Oxford-Warwick Statistics Programme

UNIVERSITY OF OXFORD

Stepping into my PhD research: network models, disclosure risk assessment and a bit of fairness

Francesca Panero

LSE Statistics Research Showcase. 14th - 15th June 2022

Hello!

My “professional” life

- High school with focus on humanities

My “professional” life

- High school with focus on humanities

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino
- A bit of Economics @ Collegio Carlo Alberto

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino
- A bit of Economics @ Collegio Carlo Alberto
- PhD in Stats @ University of Oxford

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino
- A bit of Economics @ Collegio Carlo Alberto
- PhD in Stats @ University of Oxford **(my viva is tomorrow)**

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino
- A bit of Economics @ Collegio Carlo Alberto
- PhD in Stats @ University of Oxford (my viva is tomorrow)
- Visiting period @ Duke (NC)

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino
- A bit of Economics @ Collegio Carlo Alberto
- PhD in Stats @ University of Oxford (my viva is tomorrow)
- Visiting period @ Duke (NC)
- Intern @ Dalle Molle Institute for AI (Lugano, Switzerland) and JP Morgan (London)

My “professional” life

- High school with focus on humanities
- Bachelor and MSc in Maths/Probability @ University of Torino
- A bit of Economics @ Collegio Carlo Alberto
- PhD in Stats @ University of Oxford (my viva is tomorrow)
- Visiting period @ Duke (NC)
- Intern @ Dalle Molle Institute for AI (Lugano, Switzerland) and JP Morgan (London)

