Disentangling the structure of ecological bipartite networks from observation processes

Pierre BARBILLON
Joint work with E. Anakok, C. Fontaine, E. Thébault

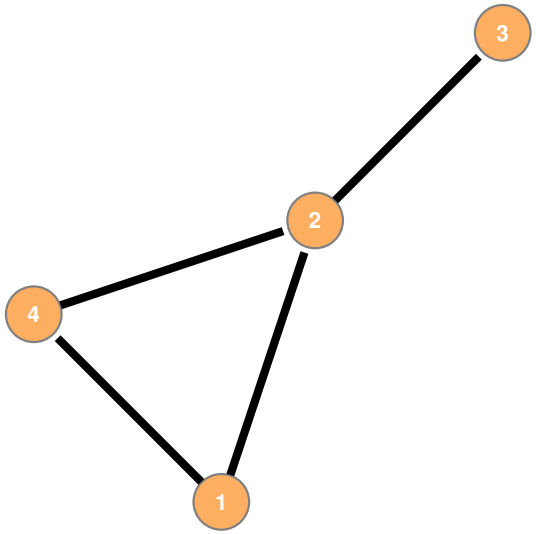
Rochebrune 2026

Outline

Outline

Networks

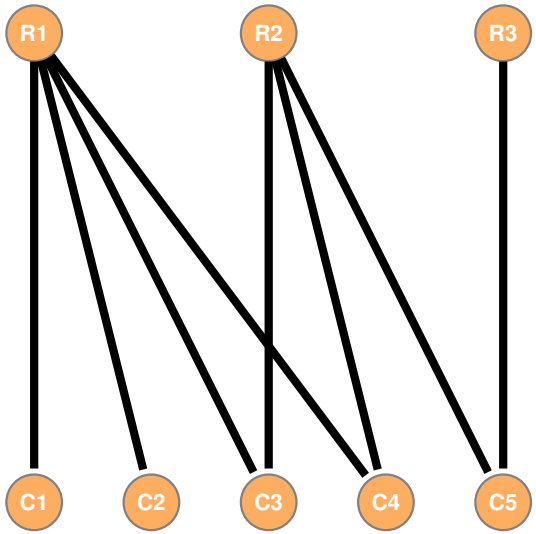
Simple network



Adjacency matrix:

