Advancing Causal Discovery in Spatio-Temporal Systems: Methods and Applications

Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Oct 29, 2024
STCausal Workshop 2024

NSF

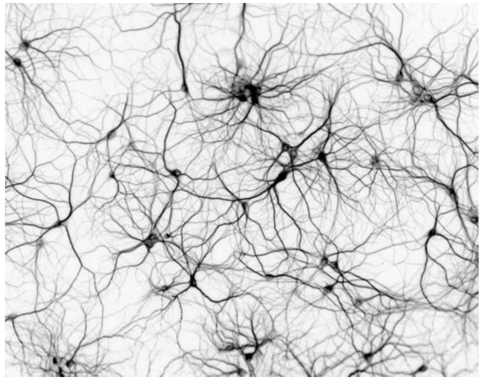
NIH
National Institutes of Health

ATLANTA POLICE FOUNDATION

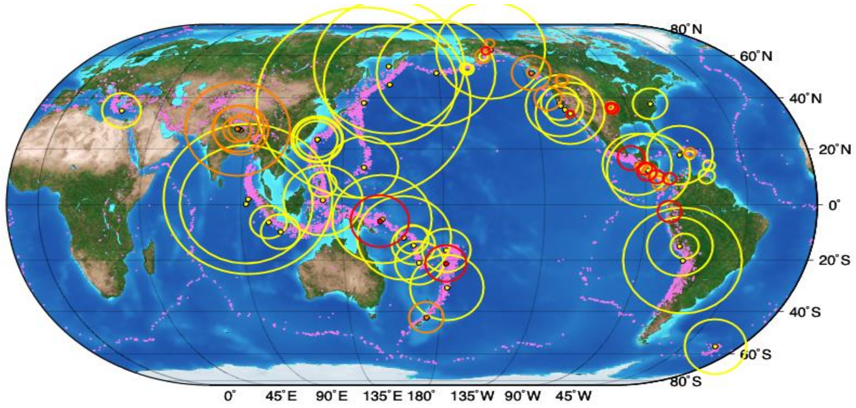
Roadmap

- Basic setup
- Granger network estimation
  - Structural constraints
  - Uncertainty quantification
- General: continuous space
- Other approaches

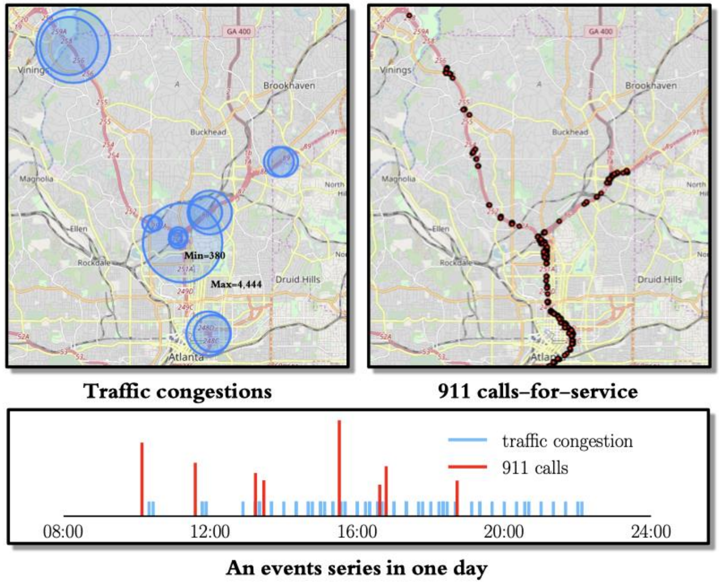
Spatial-temporal data



Neuronal networks

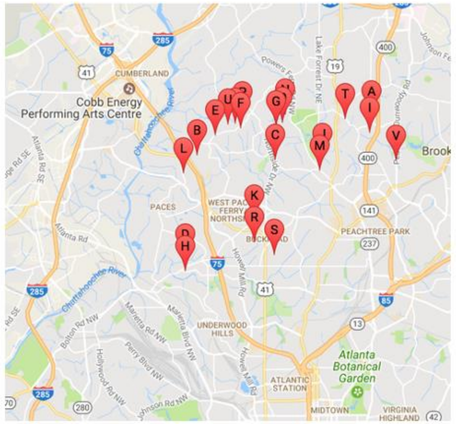


Seismic activities



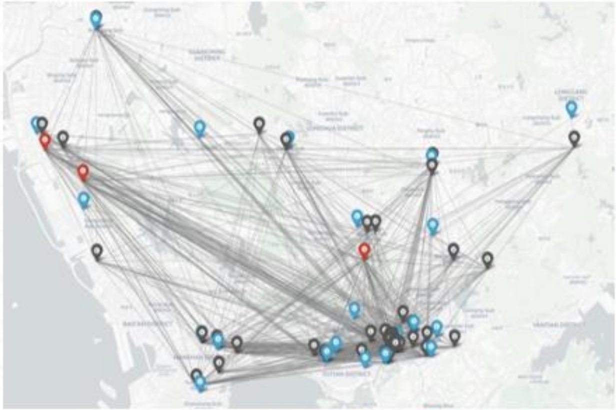
Traffic incidents

22 cases of Buckhead burglary



Crime activities

COVID spread

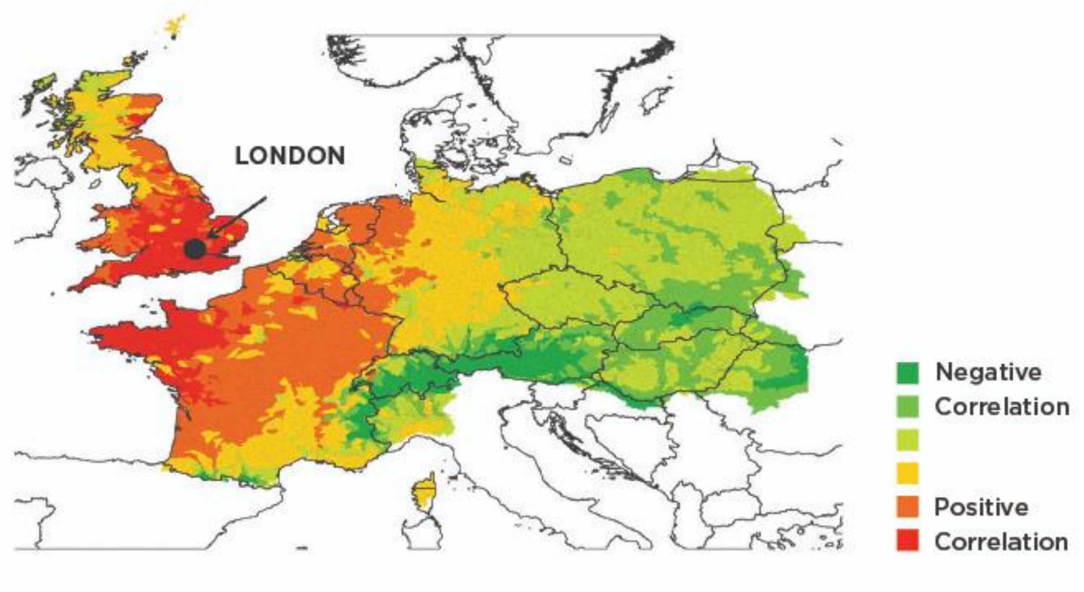


40 access hubs 21 local hubs 3 gateway hubs

Supply chain networks

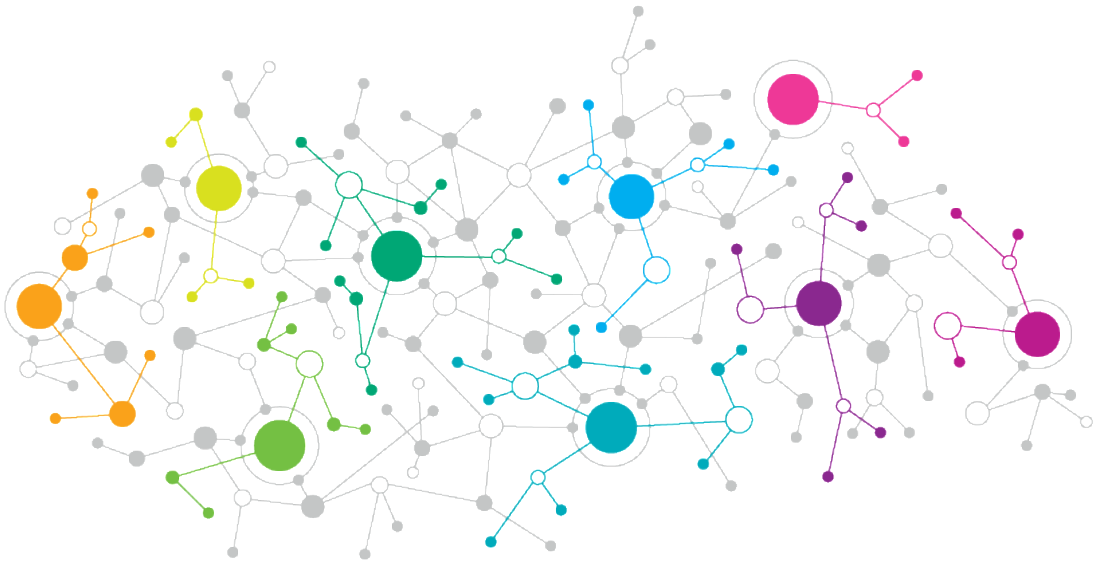
Spatial correlation as network connectivity

- Traditional spatial correlation



Underlying space is Euclidean

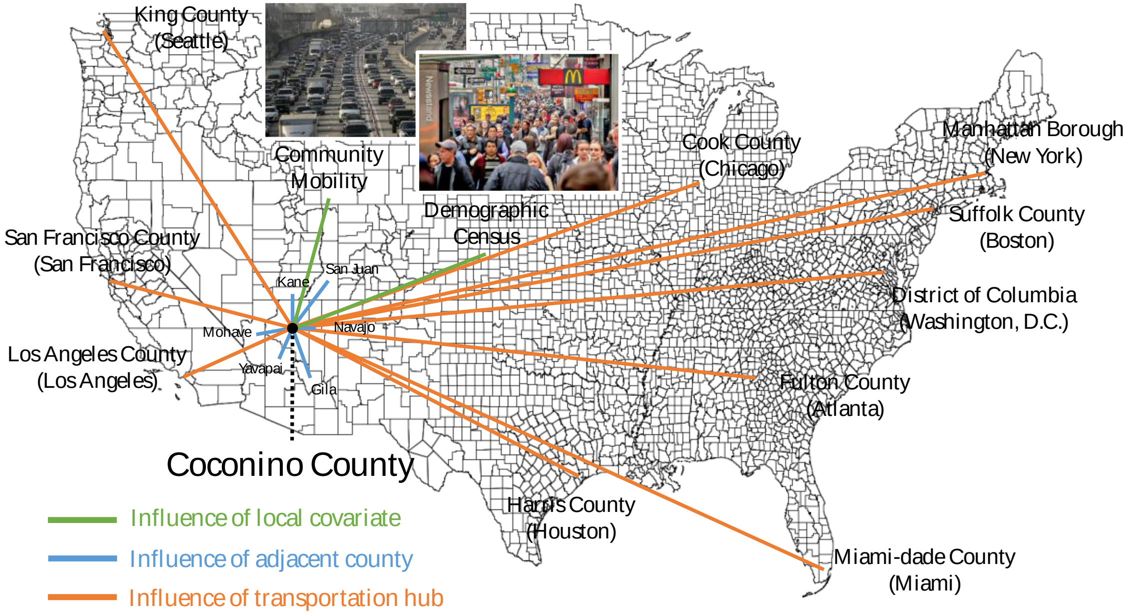
- Network influence



Underlying space is a graph

COVID-19 cases over US counties

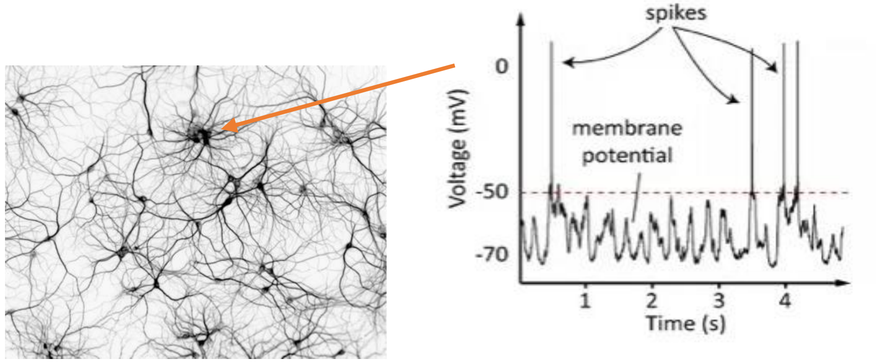
- **Influence**:
  - nearby locations, major cities, and transportation hubs have a larger influence
  - Influence may change over time



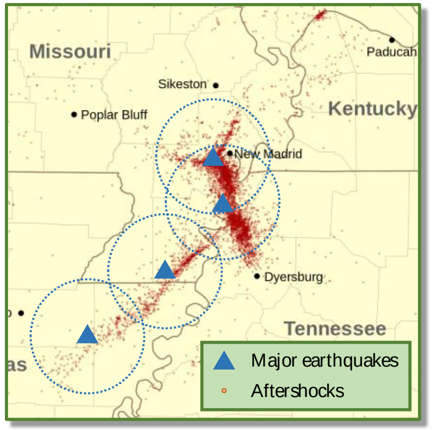
"High-resolution spatio-temporal model for county-level COVID-19 activity in the U.S." Zhu, Bukharin, Xie, Santillana, Yang, **X**. *ACM Transactions on Management Information Systems (TMIS)*, 2021.
"Early detection of COVID-19 hotspots using spatio-temporal data." Zhu, Bukharin, Xie, Yamin, Yang, Keskinocak, and **X**. *IEEE Journal Selected Topics in Signal Processing (JSTSP)* 2022, *ICML Time Series Workshop (Best Paper Award, 2nd Place)* 2021.