Other random facts

- I am from Turin

Other random facts

- I am from Turin

Other random facts

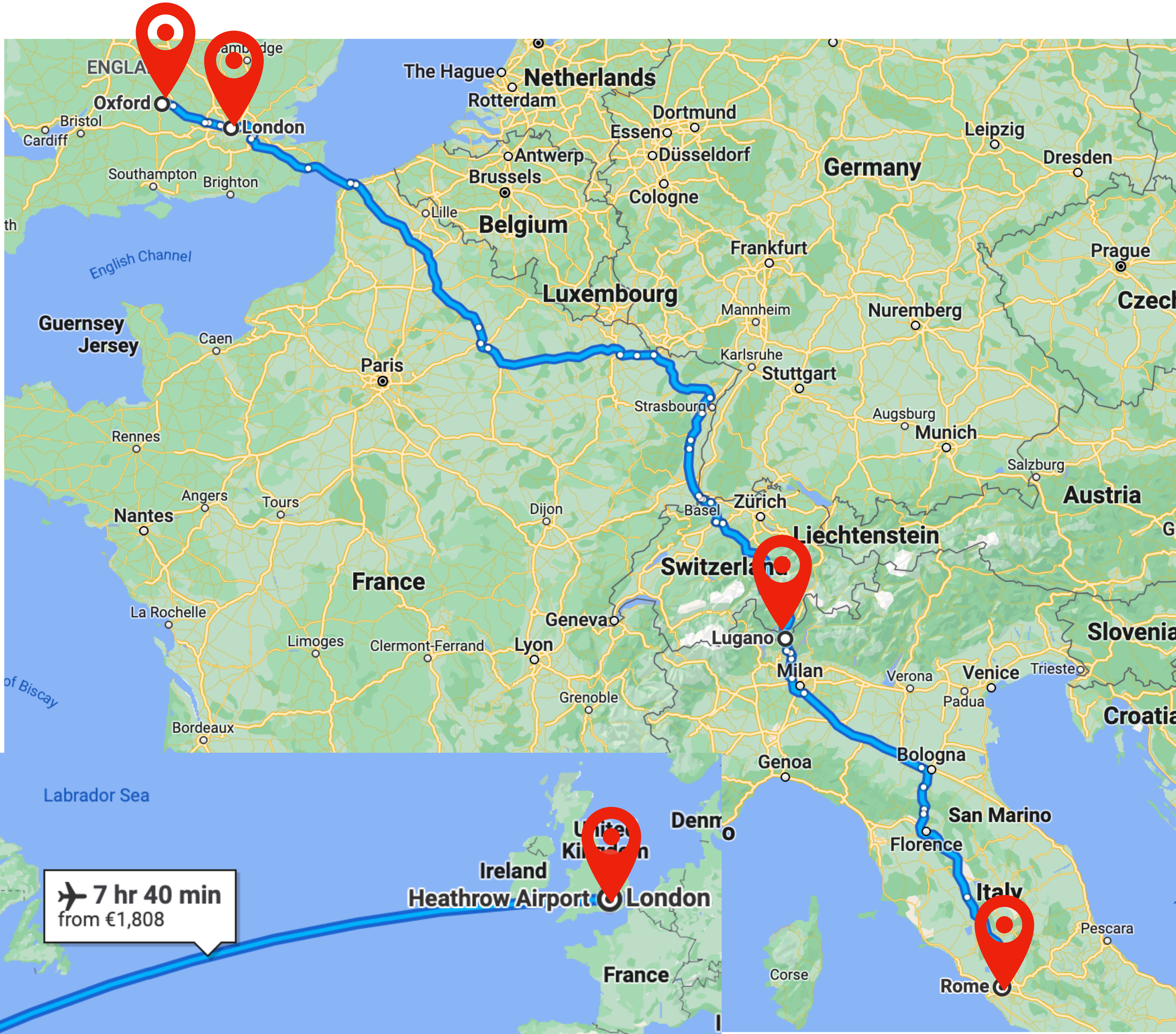
- I am from Turin

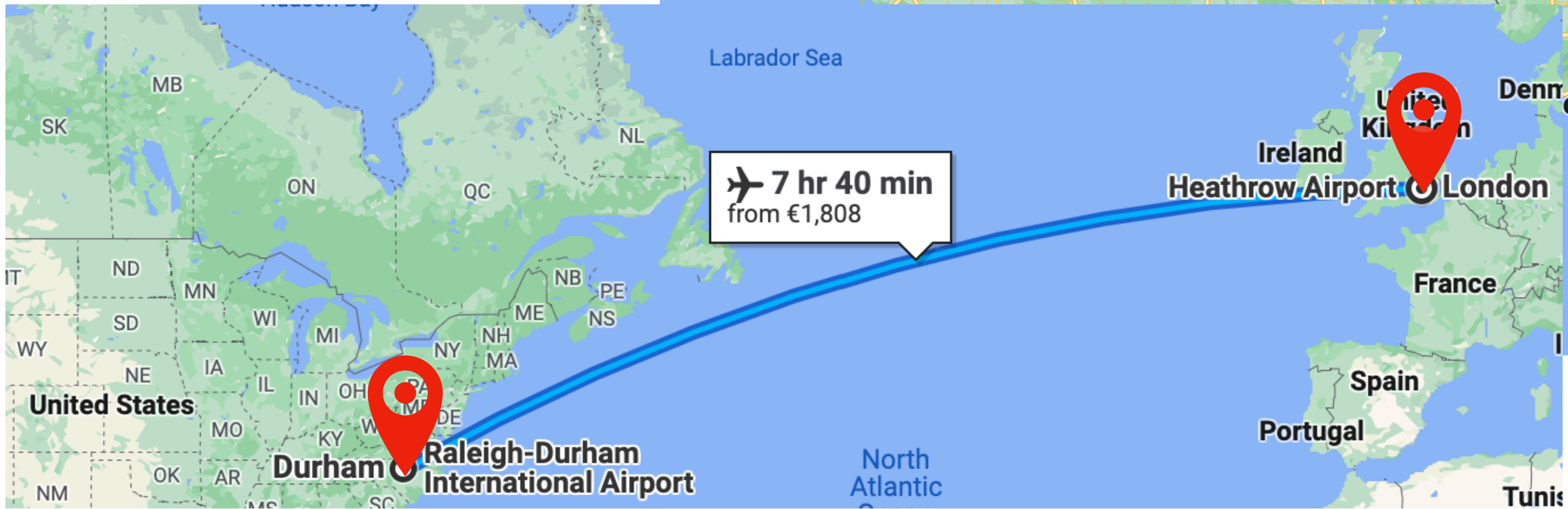
Other random facts

- I am from Turin

Other random facts

- I am from Turin
- I changed 4 countries during the pandemic

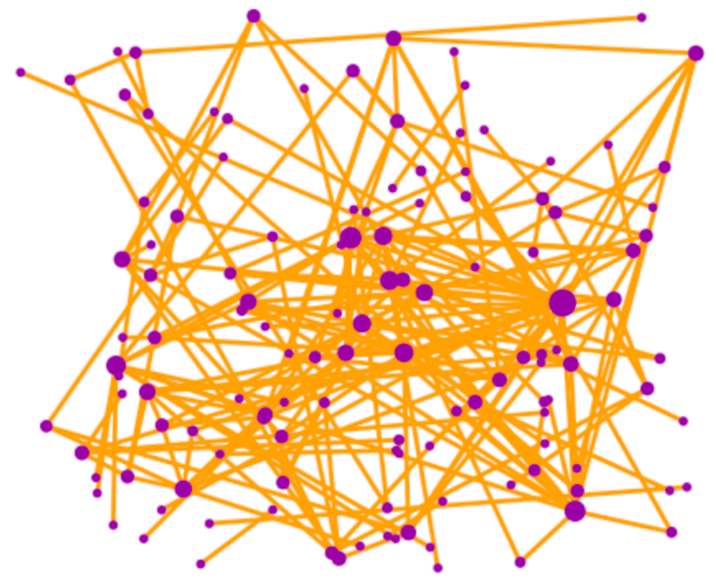




Other random facts

- I am from Turin
- I changed 4 countries during the pandemic
- I like running and yoga
  (but wish I'd do more)
- I like choirs

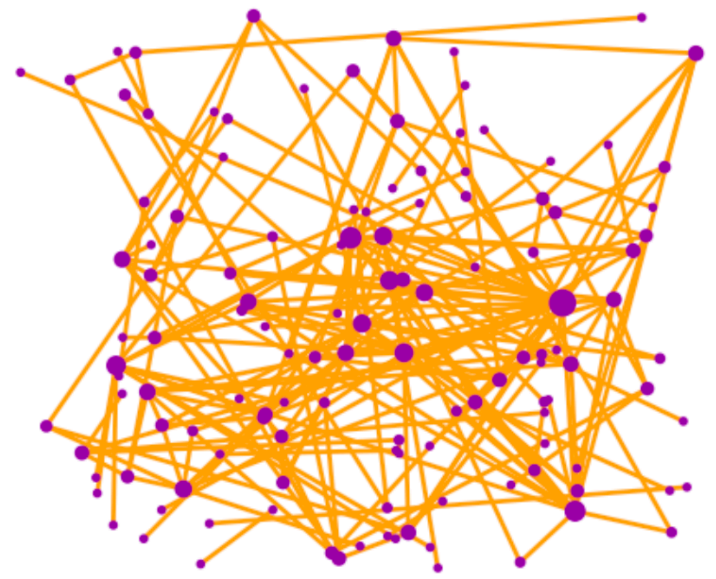
Statistical network models and their properties



- **Sparse spatial random graphs**
  F. Panero, François Caron, Judith Rousseau (ongoing work)

- **On sparsity, power-law and clustering properties of graphex processes**
  François Caron, F. Panero, Judith Rousseau (under revision)

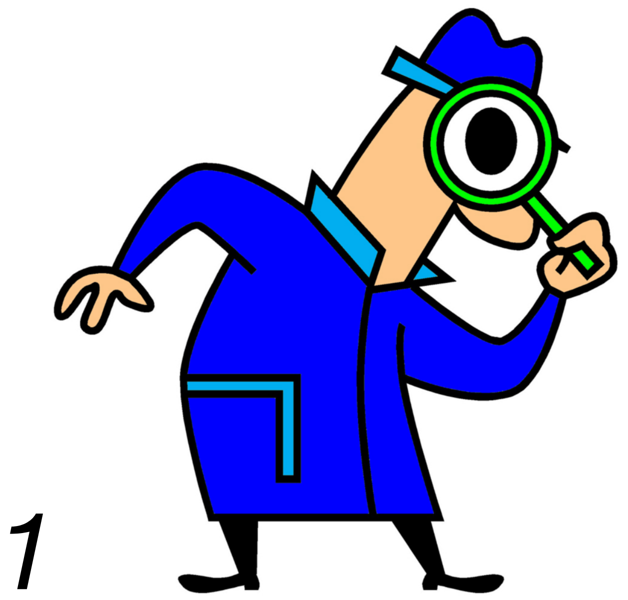
Statistical network models and their properties



- **Sparse spatial random graphs**
  F. Panero, François Caron, Judith Rousseau (ongoing work)

- **On sparsity, power-law and clustering properties of graphex processes**
  François Caron, F. Panero, Judith Rousseau (under revision)

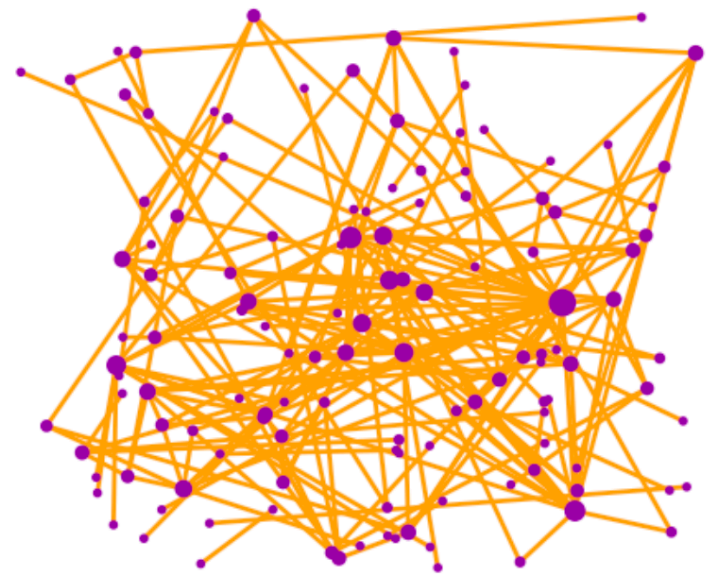
Disclosure risk assessment



- **Bayesian nonparametric disclosure risk assessment.**
  Stefano Favaro, F. Panero, Tommaso Rigon. Electron. J. Stat., 15(2), 5626-5651, 2021

- **Optimal disclosure risk assessment.**
  Federico Camerlenghi, Stefano Favaro, Zacharie Naulet, F. Panero.
  The Annals of Statistics, 49(2) 723-744, April 2021

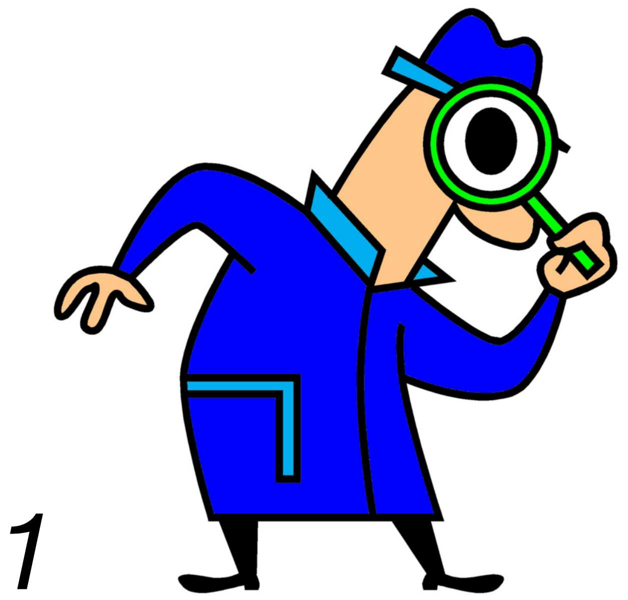
Statistical network models and their properties