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

Bipartite network

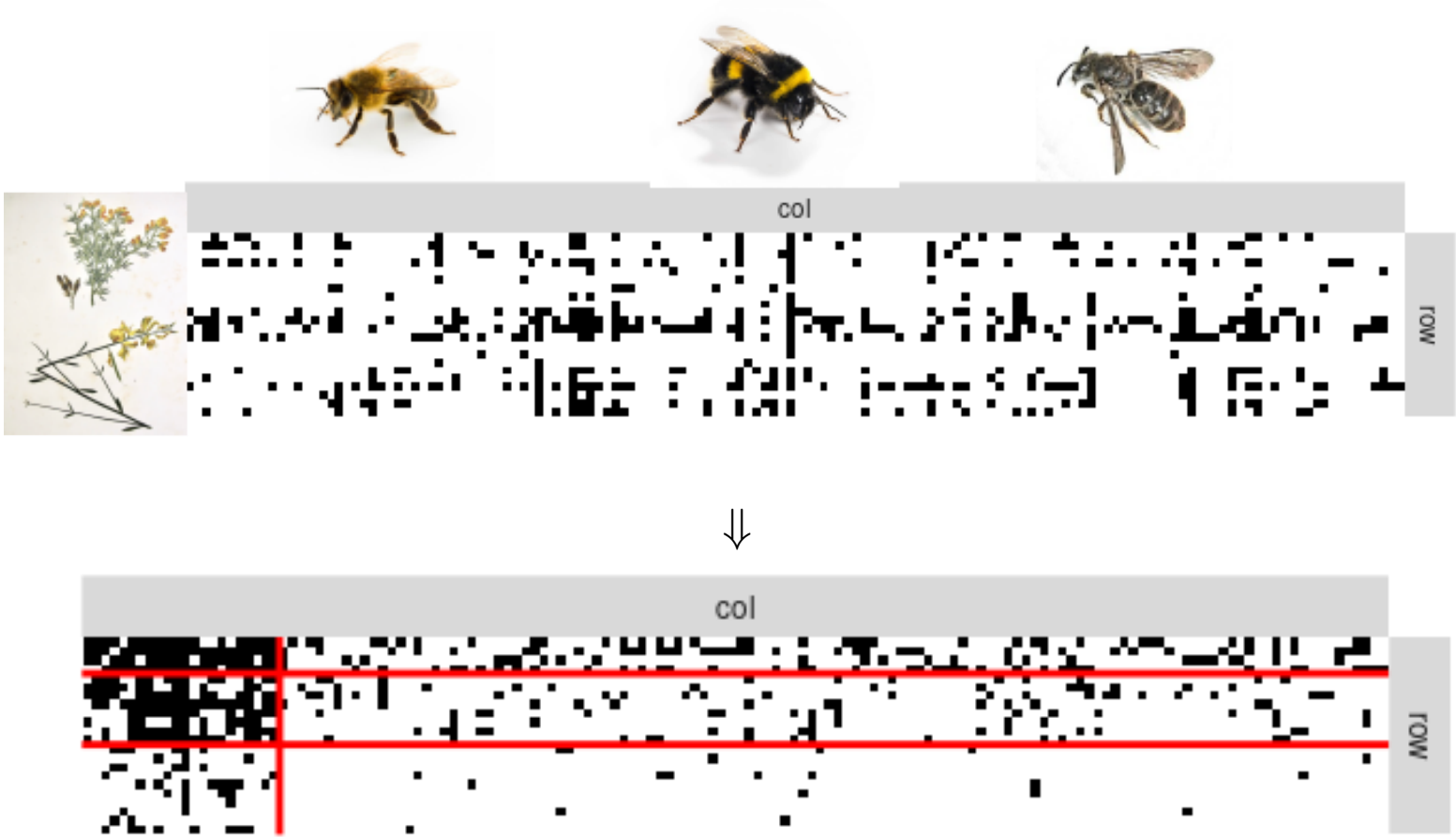


Incidence matrix:

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

Network may also have weighted edges.

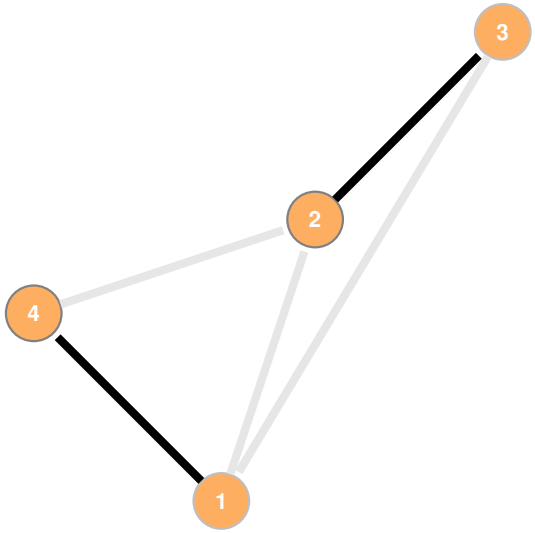
Detecting structure



LBM $\rightarrow$ Clustering of nodes.

Networks with missing data

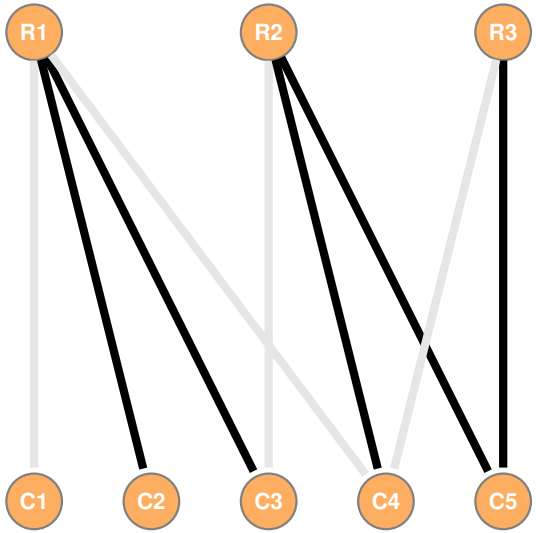
Simple network



Adjacency matrix:

$$A = \begin{pmatrix} 0 & \texttt{NA} & \texttt{NA} & 1 \\ \texttt{NA} & 0 & 1 & \texttt{NA} \\ \texttt{NA} & 1 & 0 & 0 \\ 1 & \texttt{NA} & 0 & 0 \end{pmatrix}$$