Discrete event data (time, marks)

Tweets

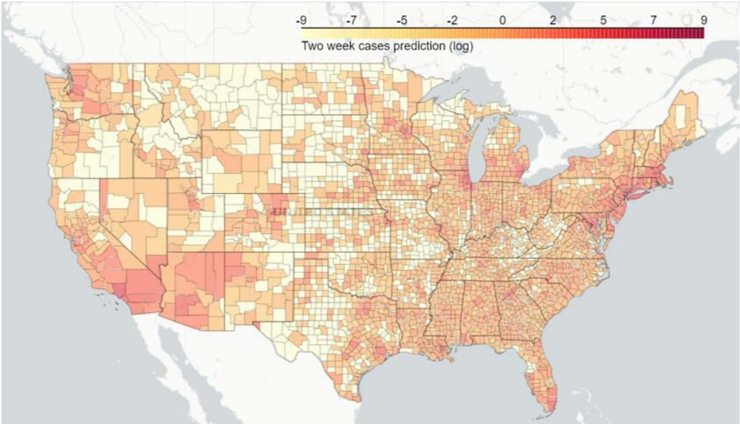


Neural spike trains

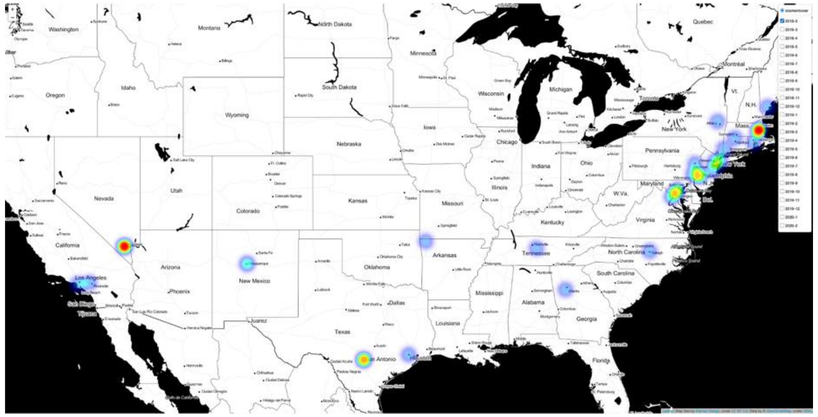


Earthquake catalog

Police reports



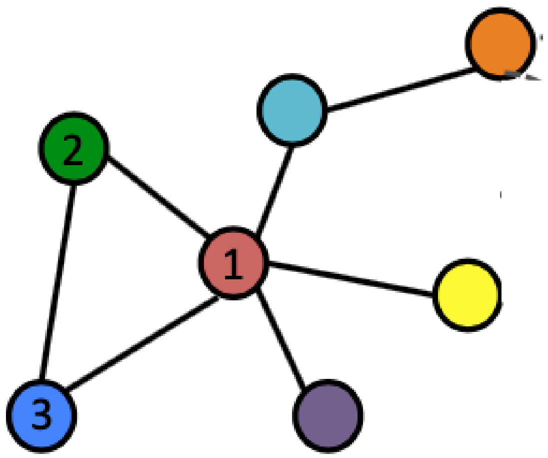
Daily #case at US counties

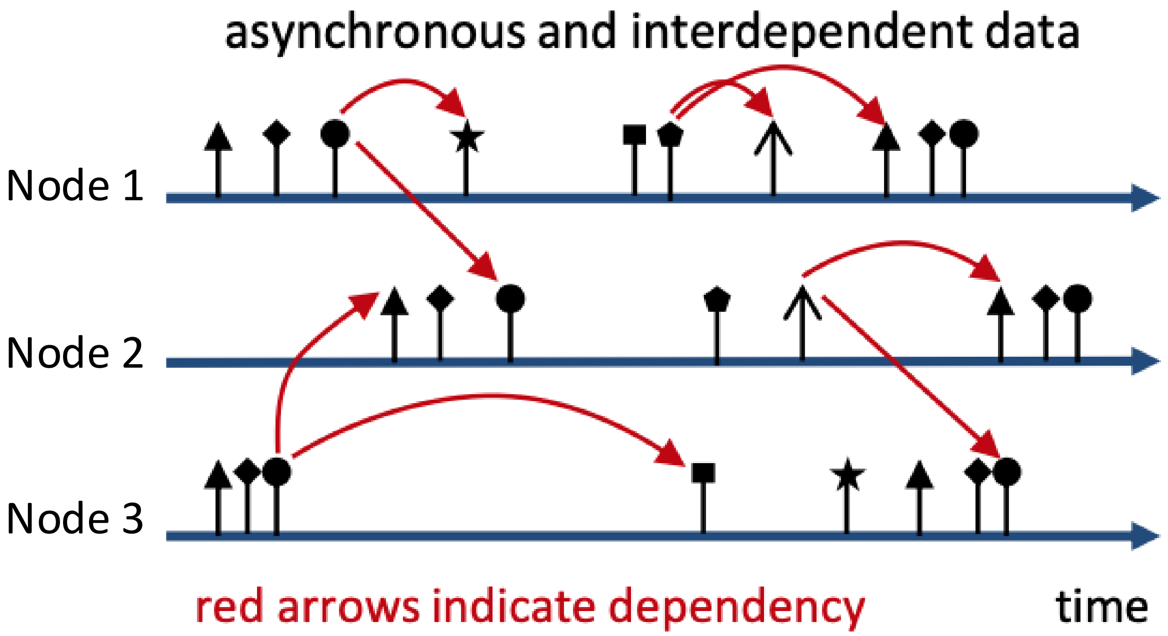


Demand over networks

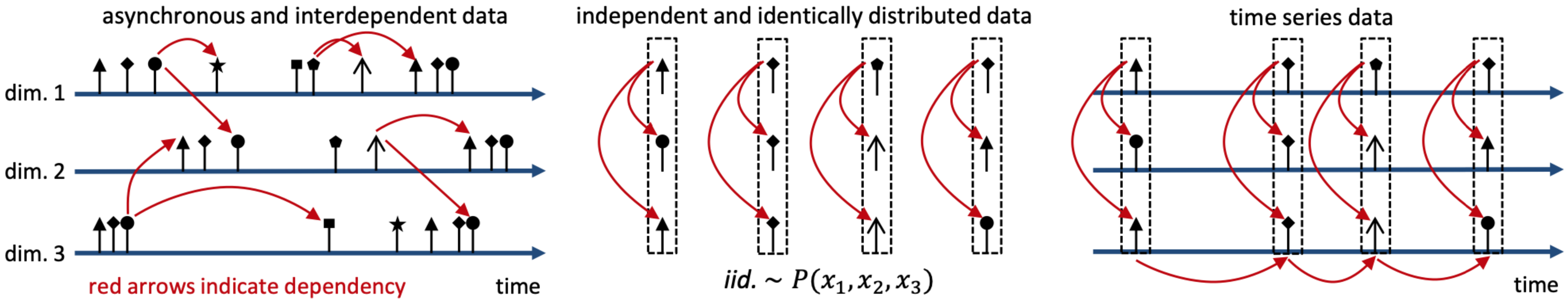
Discrete events data: "Dots"

- Discrete events data: A sequence of (time, marks)
- Asynchronously occur over time and mark space
- Marks contain additional information: location, category, description--can be high-dimensional





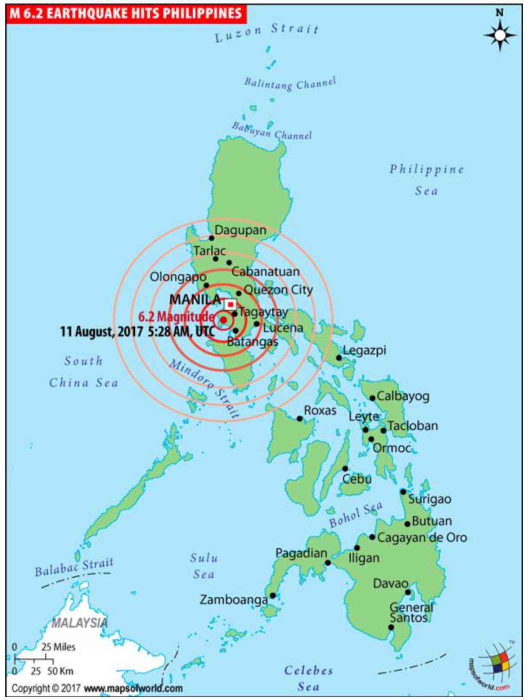
Different from i.i.d. data and classic time-series



- Asynchronously recorded data
- Interrelated over space and time
- Timing of data point carries information

Influence

- “Triggering” or ”inhibition effect” of an event over **space and time**
- Granger causality



Time passes

Zhu, Li, Peng, **X**. Imitation learning of spatio-temporal point processes.
IEEE-TKDE, 2022. *NeurIPS AI for Earth Sciences Workshop*, 2020.

Crime data

- “Broken window effect”

Once a neighborhood has a crime incident, similar crime is more likely to happen.

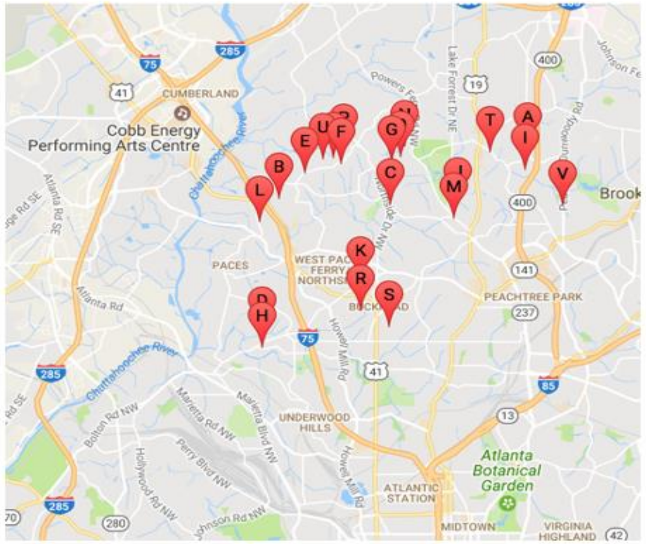
- “Buckhead burglary” in Atlanta, 2017

22 cases committed by a serial offender.

(Zhu, and **X**., *Annals of Applied Statistics*, 2022
Presented at JSM “Best of AOAS, 2021.”)

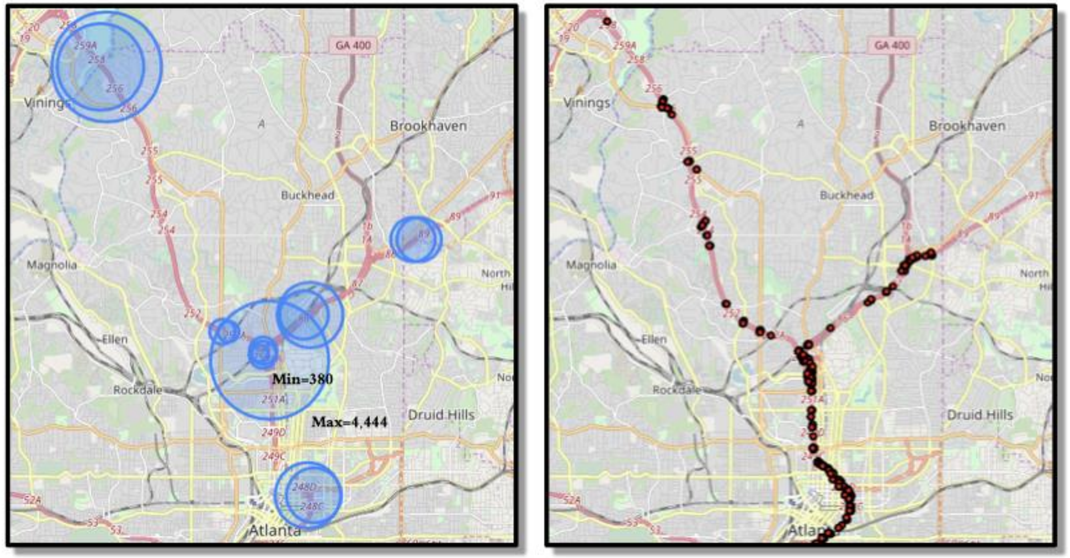
(Collaboration with *Atlanta Police Department*.)