- **Sparse spatial random graphs**
  F. Panero, François Caron, Judith Rousseau (ongoing work)

- **On sparsity, power-law and clustering properties of graphex processes**
  François Caron, F. Panero, Judith Rousseau (under revision)

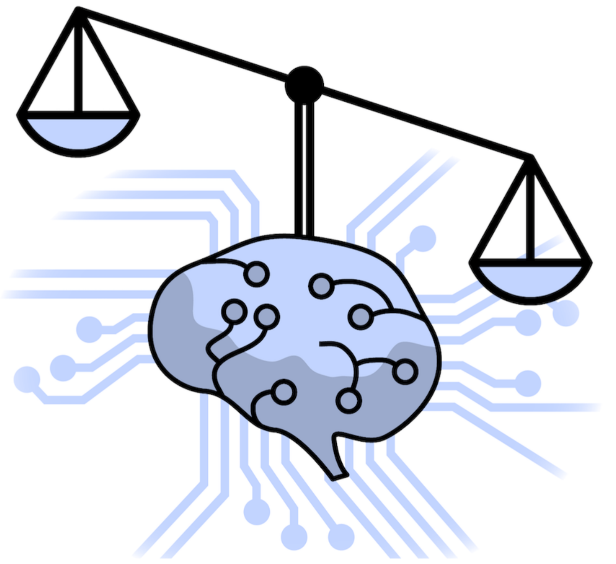
Disclosure risk assessment



- **Bayesian nonparametric disclosure risk assessment.**
  Stefano Favaro, F. Panero, Tommaso Rigon. Electron. J. Stat., 15(2), 5626-5651, 2021

- **Optimal disclosure risk assessment.**
  Federico Camerlenghi, Stefano Favaro, Zacharie Naulet, F. Panero.
  The Annals of Statistics, 49(2) 723-744, April 2021

Fair ML

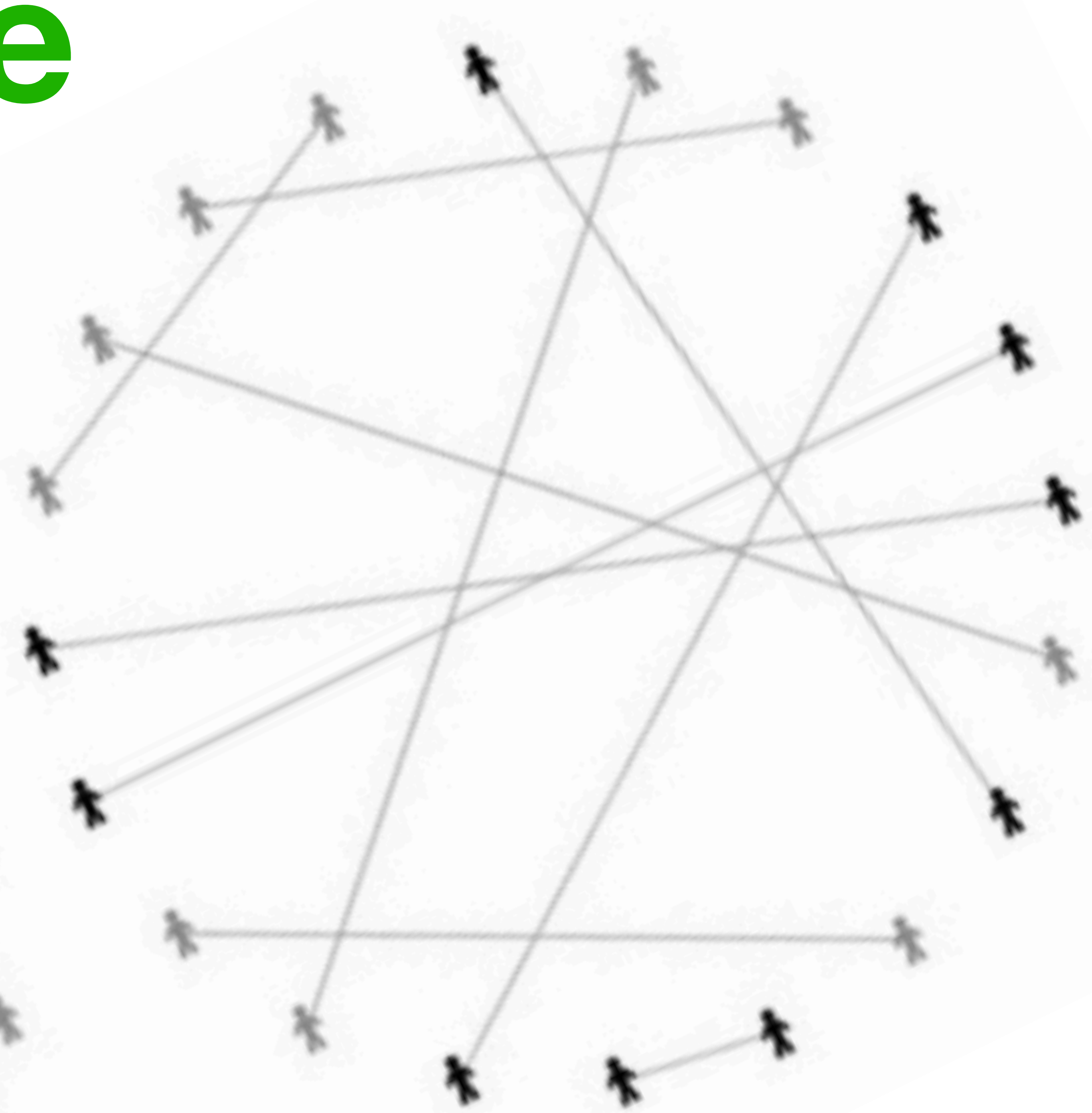


- **Achieving fairness with a simple ridge penalty.**
  Marco Scutari, F. Panero, Manuel Proissl (under revision)