Bipartite network

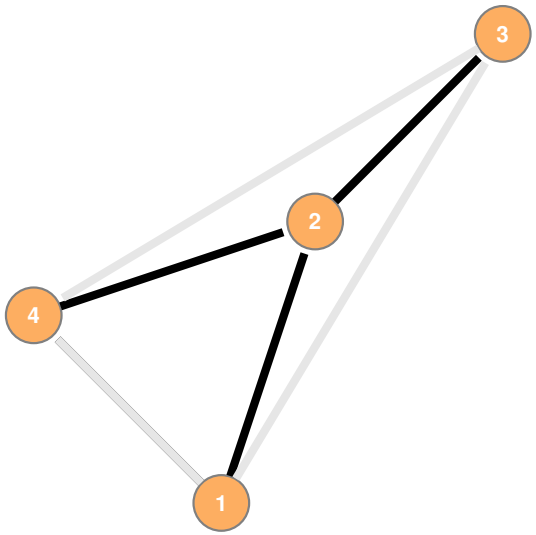


Incidence matrix:

$$B = \begin{pmatrix} \texttt{NA} & 1 & 1 & \texttt{NA} & 0 \\ 0 & 0 & \texttt{NA} & 1 & 1 \\ 0 & 0 & 0 & \texttt{NA} & 1 \end{pmatrix}$$

Networks with incomplete data

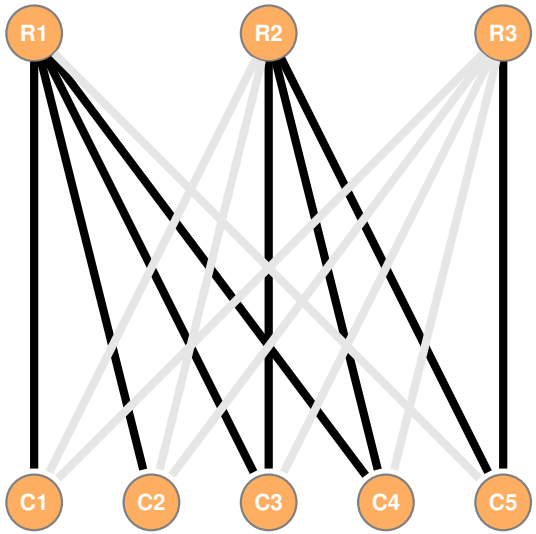
Simple network



Adjacency matrix:

$$A = \begin{pmatrix} 0 & 1 & \texttt{NA} & \texttt{NA} \\ 1 & 0 & 1 & 1 \\ \texttt{NA} & 1 & 0 & \texttt{NA} \\ \texttt{NA} & 1 & \texttt{NA} & 0 \end{pmatrix}$$

Bipartite network



Incidence matrix:

$$B = \begin{pmatrix} 1 & 1 & 1 & 1 & \texttt{NA} \\ \texttt{NA} & \texttt{NA} & 1 & 1 & 1 \\ \texttt{NA} & \texttt{NA} & \texttt{NA} & \texttt{NA} & 1 \end{pmatrix}$$

Outline

Stochastic Block Model and Latent Block Model

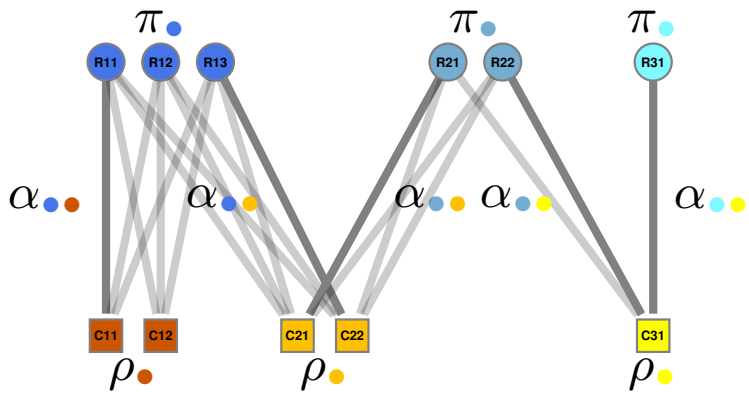
Model on a simple network with n nodes:
SBM: [Nowicki and Snijders, 2001]

- Q blocks of nodes sharing similar connection structure,
- $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ independent latent variables s.t. $\mathbb{P}(Z_i = k) = \pi_k$ for $k \in \{1, \ldots, Q\}$ and $i \in \{1, \ldots, n\}$,
- $Y_{ij} | Z_i, Z_j \overset{ind}{\sim} \mathcal{F}(\alpha_{Z_i, Z_j})$ for all dyads (i, j)

Model on a bipartite network with n_1 and n_2 nodes:
LBM: [Govaert and Nadif, 2010]