22 cases of Buckhead burglary

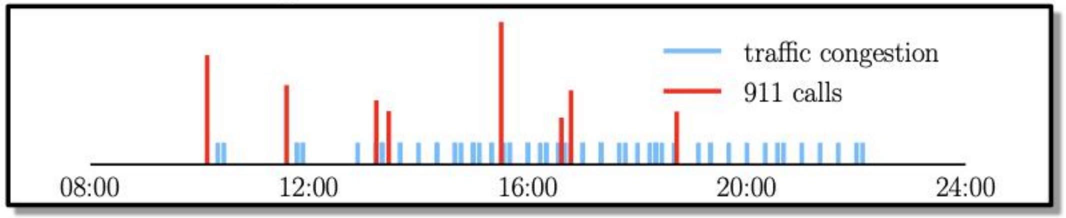


Traffic data

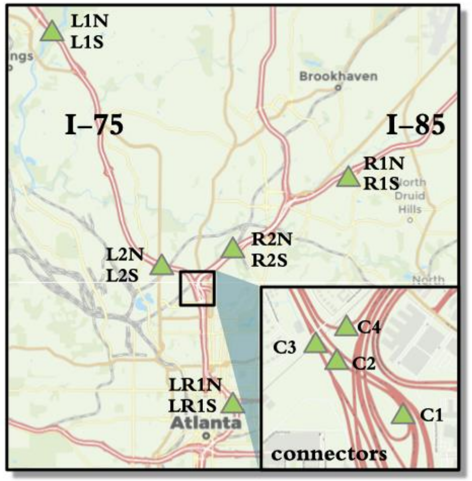
- Traffic congestion events
- Two triggering mechanisms:
  - Traffic congestion triggers future congestion
  - Traffic incidents trigger congestion



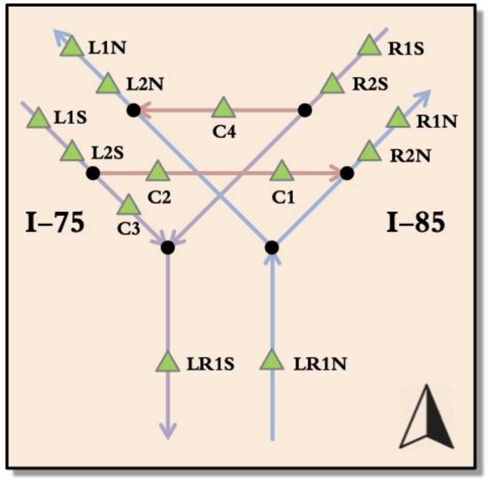
Traffic congestions 911 calls–for–service



An events series in one day



Traffic sensor map in Atlanta

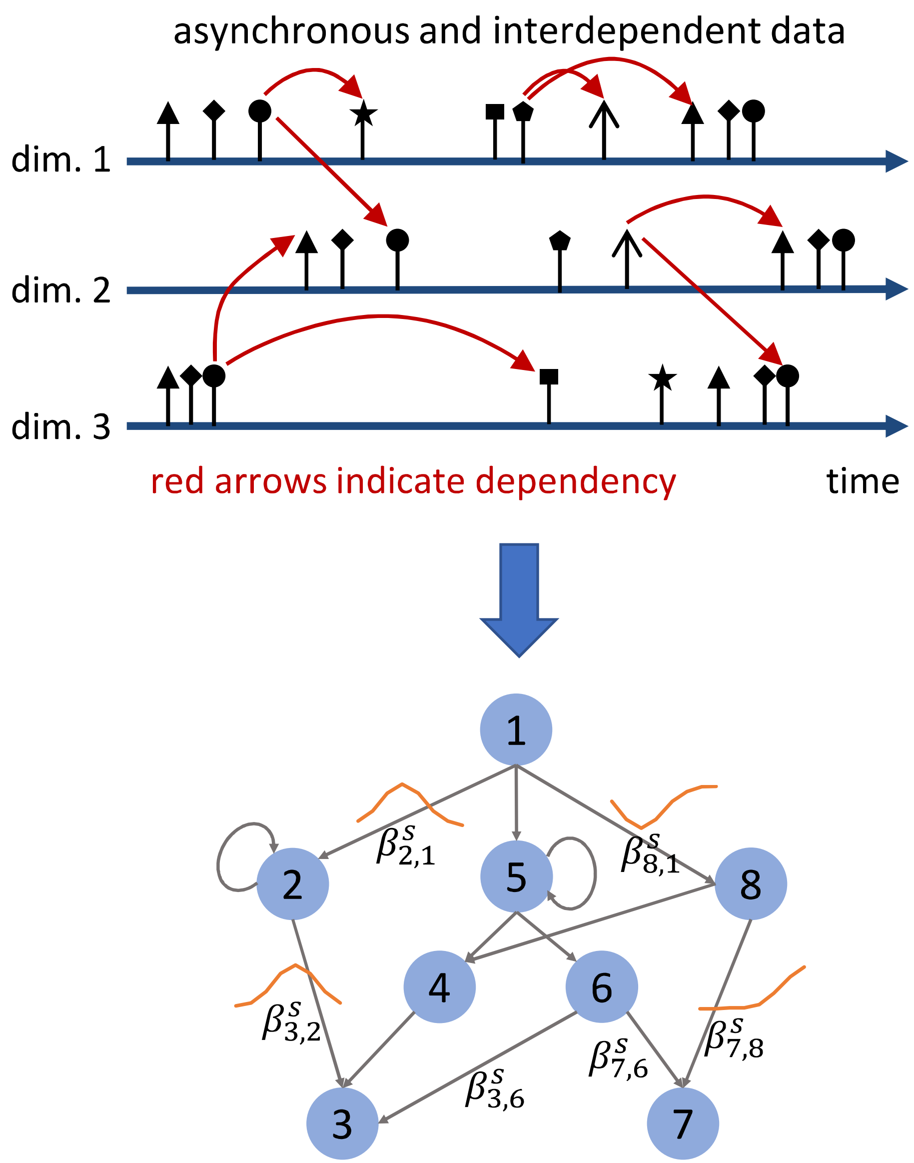


Traffic network

Spatio-temporal point processes with attention for traffic congestion event modeling. Zhu, Ding, Van Hentenryck, and **X.** *IEEE Transactions on Intelligent Transportation Systems*, 2022.

Goals

- Use discrete event data, recover spatio-temporal **influence: Granger causality**
  - **Interpretation**: Understanding underlying influence network and temporal influence
  - **Prediction**: predicting the chance of a future event
  - **Monitoring**: detecting changes – anomalies and novelty
  - **Decision**: intervention, optimization



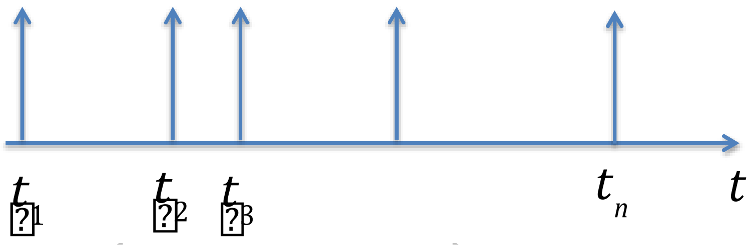
How to model influence?

- **Hawkes processes** (Hawkes 1971)
- Point-process: a sequence of random events at times $\{t_1, t_2, \dots\}$

history

$$\lambda(t)dt = P\{\text{event in } [t, t+dt) | H_t\}$$

$$\lambda(t|H^t) = \lim_{\Delta t \to 0} \frac{E[N(t+\Delta t)|H_t]}{\Delta t}$$



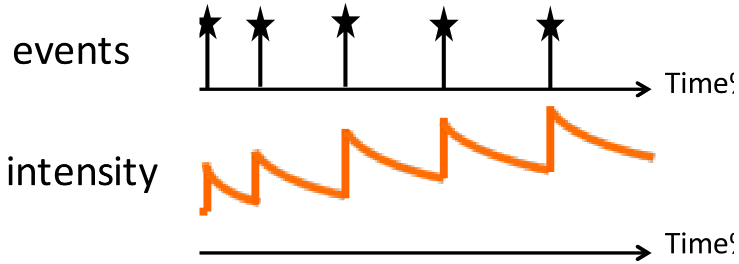

Alan Hawkes

Hawkes, Alan G. "Spectra of some self-exciting and mutually exciting point processes." *Biometrika* 1971.

Common point processes

- Poisson process: $\lambda(t) = \mu(t)$ deterministic
- Hawkes process: conditional intensity depends on history

$$\lambda(t) = \mu(t) + \text{influence from past}$$



- Self-correcting process

$$\lambda(t) = \mu(t) - \text{influence from past}$$

Hawkes process

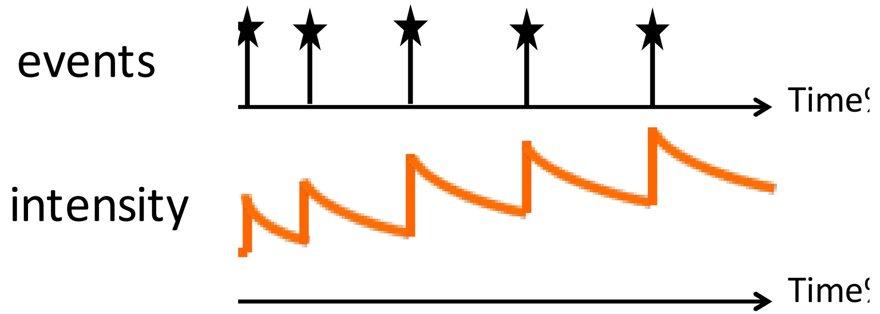
- Conditional intensity function

$$\lambda(t) = \mu(t) + \alpha \sum_{t_k < t} \phi(t - t_k)$$

Baseline intensity

Magnitude of influence

Influence kernel function



Hawkes process over networks

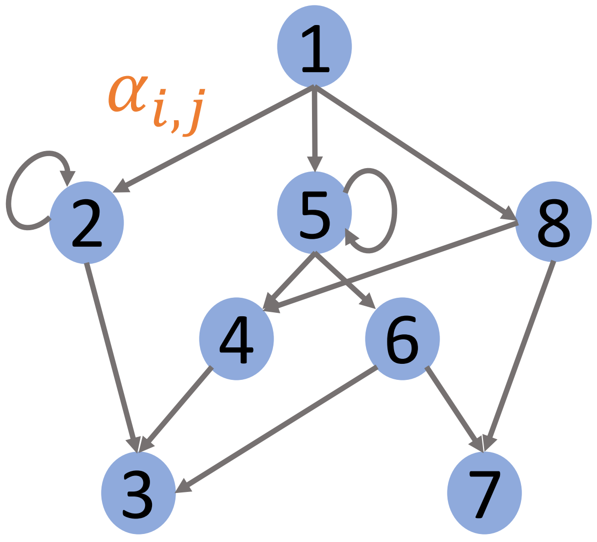
- Events on K nodes $(t_1, u_1), (t_2, u_2), \ldots$

$$\lambda_i(t) = \mu_i(t) + \sum_{t_k < t} \alpha_{i,u_k} \phi(t - t_k)$$

Baseline intensity at node i

Influence between node α_{ij}

Temporal influence kernel function



- Commonly assumed: Exponential decay influence

$$\phi(t) = \beta e^{-\beta t}, t \geq 0 \text{ (Markovian)}$$

Hawkes processes literature

- Single and multi-dimensional Hawkes processes

(Alan Hawkes 1971) (review, Reinhart 2018)

- Continuous spatio-temporal modeling with diffusion kernel (ETAS)

(Ogata 1999) (Zhu et al. 2020)

- Asymptotic convergence results of linear and non-linear processes

(Bacry, Dayri, Muzy 2012) (L. Zhu 2013) (L. Zhu 2015)