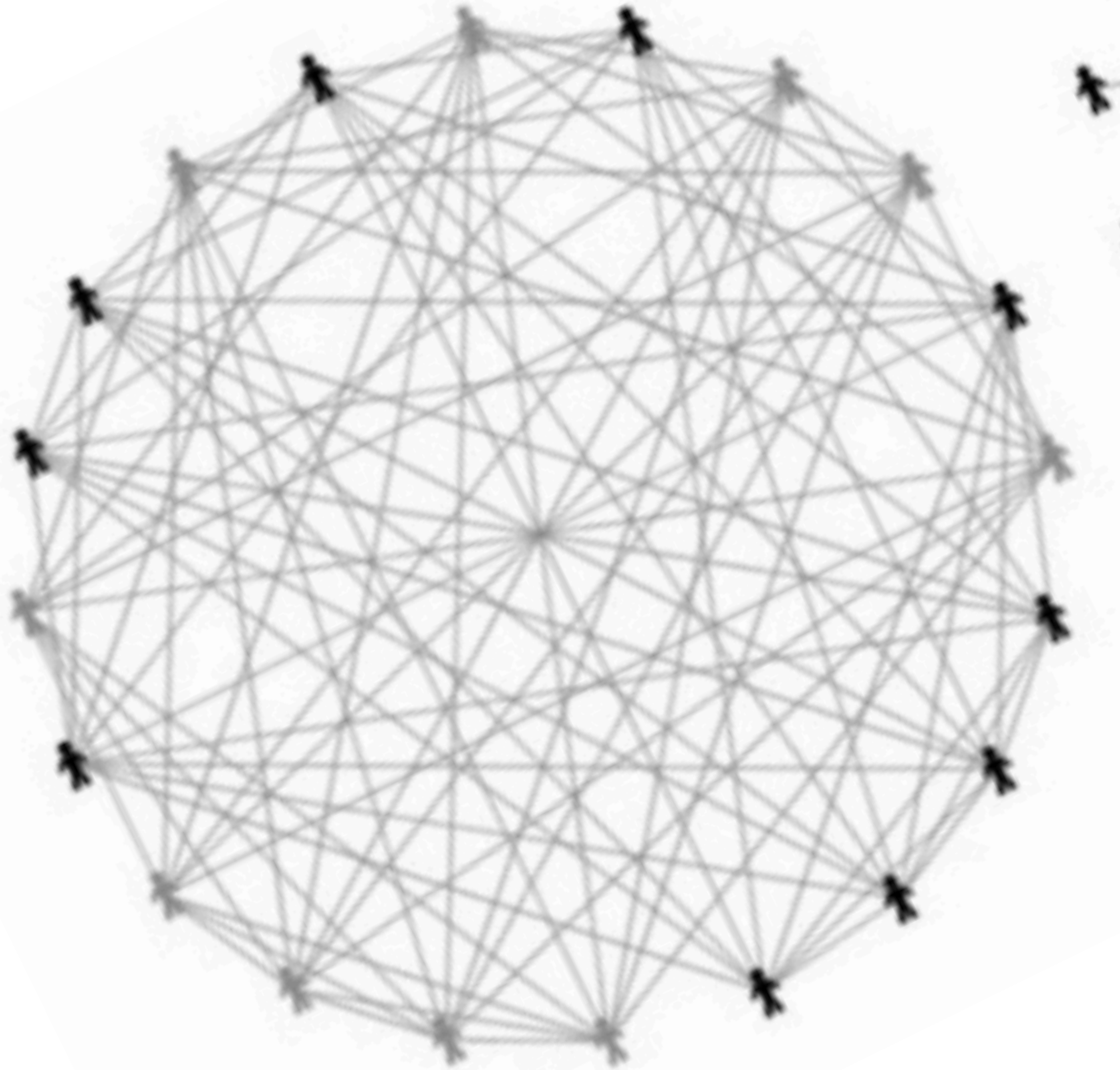
Sparse
Spatial
Random Graphs

Francesca Panero, François Caron, Judith Rousseau

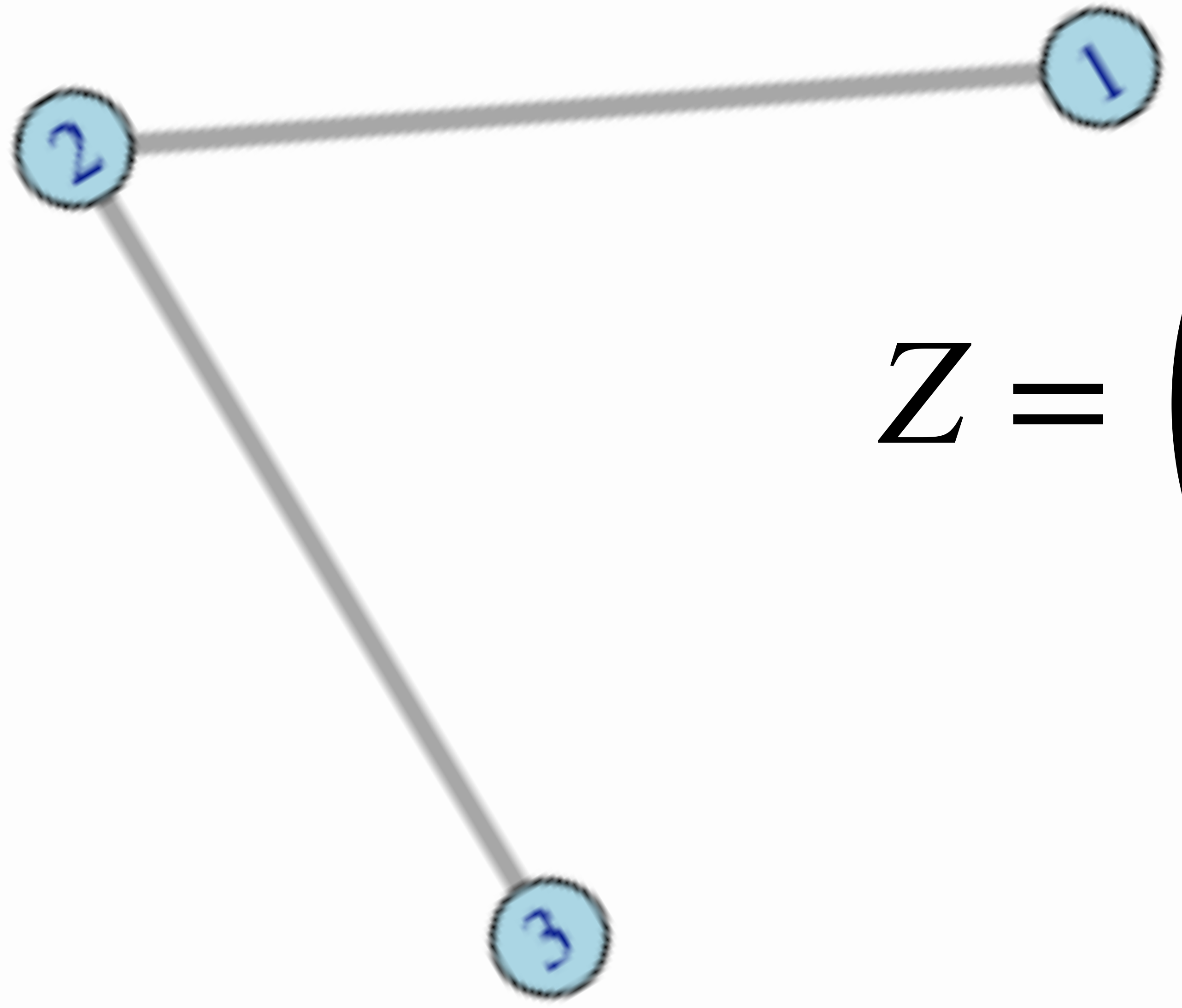
Sparse



Dense

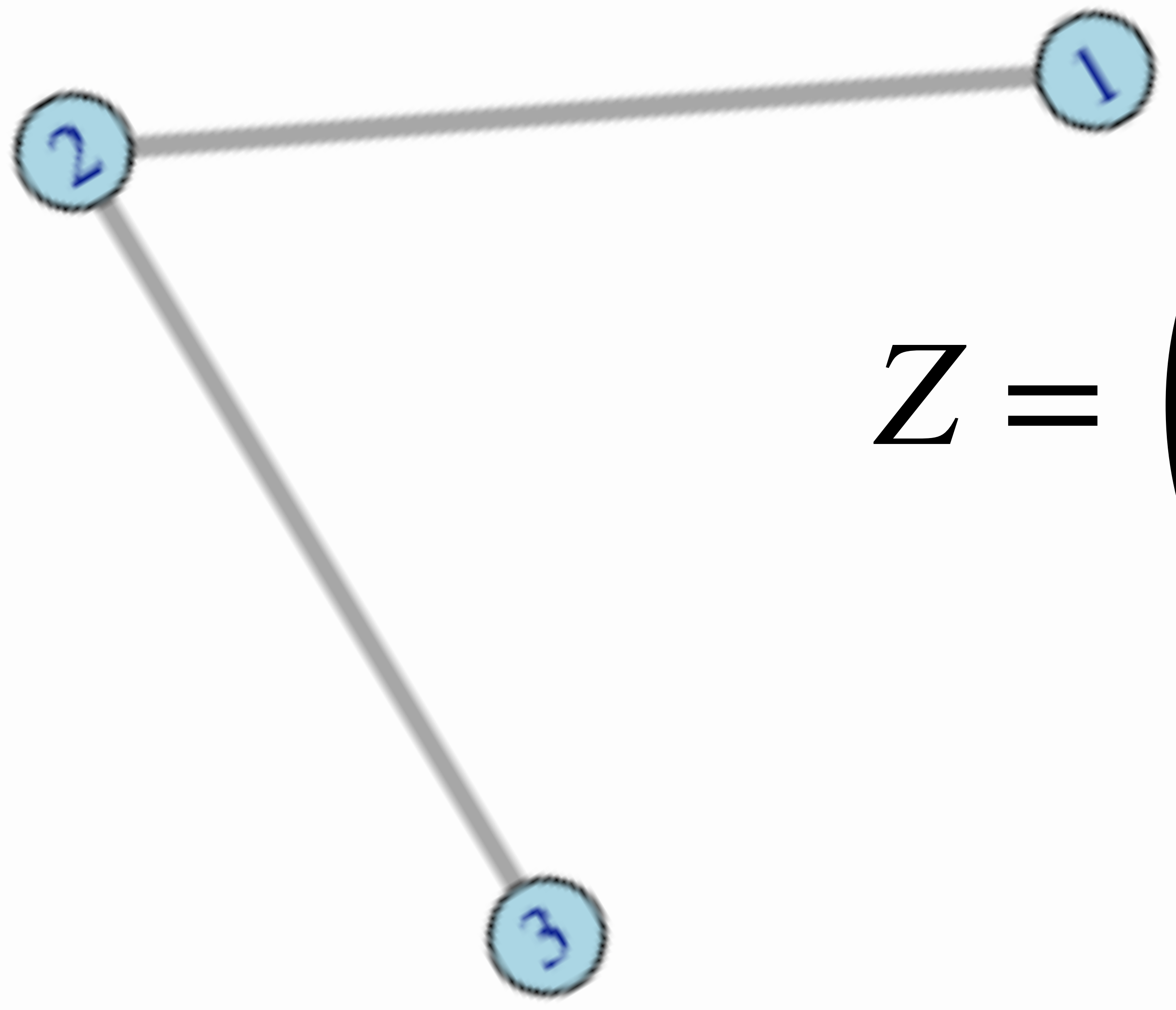


