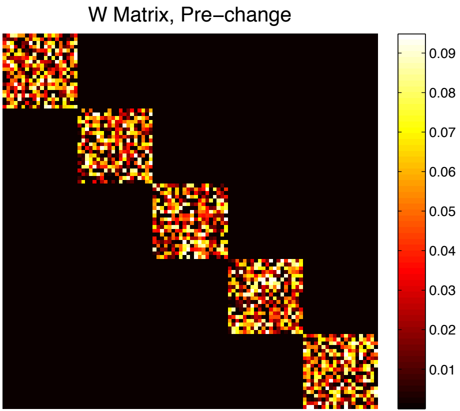
Detection of strong network edges

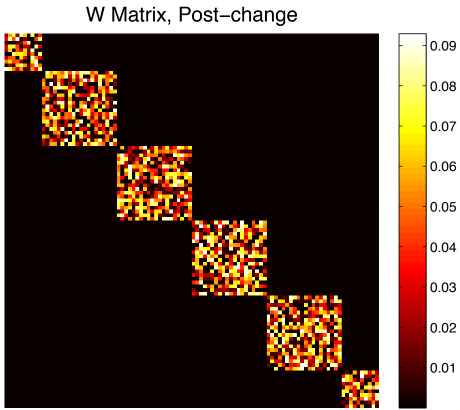


Detection of strong network edges

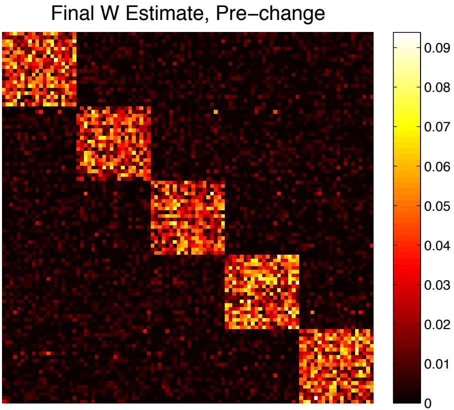


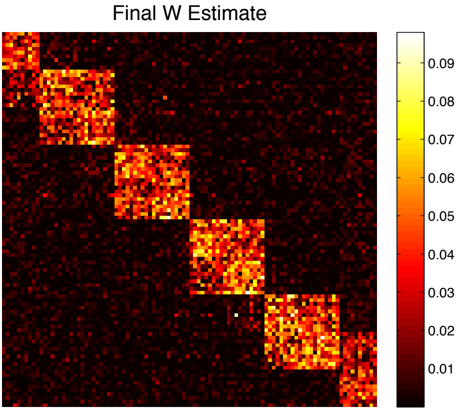
Tracking a changing network



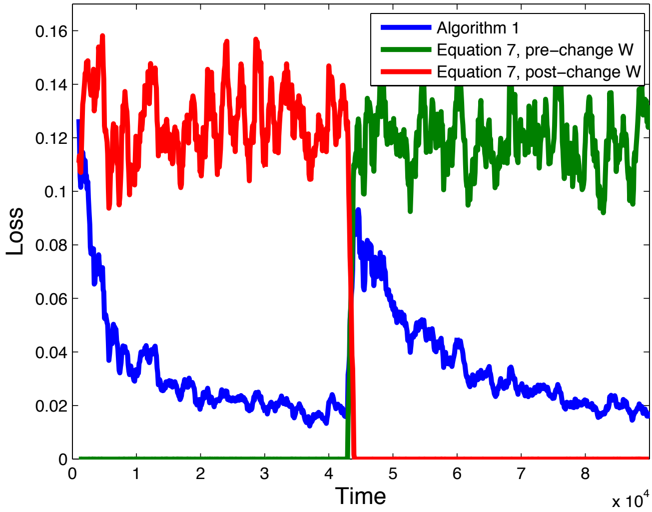


Tracking a changing network



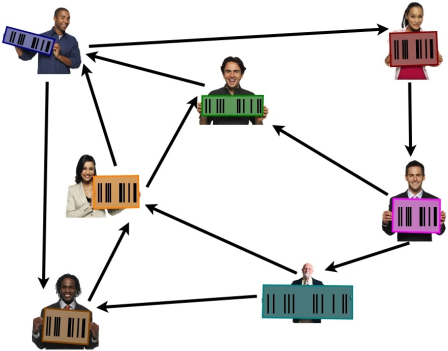


Tracking changing network



Conclusions

- Our techniques offer principled mechanisms for using streaming event observations to track dynamic networks
- Computation scales well with network size
- Theoretical performance bounds are robust to model mismatch and changing networks
- Interesting open questions remain!



Thank you.