- Estimate network interactions, assuming known influence function:

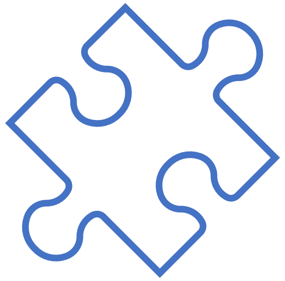
(Stomakhin, Short, Bertozzi 2011), (Myers, Leskovec, 2014), (Rodriguez et al. 2011) (Yang, Zha 2013) (Hall, Willett 2016) (Chen et al. 2017) (Li et al. 2018) (Yuan et al. 2019)

- Causal inference and testing for purely temporal process

(Chen, Witten, Shojaie 2017) (Xu, Farajtabar, Zha, 2016) (Achab et al. 2017)

- Bayesian model

(Rasmussen 2013) (Linderman, Wang Blei 2017) (Donnet, Rivoriard, Rosseau 2020)



- Uncertainty quantitation
- Structural assumptions
- General influence kernel

Roadmap

- Basic setup
- Granger network estimation
    - Structural constraints
    - Uncertainty quantification
- General: continuous space
- Other approaches

Maximum likelihood

- Parameters are solved by maximum likelihood

$$\max_{\theta} \ \ell(\theta)$$

- Property of the optimization problem
  - When $\theta = \{\mu, \alpha_{ij}\}$, and β is fixed, it can be shown that $\ell(\theta)$ is convex in θ
  - When $\theta = \{\mu, \alpha_{ij}, \beta\}$, problem is non-convex
  - When influence $\neq$ exponential decay, may not have closed-form integration

Maximum likelihood estimate for $\boldsymbol{\alpha}_i$

- Define coefficient for i-th node as $\boldsymbol{\alpha}_i$

$$\max_A \ell(A) = \sum_{i=1}^{K} \ell_i(\boldsymbol{\alpha}_i)$$

Decoupled in nodes, enable decentralized estimation

- Log-likelihood for the i-th node

$$\ell_i(\alpha_i) = -\int_0^T \lambda_i(t)dt + \int_0^T \log(\lambda_i(t))dN_t^i$$

- Assuming known influence function $\phi(t)$, $\ell_i(\boldsymbol{\alpha}_i)$ convex function in $\boldsymbol{\alpha}_i$
- Can be solved efficiently to global solution (e.g., gradient descent)

Likelihood function for Hawkes networks

- Log-likelihood function for Hawkes network, exponential influence
- Data $(t_i, u_i), i = 1, \dots, n$

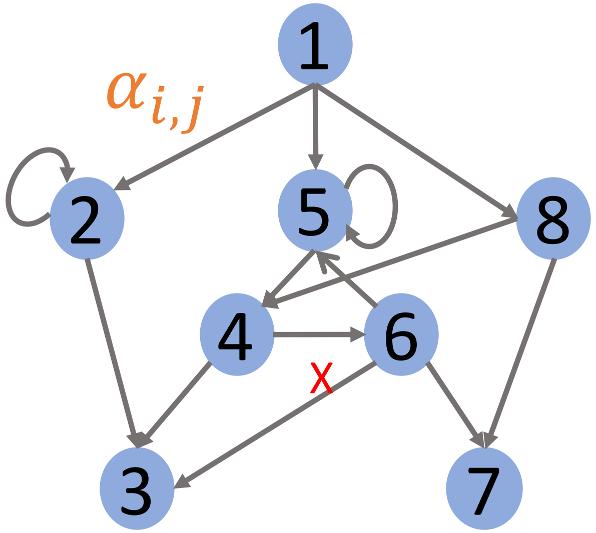
$$\ell(\theta) = \sum_{i=1}^{n} \log\left[\mu_{u_i} + \sum_{t_j<t_i} \alpha_{u_i,u_j} \beta e^{-\beta(t_i - t_j)}\right] - \sum_{j=1}^{K} \mu_j t$$

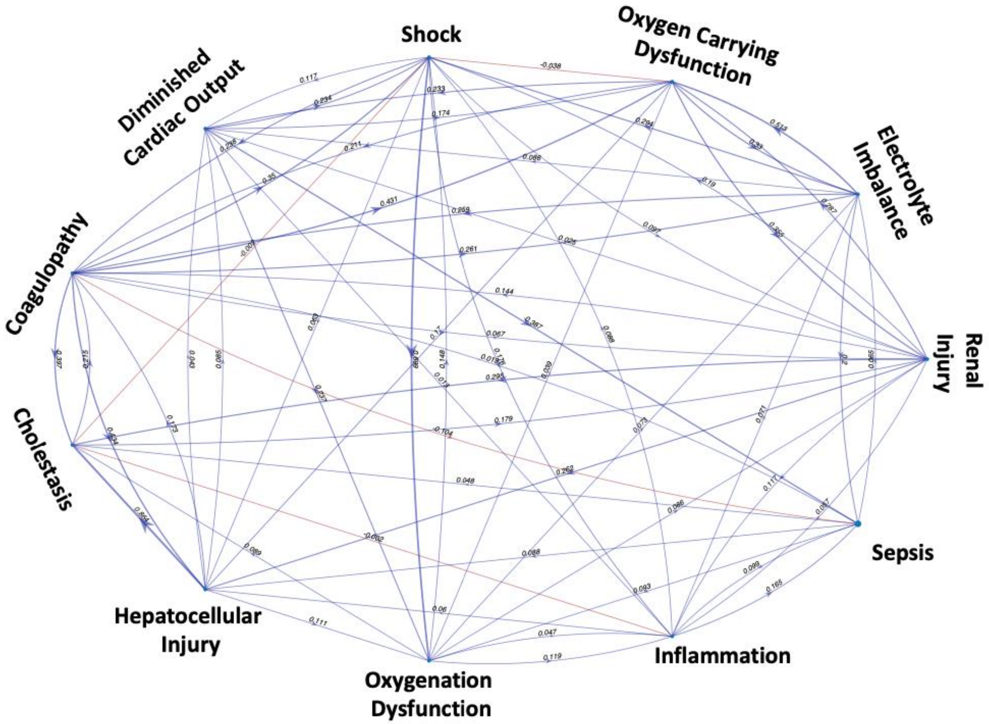
$$- \sum_{j=1}^{K} \sum_{t_i<t} \alpha_{u_i,j} [1 - e^{-\beta(t-t_i)}]$$

- Parameters $\theta = (A, \mu)$ are solved by maximum likelihood: $\max_{\theta} \ \ell(\theta)$
- Convex

Granger causality: Real-time sepsis prediction

- Add Directed Acyclic Graph (DAG) constraints to remove cycles

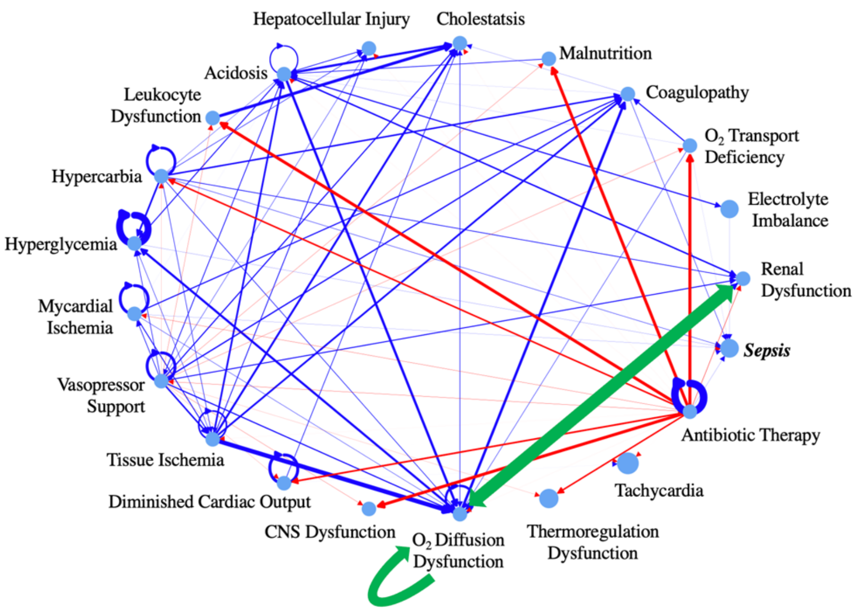


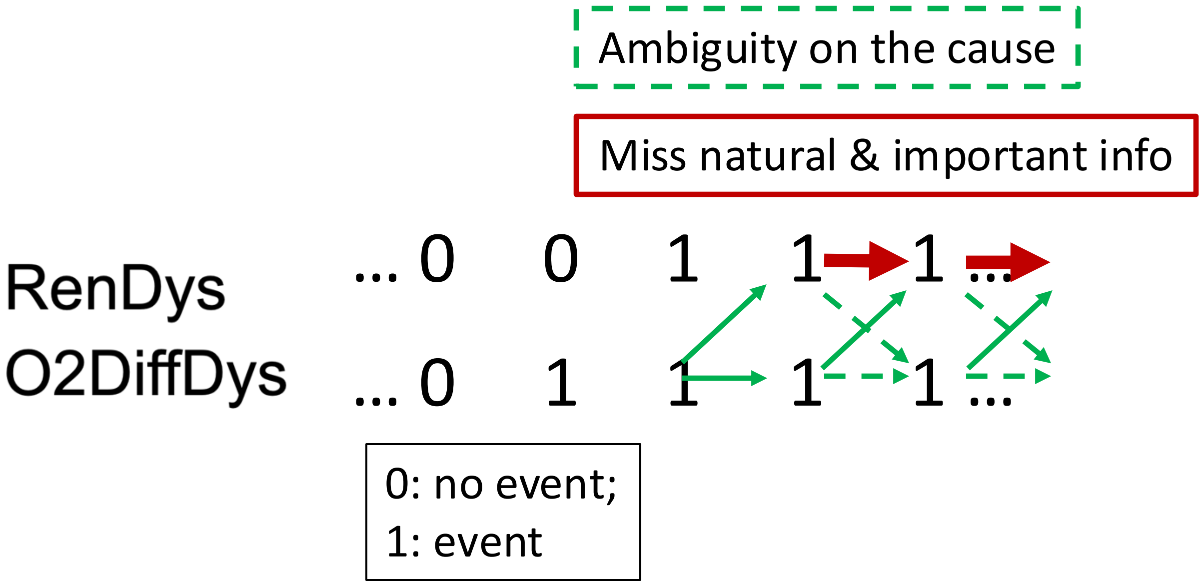


- Granger causal chain discovery for sepsis-associated derangements via multivariate Hawkes processes. Wei, Xie, Josef, Kamaleswaran. KDD 2023.
- Causal graph discovery from self and mutually exciting time series. Wei, Xie, Josef, and Kamaleswaran. IEEE Selected Areas in Information Theory (JSAIT). Vol. 4, pp. 747-761. 2023.

Why need structural assumption

- Our first attempt on Granger causal graph discovery [2] returns **<u>cyclic</u>** patterns, and therefore less reasonable causal interpretations





DAG-encouraging regularization

Graph adjacency matrix A constant d

Motivation
[Zheng et al. (2018)]

$$tr(e^A) = tr(I) + tr(A) + \frac{1}{2}tr(A^2) + \cdots$$

Length-1 cycles **Length-2 cycles** **…**

Zheng, Xun, et al. "Dags with no tears: Continuous optimization for structure learning." Advances in neural information processing systems 31 (2018).

DAG-encouraging regularization

Graph adjacency matrix A   constant d

Motivation
[Zheng et al. (2018)]

$$tr(e^A) = tr(I) + tr(A) + \frac{1}{2}tr(A^2) + \cdots$$

Linear relaxation
(convex!)

[No penalty on length-1 cycles]

$$\alpha_{12} + \alpha_{21} \leq \delta_1$$
$$\alpha_{23} + \alpha_{32} \leq \delta_2$$

… …

…

DAG-encouraging regularization

Graph adjacency matrix A   constant d

Motivation
[Zheng et al. (2018)]

$$tr(e^A) = tr(I) + tr(A) + \frac{1}{2}tr(A^2) + \cdots$$

Length-1 cycles   Length-2 cycles   ...

Linear relaxation
(convex!)