Adjacency matrix



$$Z = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Adjacency matrix



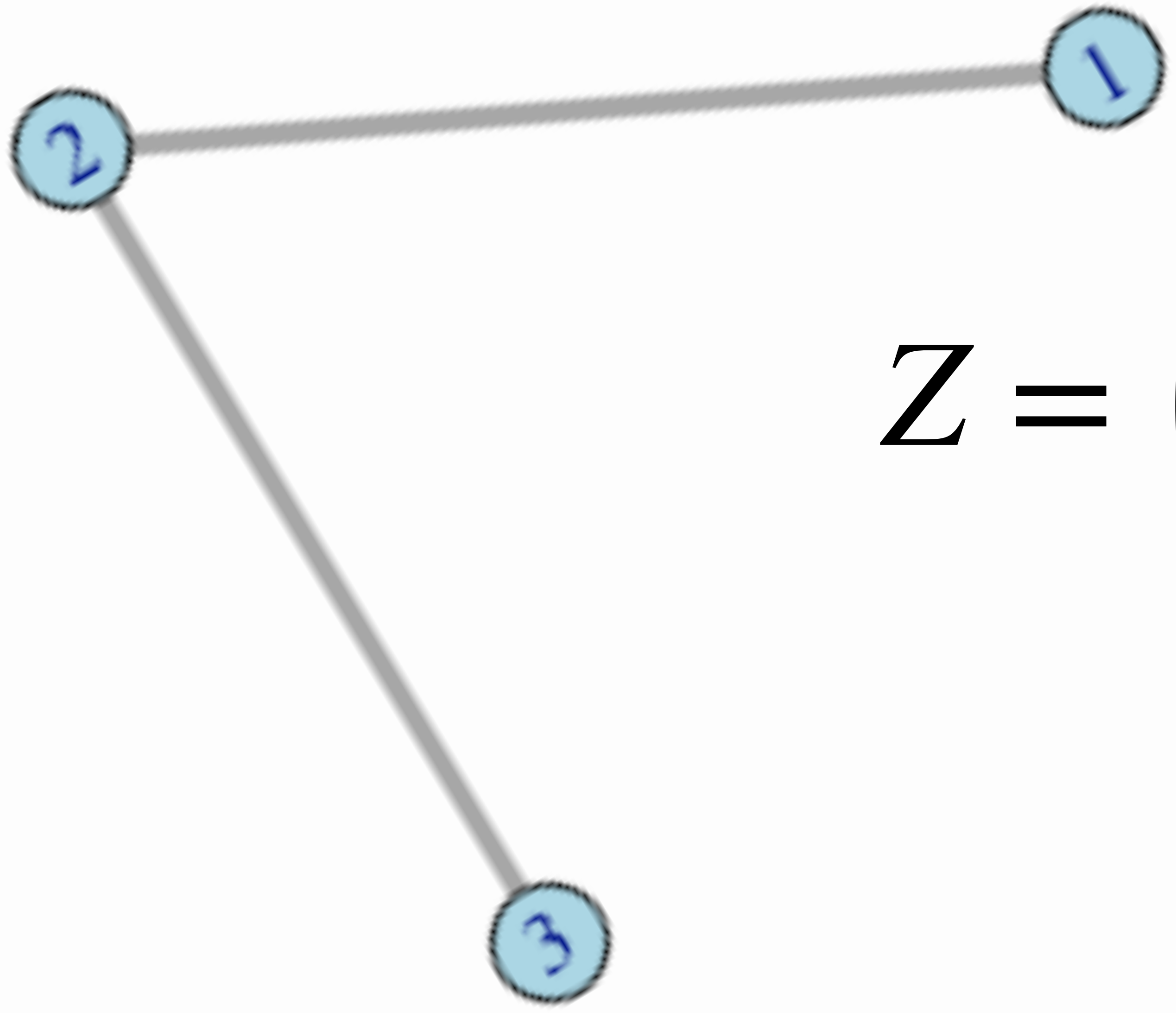
$$Z = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Point process