- Q_1 and Q_2 blocks of nodes sharing similar connection structure,
- $\boldsymbol{Z}^1 = (Z_1^1, \ldots, Z_n^1)$ and $\boldsymbol{Z}^2 = (Z_1^2, \ldots, Z_m^2)$ independent latent variables s.t. $\mathbb{P}(Z_i^1 = k) = \pi_k^1$ for all $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, Q_1\}$ and $\mathbb{P}(Z_j^2 = l) = \pi_l^2$ for all $j \in \{1, \ldots, m\}$, $l \in \{1, \ldots, Q_2\}$
- $Y_{ij} | Z_i^1, Z_j^2 \overset{ind}{\sim} \mathcal{F}(\alpha_{Z_i^1, Z_j^2})$ for all dyads (i, j).

LBM or BiSBM: illustration



With

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ row blocks
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ column blocks

Parameters to be estimated

- $\pi_\bullet = \mathbb{P}(Z_i^1 = \bullet)$ for rows and $\rho_\bullet = \mathbb{P}(Z_j^2 = \bullet)$ for columns
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i^1 = \bullet, Z_j^2 = \bullet)$

Inference of the LBM

Observed likelihood:

$$\log \ell(\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}); \mathbf{Y}) = \sum_{(\boldsymbol{Z}^1, \boldsymbol{Z}^2) \in \{1,\dots,Q_1\}^{n_1} \times \{1,\dots,Q_2\}^{n_2}} \log \ell_c(\boldsymbol{\theta}; \mathbf{Y}, \boldsymbol{Z}^1, \boldsymbol{Z}^2)\,.$$

Difficulties:

- Sum over $\boldsymbol{Z}^1, \boldsymbol{Z}^2$ is then intractable as soon as either n or Q_1, Q_2 grows,
- $p(\boldsymbol{Z}^1, \boldsymbol{Z}^2 | \mathbf{Y}; \boldsymbol{\theta})$ not tractable so EM algorithm not possible.

Solution: Variational EM algorithm [Daudin et al., 2008] $\rightarrow$ Replace $p(\boldsymbol{Z}^1, \boldsymbol{Z}^2 | \mathbf{Y}; \boldsymbol{\theta})$ with

$$\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}}(\boldsymbol{Z}^1, \boldsymbol{Z}^2) = \prod_{i=1}^{n} \prod_{k=1}^{Q_1} (\tau_{ik}^1)^{\mathbb{I}_{Z_i^1 = k}} \times \prod_{j=1}^{m} \prod_{l=1}^{Q_2} (\tau_{jl}^2)^{\mathbb{I}_{Z_j^2 = l}},$$

where $\tau_{ik}^1 = \mathbb{P}_{\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}^1}}(Z_i^1 = k)$ and $\tau_{jl}^2 = \mathbb{P}_{\mathcal{R}_{\mathbf{Y},\boldsymbol{\tau}^2}}(Z_j^2 = l)$.
Maximizing alternatively in $\boldsymbol{\theta}$ and $\mathcal{R}$:

$$\mathcal{I}_{\boldsymbol{\theta}}(\mathcal{R}) = \log \ell(\boldsymbol{\theta}; \mathbf{Y}) - \mathrm{KL}[\mathcal{R}, \mathbb{P}(\cdot | \mathbf{Y}; \boldsymbol{\theta})].$$

Selecting the number of blocks

Integrated Classification Likelihood [Biernacki et al., 2000, Daudin et al., 2008]

$$\text{ICL}(Q_1, Q_2) = \log \ell_c(\hat{\theta}_{Q_1,Q_2}; \mathbf{Y}, \hat{\boldsymbol{Z}}^1, \hat{\boldsymbol{Z}}^2) - \text{pen}(Q_1, Q_2)$$

where

$$\text{pen}(Q_1, Q_2) \quad = \quad \frac{1}{2}\left\{(Q_1 - 1)\log(n) + (Q_2 - 1)\log(m) + (Q_1 \cdot Q_2)\log\left(n \times m\right)\right\} \ ,$$

for a network with a one-parameter distribution on dyads.