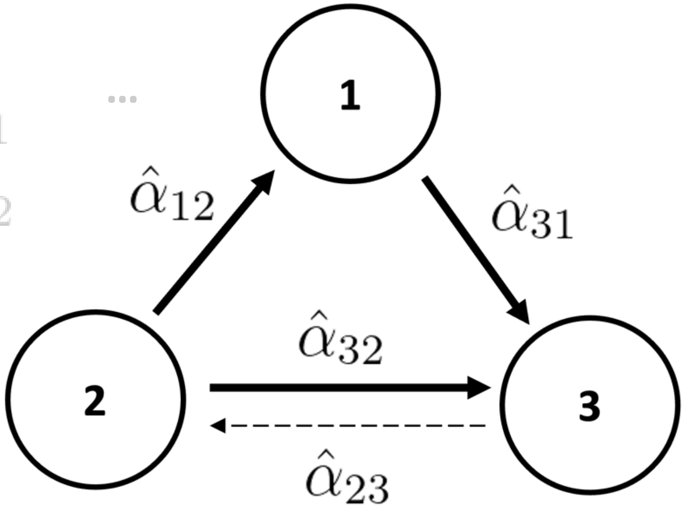
[No penalty on length-1 cycles]

$$\alpha_{12} + \alpha_{21} \leq \delta_1$$
$$\alpha_{23} + \alpha_{32} \leq \delta_2$$

... ...

...

Proposed data-adaptive modification [3]



Step 1: Rough estimation $\hat{A} = (\hat{\alpha}_{ij})$

Step 2: Data-adaptive linear DAG-encouraging constraints

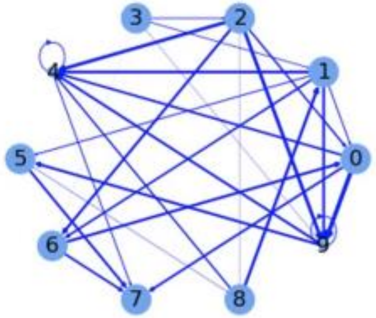
$$\alpha_{23} + \alpha_{32} \leq \hat{\alpha}_{32} \quad \& \quad \alpha_{12} + \alpha_{23} + \alpha_{31} \leq \hat{\alpha}_{12} + \hat{\alpha}_{31}$$

[3] Wei, Song, et al. "Causal graph discovery from self and mutually exciting time series." *IEEE Journal on Selected Areas in Information Theory* (2023).

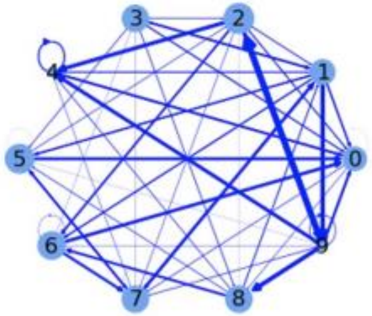
Simulation

Regularization	None	Proposed	DAG	DAG-Variant	ℓ_1	Ada. ℓ_1
A err.	.3874	**.2094**	.3541	.2949	*.2501*	.3022
ν err.	.1175	**.0775**	.0895	*.0841*	.0884	.1251
$h(A_0)$	.1223	.0308	.0337	*.0242*	.0274	**.0232**
SHD	41	**25**	32	34	41	*29*

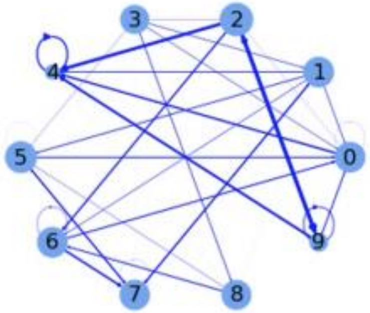
- DAG regularization removes suspicious links and helps parameter recovery



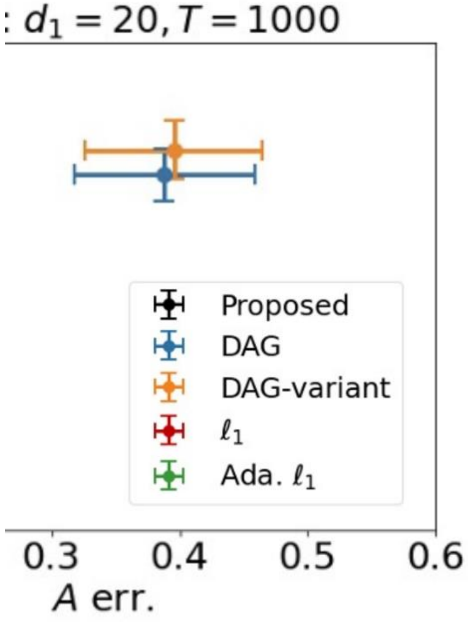
Ground truth

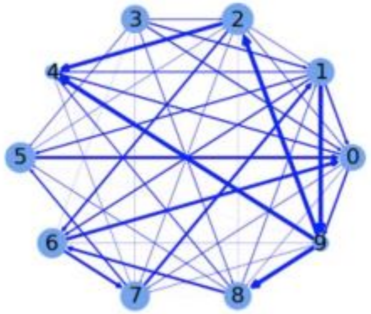


No regularization

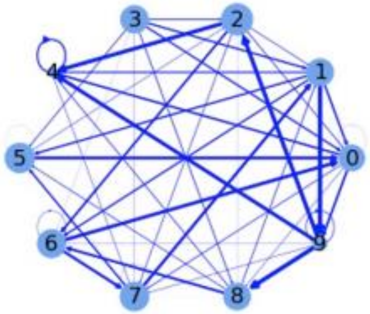


Proposed data-adaptive linear regularization

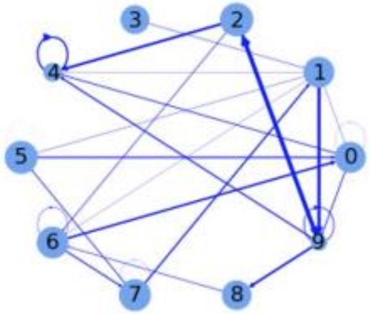




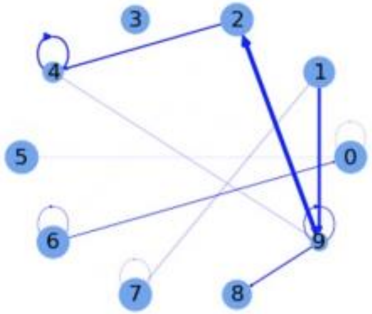
DAG regularization



DAG regularization variant



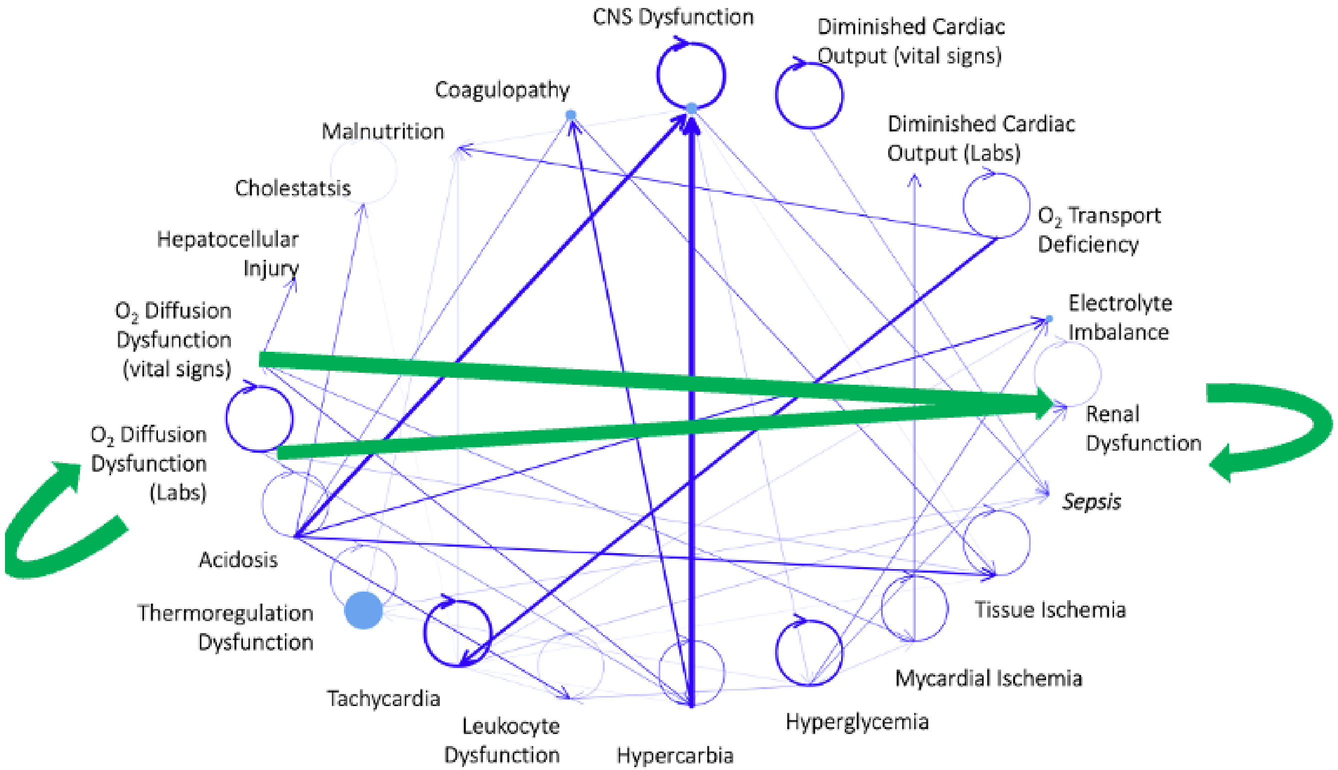
ℓ_1 regularization

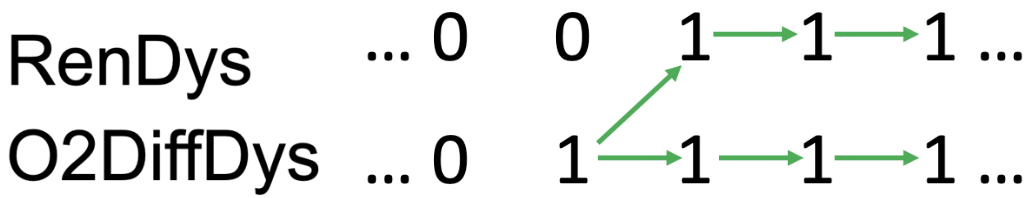


Adaptive ℓ_1 regularization

Real-data experiment

- Resulting Causal DAG





0: no event;
1: event

Roadmap

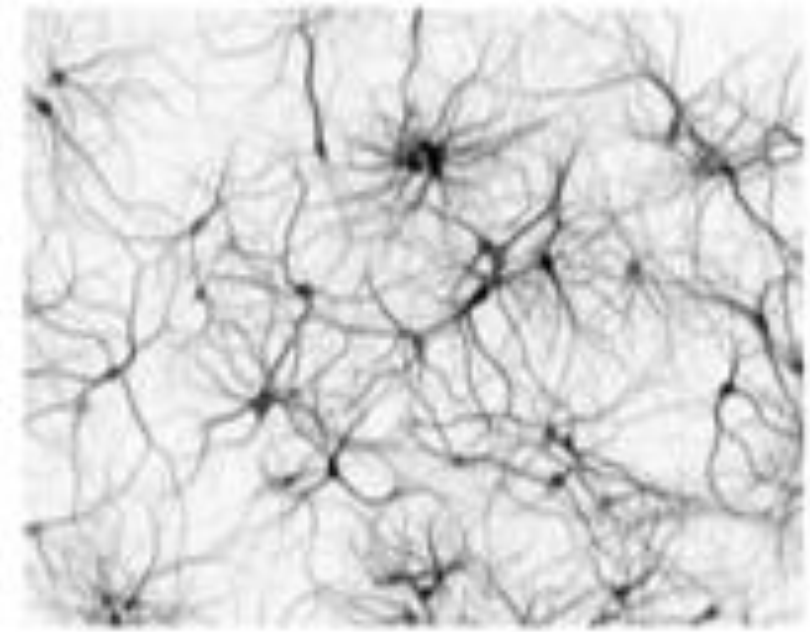
- Basic setup
- Granger network estimation
  - Structural constraints
  - Uncertainty quantification
- General: continuous space
- Other approaches