F. Caron, E. Fox (2017)

$$Z = \sum_{i,j} Z_{ij} \delta_{(\theta_i, \theta_j)}$$

Adjacency matrix



$$Z = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

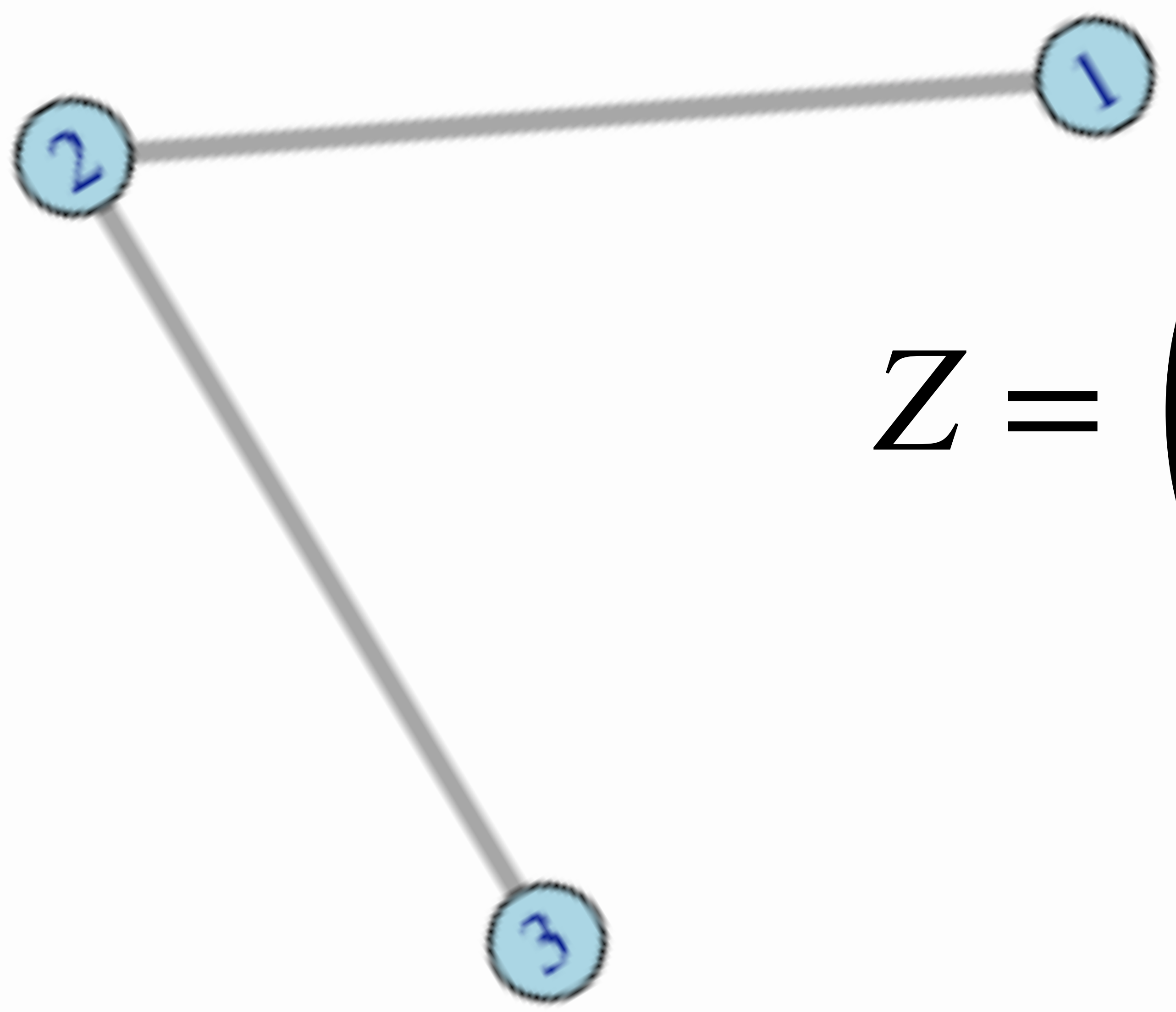
Point process

F. Caron, E. Fox (2017)

Edge $\in \{0,1\}$

$$Z = \sum_{i,j} Z_{ij} \delta_{(\theta_i, \theta_j)}$$

Adjacency matrix



$$Z = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Point process

F. Caron, E. Fox (2017)

$$Z = \sum_{i,j} Z_{ij} \delta_{(\theta_i, \theta_j)}$$

Edge $\in \{0,1\}$

Label ≥ 0

The spatial model

The spatial model

$$Z = \sum_{ij} Z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)}$$

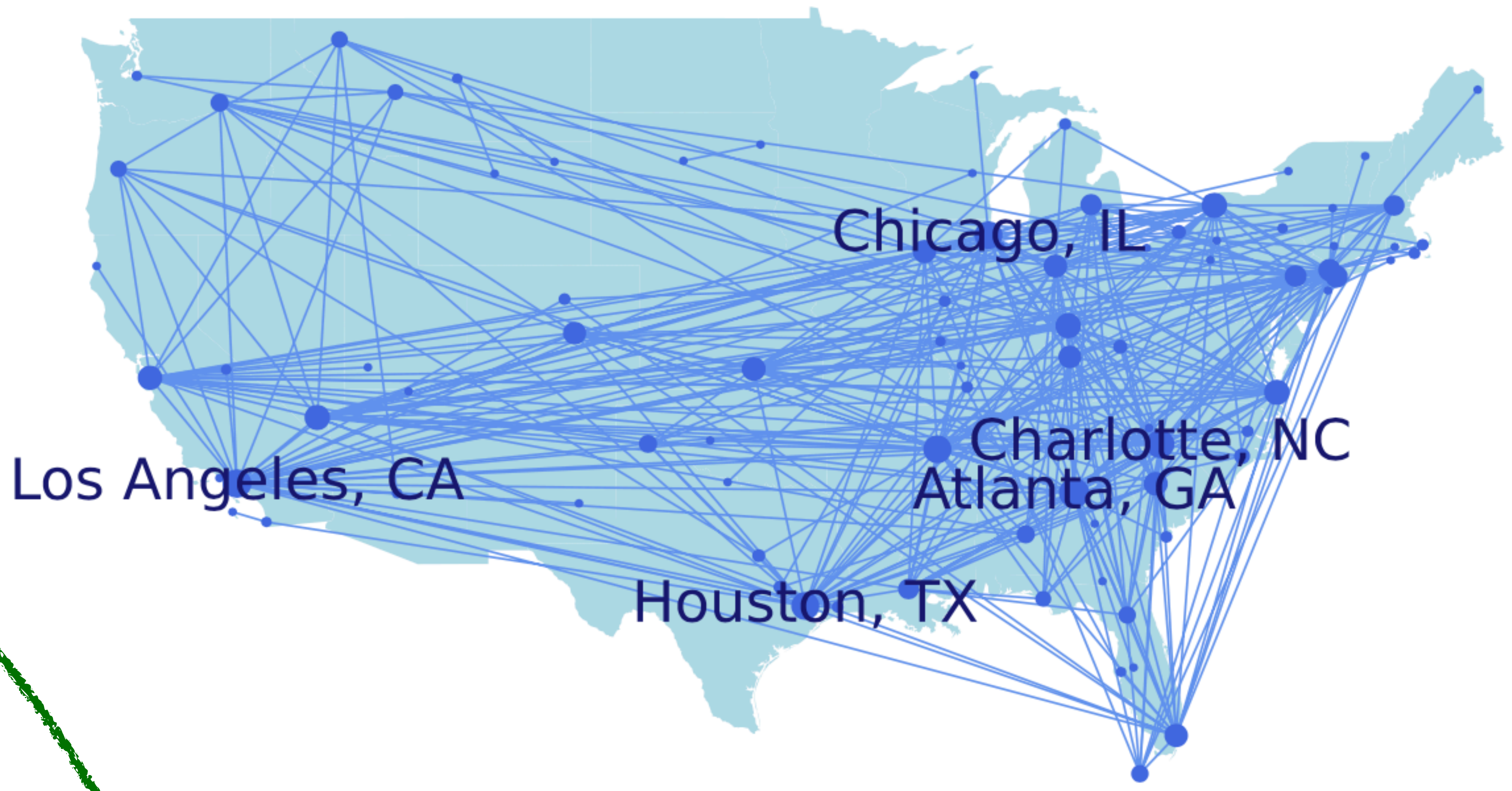
Location

The spatial model

$$Z = \sum_{ij} Z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)}$$

Location

The spatial model



$$Z = \sum_{ij} Z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)}$$

Location

$$Z_{ij} \mid (\theta_k, w_k, x_k)_{k \geq 1} \sim \text{Bernoulli}\left(1 - e^{-\frac{2 w_i w_j}{(1 + |x_i - x_j|)^{\beta}}}\right)$$

The spatial model

$$Z = \sum_{ij} Z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)}$$

Location

$$Z_{ij} \,|\, (\theta_k, w_k, x_k)_{k \geq 1} \sim \text{Bernoulli}\left(1 - e^{-\frac{2 w_i w_j}{(1 + |x_i - x_j|)^{\beta}}}\right)$$

The spatial model

$$Z = \sum_{ij} Z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)}$$

Location

Sociability > 0

$$Z_{ij} \mid (\theta_k, w_k, x_k)_{k \geq 1} \sim \text{Bernoulli}\left(1 - e^{-\frac{2 w_i w_j}{(1 + |x_i - x_j|)^\beta}}\right)$$

The spatial model

$$Z = \sum_{ij} Z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)}$$

Location

Sociability > 0

$$Z_{ij} \mid (\theta_k, w_k, x_k)_{k \geq 1} \sim \text{Bernoulli}\left(1 - e^{-\frac{2 w_i w_j}{(1 + |x_i - x_j|)^{\beta}}}\right)$$

BNP prior inducing…

On sparsity, power-law and clustering properties of graphex processes

François Caron, Francesca Panero, Judith Rousseau

arXiv:1708.03120

Graphex process

Sparse graphon function

$$Z_{ij} \,|\, (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

Graphex process

Sparse graphon function

$$Z_{ij} \,|\, (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

Assumption

$$\mu(\vartheta) := \int_0^{+\infty} W(\vartheta, \vartheta') d\vartheta'$$ Marginal sparse graphon function

Graphex process

Sparse graphon function

$$Z_{ij} \mid (\theta_k, \vartheta_k)_{k=1,2,\dots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

Assumption

$$\mu(\vartheta) := \int_0^{+\infty} W(\vartheta, \vartheta') d\vartheta'$$ Marginal sparse graphon function

$\mu^{-1}(\vartheta) \sim \ell(1/\vartheta)\vartheta^{-\sigma}$ **as** $\vartheta \to 0$

Graphex process

Sparse graphon function

$$Z_{ij} \mid (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

Assumption

$$\mu(\vartheta) := \int_0^{+\infty} W(\vartheta, \vartheta')d\vartheta'$$ Marginal sparse graphon function

$$\mu^{-1}(\vartheta) \sim \ell(1/\vartheta)\vartheta^{-\sigma} \textbf{ as } \vartheta \to 0$$

Graphex process

Sparse graphon function

$$Z_{ij} \mid (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

Assumption

$$\mu(\vartheta) := \int_0^{+\infty} W(\vartheta, \vartheta')d\vartheta'$$ Marginal sparse graphon function

$$\mu^{-1}(\vartheta) \sim \ell(1/\vartheta)\vartheta^{-\sigma} \textbf{ as } \vartheta \to 0$$

Graphex process

Sparse graphon function

$$Z_{ij} \mid (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

Assumption

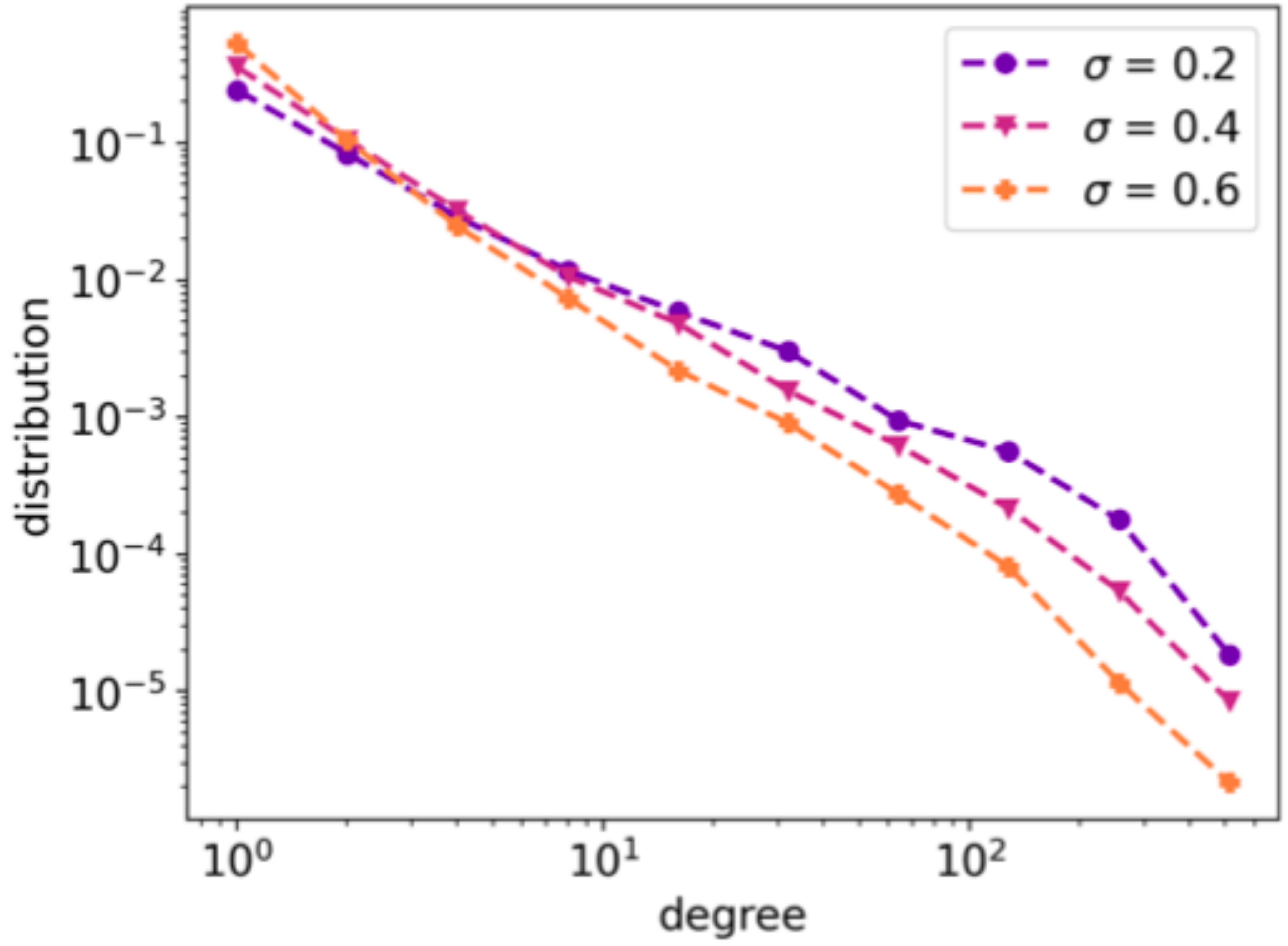
$$\mu(\vartheta) := \int_0^{+\infty} W(\vartheta, \vartheta')d\vartheta'$$ Marginal sparse graphon function

$\mu^{-1}(\vartheta) \sim \ell(1/\vartheta)\vartheta^{-\sigma}$ **as** $\vartheta \to 0$ **Regular variation at zero,** $\sigma \in [0,1]$

Results

- $\sigma = 0$ Dense graph
- $\sigma \in (0,1)$ Sparse graph + power-law degree distribution

Results



- $\sigma = 0$ Dense graph
- $\sigma \in (0,1)$ Sparse graph + power-law degree distribution

Results

- $\sigma = 0$ Dense graph
- $\sigma \in (0,1)$ Sparse graph + power-law degree distribution
- Strictly positive global clustering coefficient

Results

- $\sigma = 0$ Dense graph
- $\sigma \in (0,1)$ Sparse graph + power-law degree distribution
- Strictly positive global clustering coefficient

Results

- $\sigma = 0$ Dense graph
- $\sigma \in (0,1)$ Sparse graph + power-law degree distribution
- Strictly positive global clustering coefficient
- Central limit theorems for number of nodes and subgraphs

Optimal disclosure risk assessment

Federico Camerlenghi, Stefano Favaro, Zacharie Naulet, Francesca Panero

The Annals of Statistics (2021)

Disclosure risk

Sample

Gender	# Kids	Education	Residence
F	1	Degree	Oxford
M	7	PhD	Birmingham
F	1	Degree	Oxford
F	1	Degree	Oxford
F	3	Diploma	Manchester