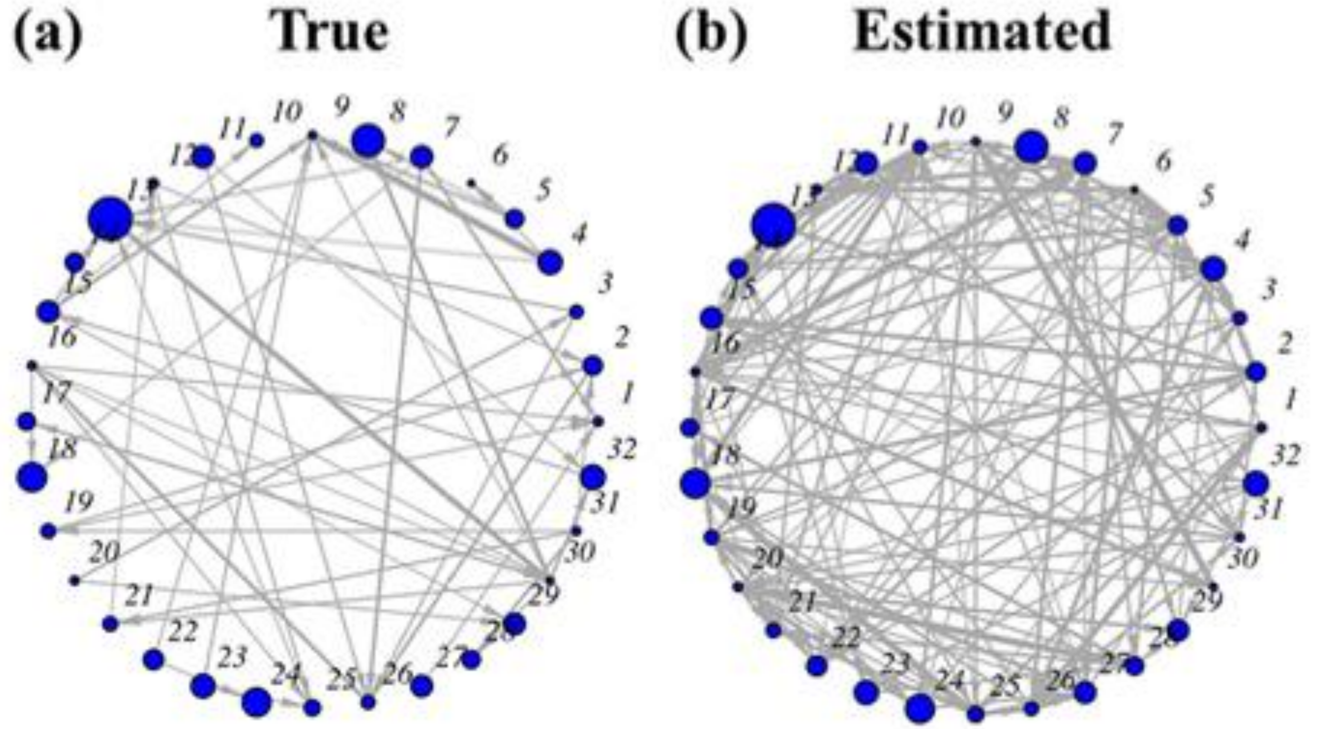
Why need uncertainty quantification?

- Causal inference: with statistical significance there exists an edge?

Granger causality graph $G(U, E)$, then $j \to i \notin E$, iff $\alpha_{ij} = 0$.

Example: Recovery neuronal networks





"Uncertainty quantification for inferring Hawkes networks." Wang, Xie, Cuozzo, Mak, **X.** *NeurIPS* 2020.

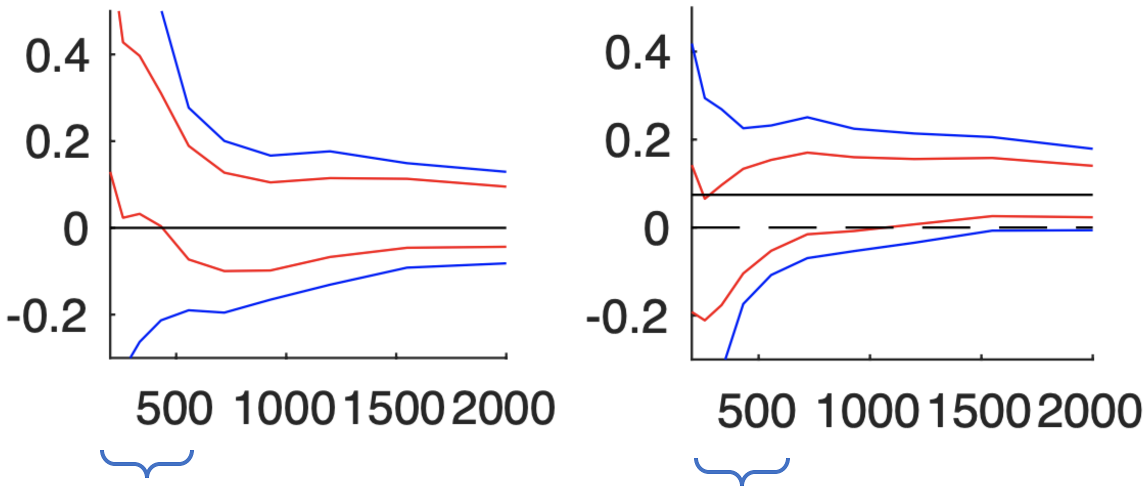
Asymptotic properties of MLE

- (Rathbun 1996) MLE is consistent with asymptotically normal as $T \to \infty$

$$\sqrt{T}(\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i) \to N(0, {I_i^*}^{-1})$$

I_i^* : Fisher information

- How good is asymptotic?



Red: asymptotic CI

Blue: Non-asymptotic CI

Can we do better than asymptotic?

- Challenge:

  Continuous time non-i.i.d. data

$$\lambda(t) = \mu(t) + \text{influence from past}$$

Standard Hoeffding or Bernstein type of concentration bound does not apply

- New approach:
  - Recent advances concentration inequality for continuous-time martingale
  - Develop more precise general **sequential confidence set** adaptive to data

"Time-uniform Chernoff bounds via nonnegative supermartingales", Howard et al., *Prob. Surveys* 2020.

UQ for estimating $\hat{\alpha}_{ij}$: Main idea

- Recall: Delta method (mean-value theorem)

$$S_i(\boldsymbol{\alpha}_i) - \underbrace{S_i(\hat{\boldsymbol{\alpha}}_i)}_{=0} = \underbrace{H_i(\boldsymbol{\alpha}_i')}_{\to T I_i^*}(\boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i)$$

$$\Rightarrow \quad \boldsymbol{\alpha}_i - \hat{\boldsymbol{\alpha}}_i \approx \frac{1}{T} {I_i^*}^{-1} S_i(\boldsymbol{\alpha}_i)$$

score function is a continuous-time martingale

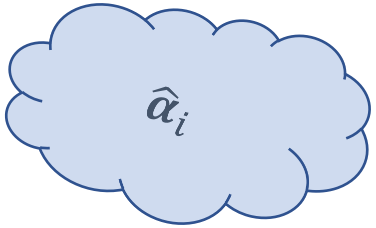
- Concentration bound for entries of ${I_i^*}^{-1} S_i(\boldsymbol{\alpha}_i) \in R^K$

Sequential confidence set

Theorem (Uncertain sets for each $\boldsymbol{\alpha}_i$)

For any $\boldsymbol{\alpha}_i$, $t \in [0, T]$, let $\hat{I}_i(\boldsymbol{\alpha}_i, t)$ be estimator for the Fisher Information given data up to time t. Then

$$C_{i,\varepsilon} = \{\boldsymbol{\alpha}_i \in R^K : g_k(\boldsymbol{\alpha}_i) \le \ln(2K/\varepsilon)\,, k = 1, \dots, 2K\}$$



is a confidence set for $\boldsymbol{\alpha}_i$ at level $1 - \varepsilon$.

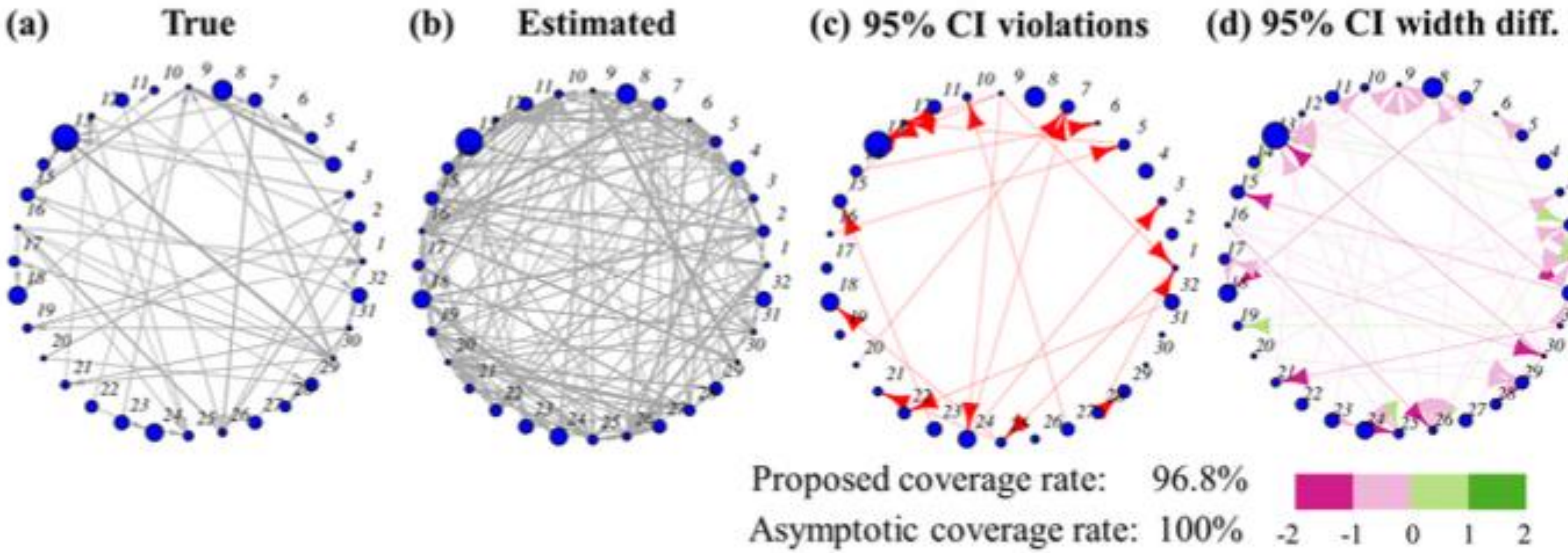
Corollary (Width of confidence interval, asymptotically optimal)

Width of $C_{i,\varepsilon}$ in the direction of $\alpha_{ij} \to 2\sqrt{2\ln(2K/\varepsilon)\sigma_{ij}^2/T}$.

$$g_k(\boldsymbol{\alpha}_i) = \int_0^T \boldsymbol{z}_k^T(H_{t^-}, \boldsymbol{\alpha}_i)\,dS_{i,t}\,(\boldsymbol{\alpha}_i) - V_i(\boldsymbol{z}_k, \boldsymbol{\alpha}_i), \quad \boldsymbol{z}_k\,(H_{t^-}, \boldsymbol{\alpha}_i) \in \left\{ \pm\sqrt{\frac{2\ln(2K/\varepsilon)}{T\boldsymbol{e}_j^T \hat{I}_i^{-1}(\boldsymbol{\alpha}_i, t)\boldsymbol{e}_j}}, j = 1, \dots, K \right\}$$

Intrinsic variance: $V_i(\boldsymbol{z}_k, \boldsymbol{\alpha}_i) = \int_0^T \left(\lambda_i(t)\exp\left(\lambda_i^{-1}(t) z^T \eta_i(t)\right) - z^T \eta_i(t) - \lambda_i(t)\right)dt$

Results



- Asymptotic CI is over-covering
- Non-asymptotic CI achieves targeted coverage and has narrower bandwidth

Roadmap

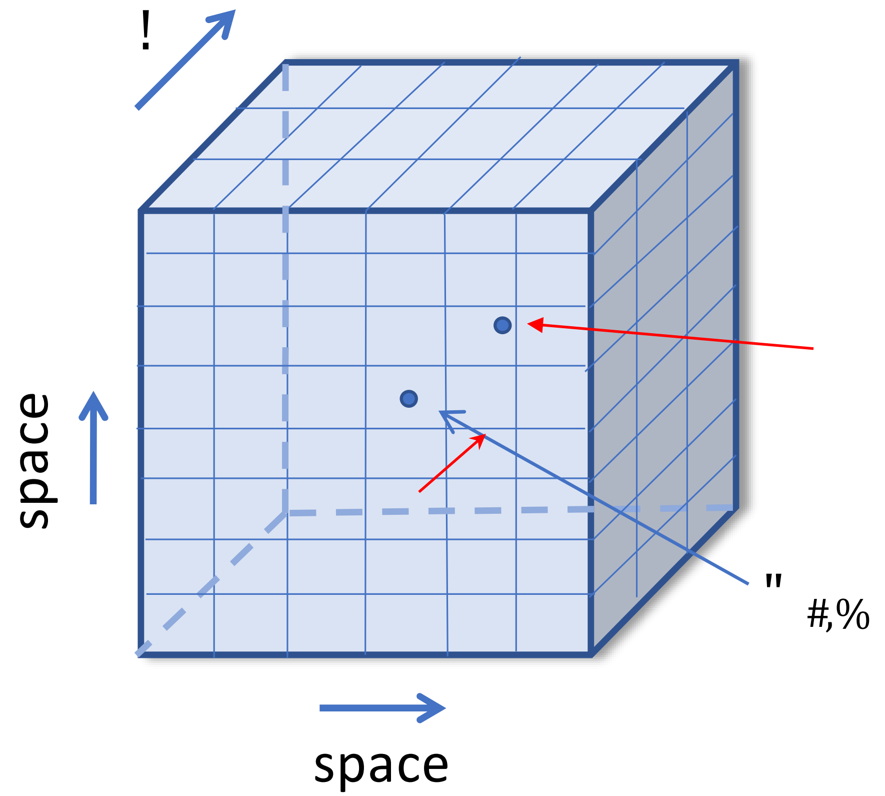
- Basic setup
- Granger network estimation
    - Structural constraints
    - Uncertainty quantification
- General: continuous space
- Other approaches

General influence kernel:
Continuous space, non-stationary

- Events $x_i = (t_i, u_i), u_i \in M$

$$\lambda(x) = \mu(x) + \sum_{x':t'<t} K(x, x')$$

Can we build general model for $K(x', x)$?



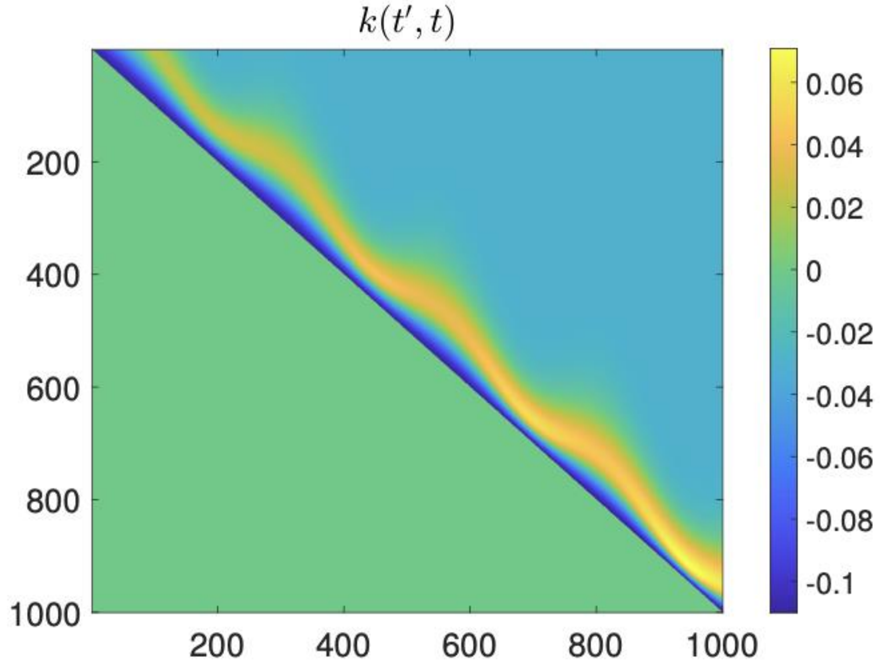
"Neural Spectral Marked Point Processes." Zhu, Wang, Cheng, and **X.** *ICLR 2022.*

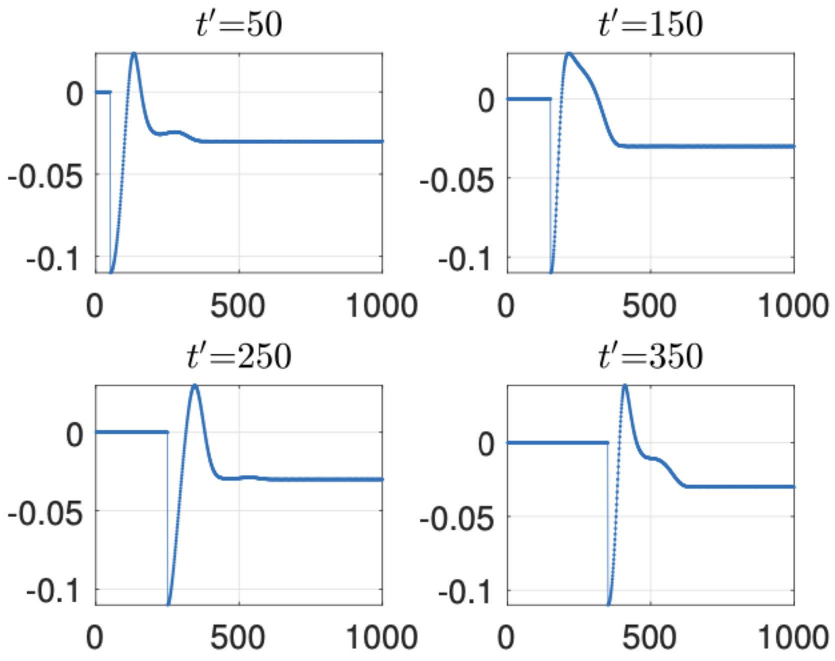
"Spatio-temporal point processes with deep non-stationary kernels". Dong, Cheng, **X.** *ICLR* 2023.

Kernel representation using deep neural networks

- Kernel representation (Mercer's theorem)

$$k(x, x') = \sum_{r=1}^{R} \nu_r \psi_r(x') \phi_r(x)$$



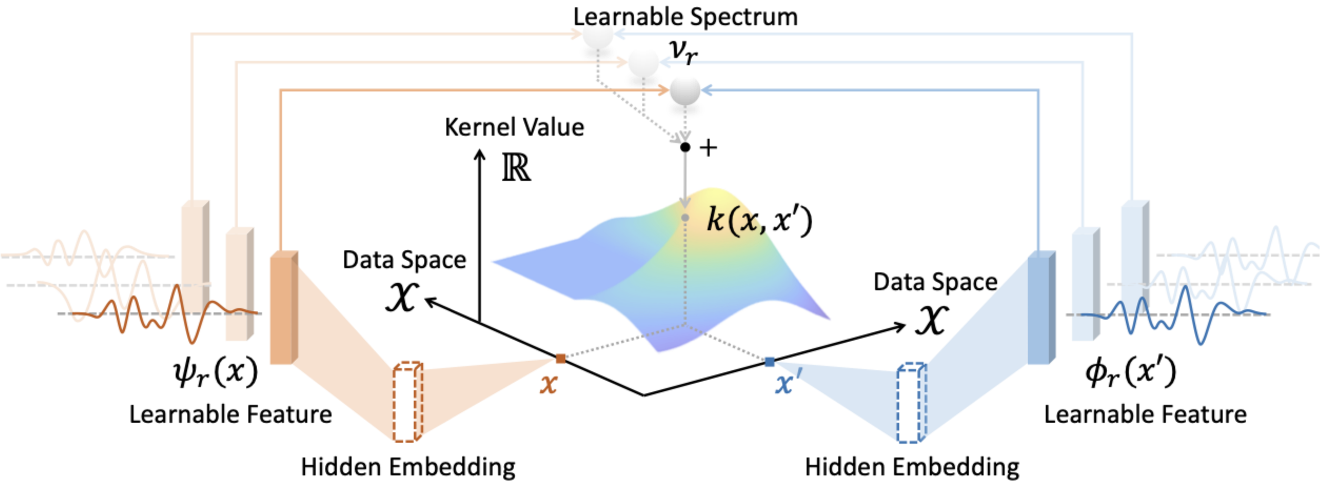


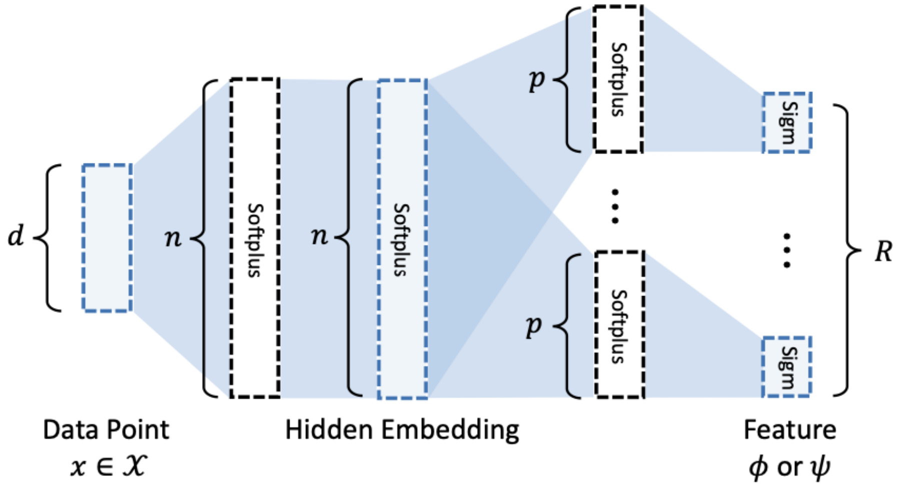
Kernel representation using deep neural networks

- Kernel representation (Mercer's theorem)

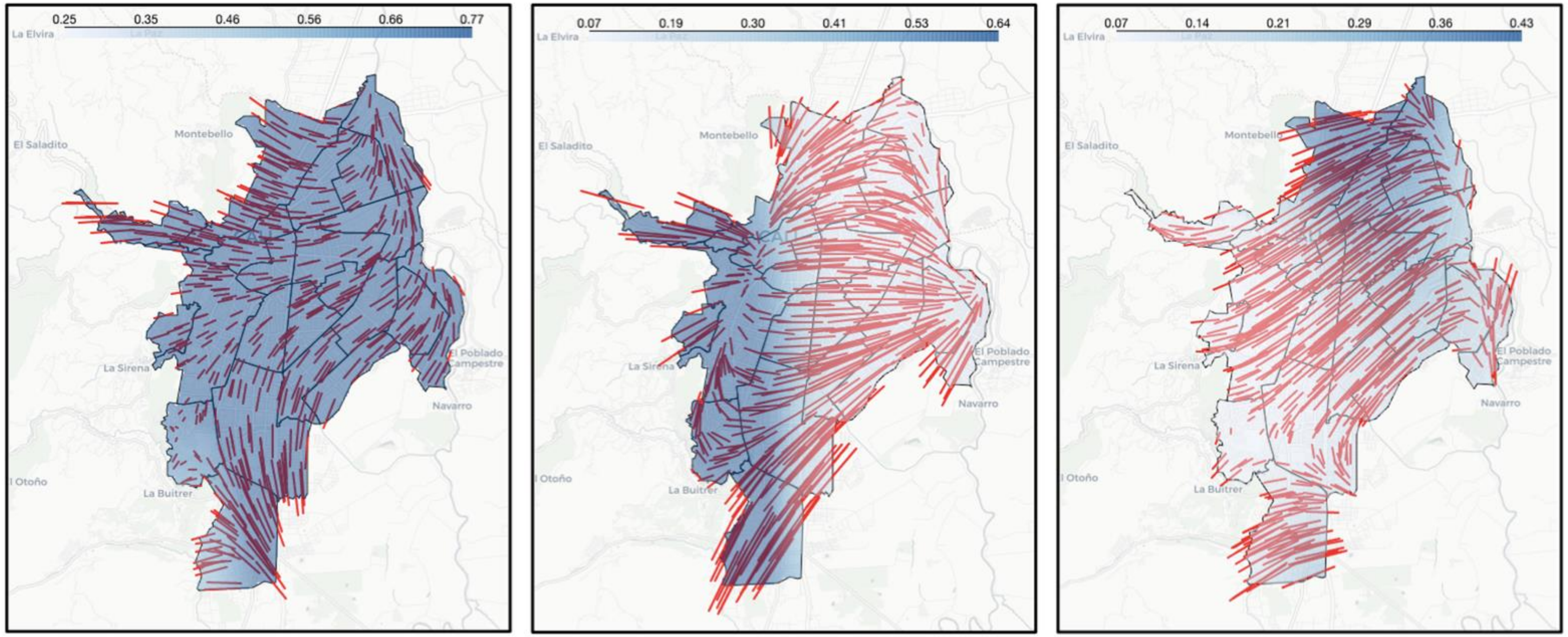
$$k(x, x') = \sum_{r=1}^{R} \nu_r \psi_r(x') \phi_r(x)$$