Disclosure risk

Sample

	Gender	# Kids	Education	Residence
✦	F	1	Degree	Oxford
❀	M	7	PhD	Birmingham
✦	F	1	Degree	Oxford
✦	F	1	Degree	Oxford
*	F	3	Diploma	Manchester

Disclosure risk

Sample

	Gender	# Kids	Education	Residence
✦	F	1	Degree	Oxford
❀	M	7	PhD	Birmingham
✦	F	1	Degree	Oxford
✦	F	1	Degree	Oxford
*	F	3	Diploma	Manchester

Disclosure risk

Sample

	Gender	# Kids	Education	Residence
✦	F	1	Degree	Oxford
❀	M	7	PhD	Birmingham
✦	F	1	Degree	Oxford
✦	F	1	Degree	Oxford
✱	F	3	Diploma	Manchester

✱

⌘

Population

Disclosure risk

Sample

Gender	# Kids	Education	Residence
F	1	Degree	Oxford
M	7	PhD	Birmingham
F	1	Degree	Oxford
F	1	Degree	Oxford
F	3	Diploma	Manchester

Population

Disclosure risk

Sample

	Gender	# Kids	Education	Residence
✦	F	1	Degree	Oxford
❀	M	7	PhD	Birmingham
✦	F	1	Degree	Oxford
✦	F	1	Degree	Oxford
✱	F	3	Diploma	Manchester

✱

⌘

Population

τ_1: sample uniques that are also population uniques

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

Size of rest of the population: $M = \lambda n$

Model and estimator

$(X_1, \ldots, X_n)$ Sample

Size of rest of the population: $M = \lambda n$

$\lambda = 1/2 \rightarrow M = n/2$

Sample

Rest of population

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

Size of rest of the population: $M = \lambda n$

$\lambda = 1 \rightarrow M = n$

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

Size of rest of the population: $M = \lambda n$

$\lambda = 2 \rightarrow M = 2n$

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

Size of rest of the population: $M = \lambda n$

$$\hat{\tau}_1^L = \sum_{i \geq 0} (-1)^i (i+1) \lambda^i Z_{i+1}(X_1, \ldots, X_n) \mathbb{P}(L \geq i)$$

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

Size of rest of the population: $M = \lambda n$

symbols with frequency $i + 1$

$$\hat{\tau}_1^L = \sum_{i \geq 0} (-1)^i (i+1) \lambda^i Z_{i+1}(X_1, \ldots, X_n) \mathbb{P}(L \geq i)$$

Model and estimator

$(X_1, \ldots, X_n)$ Sample

Size of rest of the population: $M = \lambda n$

symbols with frequency $i + 1$

$$\hat{\tau}_1^L = \sum_{i \geq 0} (-1)^i (i+1) \lambda^i Z_{i+1}(X_1, \ldots, X_n) \mathbb{P}(L \geq i)$$

Truncation random variable

Results

- Upper bound for worst-case normalised MSE of $\hat{\tau}_1^L$ goes to 0 for $\lambda < \log(n)$

Results

- Upper bound for worst-case normalised MSE of $\hat{\tau}_1^L$ goes to 0 for $\lambda < \log(n)$
- Lower bound for best worst-case normalised MSE of any nonparametric estimator vanishes for $\lambda < \log(n)$

Results

- Upper bound for worst-case normalised MSE of $\hat{\tau}_1^L$ goes to 0 for $\lambda < \log(n)$
- Lower bound for best worst-case normalised MSE of any nonparametric estimator vanishes for $\lambda < \log(n)$
- For $\lambda > \log(n)$ it is impossible to find a nonparametric estimator with vanishing lower bound

Results

Up until $\lambda \propto \log(n)$ the lower and upper bound match for $\hat{\tau}_1^L$ +
impossible to find nonparametric estimator with guarantees after $\log(n)$:
$\hat{\tau}_1^L$ is optimal!

Results

Up until $\lambda \propto \log(n)$ the lower and upper bound match for $\hat{\tau}_1^L$ +
impossible to find nonparametric estimator with guarantees after $\log(n)$:
$\hat{\tau}_1^L$ is optimal!

Dedicated to the memory of Chris Skinner

Bayesian nonparametric disclosure risk assessment

Stefano Favaro, Francesca Panero and Tommaso Rigon

Electronic Journal of Statistics (2021)

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

$p_1 = \mathbb{P}(✦), p_2 = \mathbb{P}(❀), p_3 = \mathbb{P}(✱)\ldots$

Model and estimator

$(X_1, \ldots, X_n)$ ✦ ❀ ✦ ✦ ✱ Sample

$p_1 = \mathbb{P}($✦$), p_2 = \mathbb{P}($❀$), p_3 = \mathbb{P}($✱$)\ldots$

Dirichlet process $\alpha = 0$

Pitman-Yor process prior $P_{\alpha,\theta}$

$P_{\alpha,\theta} = \sum p_i \delta_{z_i}$ $\rightarrow p_{(1)}, p_{(2)}, p_{(3)} \ldots$ decreasing order

$p_{(j)}$ as $j \rightarrow \infty$ have power-law behaviour with exponent α^{-1}

exponential decay for $\alpha = 0$

Posterior characterisation

$$\mathbb{P}(\tau_1 = x \mid X_1, \ldots, X_n) = \sum_{u=1}^{N-n} \frac{\binom{\frac{\theta+n}{1-\alpha}-1}{x}\binom{u}{m_1-x}}{\binom{\frac{\theta+n}{1-\alpha}-1+u}{m_1}} \mathbb{P}(U_{1-\alpha, \frac{\theta+n}{1-\alpha}, N-n} = u)$$

Posterior characterisation

MIXTURE!

$$\mathbb{P}(\tau_1 = x \,|\, X_1, \ldots, X_n) = \sum_{u=1}^{N-n} \frac{\binom{\frac{\theta+n}{1-\alpha}-1}{x}\binom{u}{m_1-x}}{\binom{\frac{\theta+n}{1-\alpha}-1+u}{m_1}} \mathbb{P}(U_{1-\alpha,\frac{\theta+n}{1-\alpha},N-n} = u)$$

General hypergeometric distribution

Generalised factorial distribution

Posterior characterisation

MIXTURE!

$$\mathbb{P}(\tau_1 = x \mid X_1, \dots, X_n) = \sum_{u=1}^{N-n} \frac{\binom{\frac{\theta+n}{1-\alpha}-1}{x}\binom{u}{m_1-x}}{\binom{\frac{\theta+n}{1-\alpha}-1+u}{m_1}} \mathbb{P}(U_{1-\alpha, \frac{\theta+n}{1-\alpha}, N-n} = u)$$

General hypergeometric distribution

Generalised factorial distribution

Works well in the case of power-law or exponential decaying probabilities

Achieving fairness with a simple ridge penalty

Marco Scutari, Francesca Panero, Manuel Proissl (2021)

arXiv:2105.13817

What's next?

Networks

Disclosure risk assessment

Fair ML

Else

What's next?

Networks

- Brain networks
- Other extension of Caron-Fox

Disclosure risk assessment

Fair ML

Else

What's next?

Networks

- Brain networks
- Other extension of Caron-Fox

Disclosure risk assessment

- Finding motivation to work on disclosure risk. Possibly different measures?

Fair ML

Else

What's next?

Networks

- Brain networks
- Other extension of Caron-Fox

Disclosure risk assessment

- Finding motivation to work on disclosure risk. Possibly different measures?

Fair ML

- Waiting…

Else

What's next?

Networks

- Brain networks
- Other extension of Caron-Fox

Disclosure risk assessment

- Finding motivation to work on disclosure risk. Possibly different measures?

Fair ML

- Waiting…

Else

- More applied
- ED&I

Thank you!

francesca.panero@stats.ox.ac.uk
https://francescapanero.github.io