- Feature maps represented by neural networks

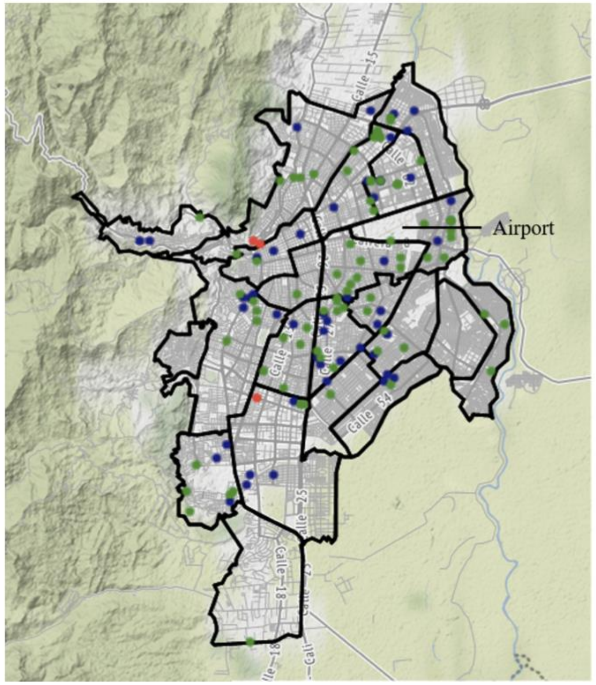




Highly interpretable influence kernels



(a) $\kappa_s^{(1)}$ (b) $\kappa_s^{(2)}$ (c) $\kappa_s^{(3)}$

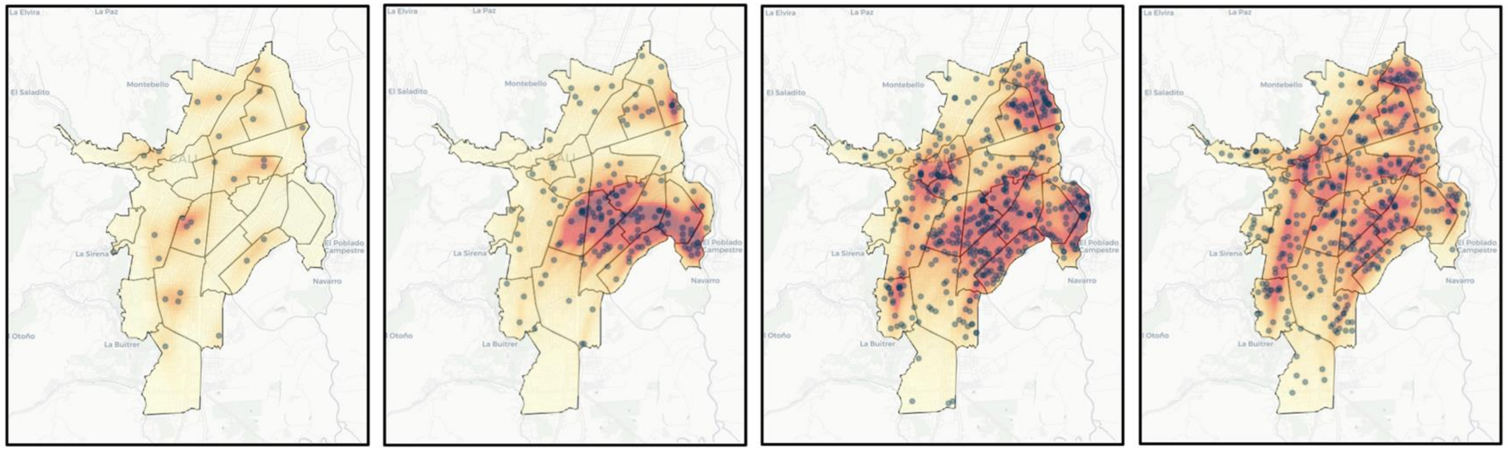


(b) Landmarks

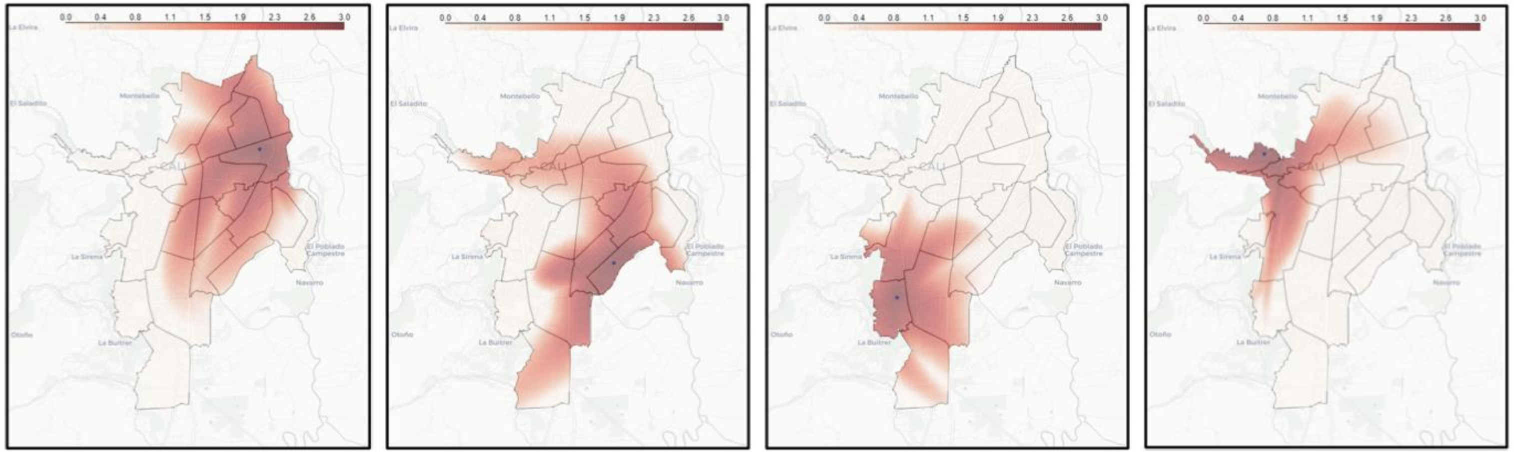
Data: 38,611 cases from March 15 to Sept. 30, 2020.
Exact location of case (residence) and date.

Red: city hall
Blue: church
Green: school

Hotspot prediction



(a) March 22, 2020 (b) May 17, 2020 (c) June 28, 2020 (d) August 30, 2020



(a) Airport (b) Center of Comuna 15 (c) Center of Comuna 18 (d) Center of Comuna 1

Table 2: Out-of-sample estimation performance.

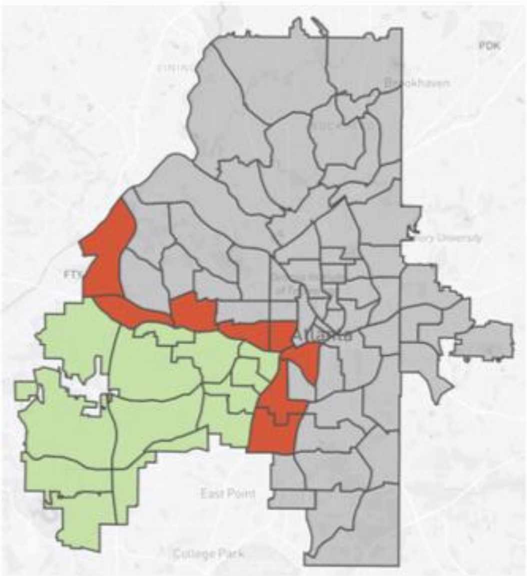
Models	MAE $Q^{\text{out}}_{0.25}$	MAE $Q^{\text{out}}_{0.5}$	MAE $Q^{\text{out}}_{0.75}$
Random	5.190	8.660	14.900
SIR	2.253	5.713	8.554
AR(3)	2.219	**3.776**	8.915
ETAS	4.413	8.234	14.153
NSSTPP−Exo (R=1)	**1.732**	6.051	8.779
NSSTPP−Exo (R=2)	1.962	5.151	8.575
NSSTPP−Exo (R=3)	1.762	5.190	8.342
NSSTPP (R=3)	2.051	4.702	**7.450**

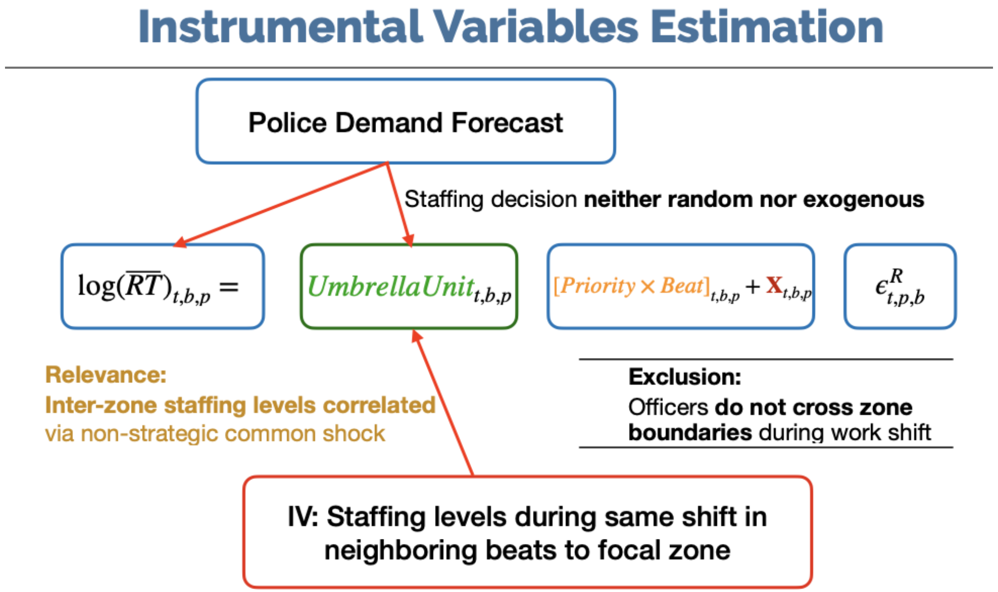
Roadmap

- Basic setup
- Granger network estimation
  - Structural constraints
  - Uncertainty quantification
- General: continuous space
- Other approaches

Police staffing and response time

- How does police staffing contribute to response time disparity and its causal impact on service quality?
- Treatment: staffing, Response: Response time
- Confounding factors: Weather, traffic, service priority….

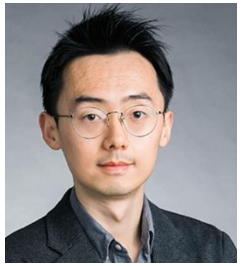


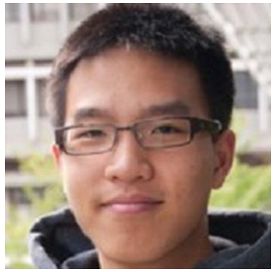


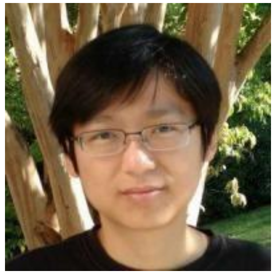
Police staffing and fairness: Evidence from Atlanta Police. Zhou, Xie, Yu. Working paper. 2024.

Summary

- Granger **causal graph** estimation from spatio-temporal discrete events
  - Structural assumptions, uncertainty quantification
- General: **continuous space** ”influence kernel”
- Other possible approaches


Woody Zhu
CMU

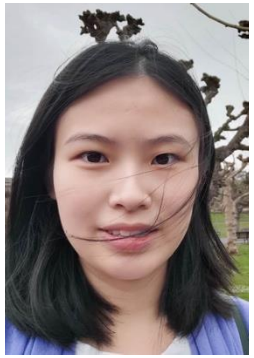

Simon Mak
Duke


Xiuyuan Cheng
Duke


Song Wei
GT


Jonathan Zhou
GT


Zheng Dong
GT


Haoyun Wang
GT


Rishi Kamaleswaran
Duke


Jorge Mateu
UJI Castellon


Qiuping Yu
Georgetown

Main References

1. Causal graph discovery from self and mutually exciting time series. Wei, Xie, Josef, Kamaleswaran. IEEE Selected Areas in Information Theory (JSAIT). 2023.
2. Granger causal chain discovery for sepsis-associated derangements via multivariate Hawkes processes. Wei, Xie, Josef, Kamaleswaran. KDD 2023.
3. Causal structural learning from time series: A convex optimization approach. Wei, Xie. Asilomar 2023.
4. Uncertainty quantification for inferring Hawkes networks. Wang, Xie, Cuozzo, Mak, and Xie. NeurIPS 2020.
5. Spatio-temporal point processes with deep non-stationary kernels. Dong, Cheng, Xie. ICLR 2023.
6. Non-stationary spatio-temporal point process modeling for high-resolution COVID-19 data. Dong, Zhu, Xie, Mateu, Rodriguez-Cortes. Journal of the Royal Statistical Society: Series C. 2023.
7. Police staffing and fairness: Evidence from Atlanta Police. Zhou, Xie, Yu. Working paper. 2024.