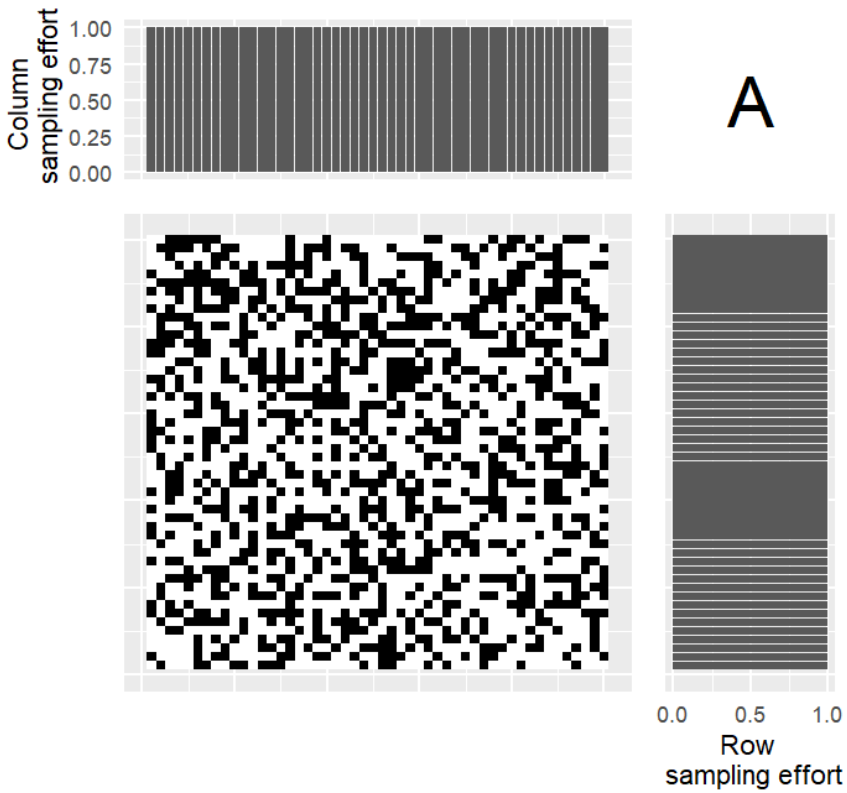
Statistical inference

- Selection of the number of clusters Q_1, Q_2 for SBM
- Estimation of the parameters $\theta_{Q_1,Q_2} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ for a given number of clusters Q_1, Q_2
- Clustering $\widehat{\boldsymbol{Z}}^1, \widehat{\boldsymbol{Z}}^2$,

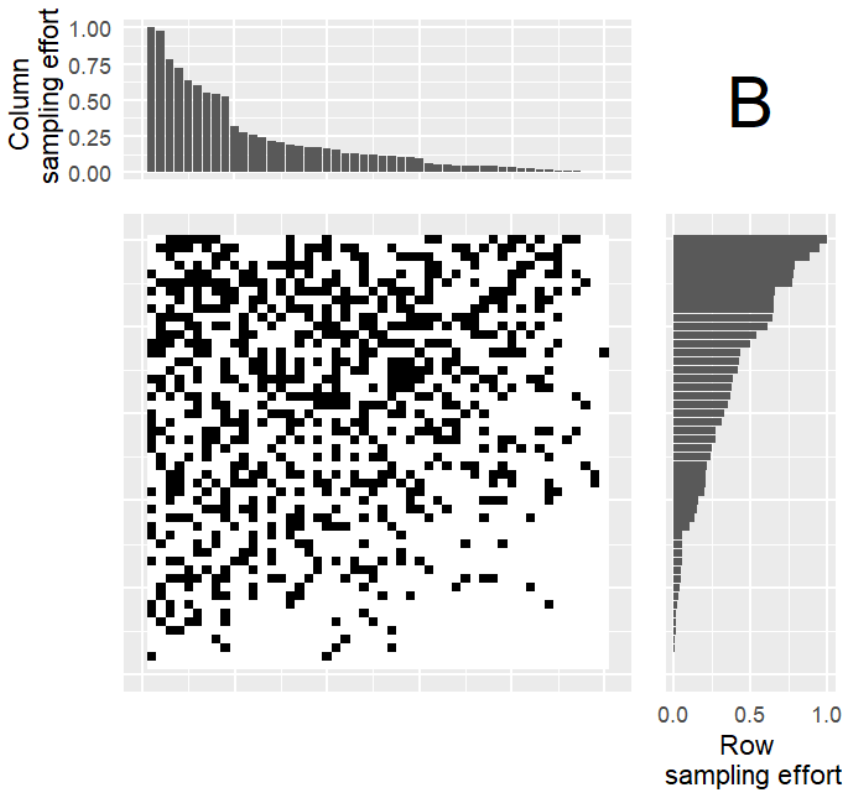
Outline

Simulation and sampling example

- Assumption: $Q_1 = Q_2 = 1, \alpha_{k,l} = c_0$



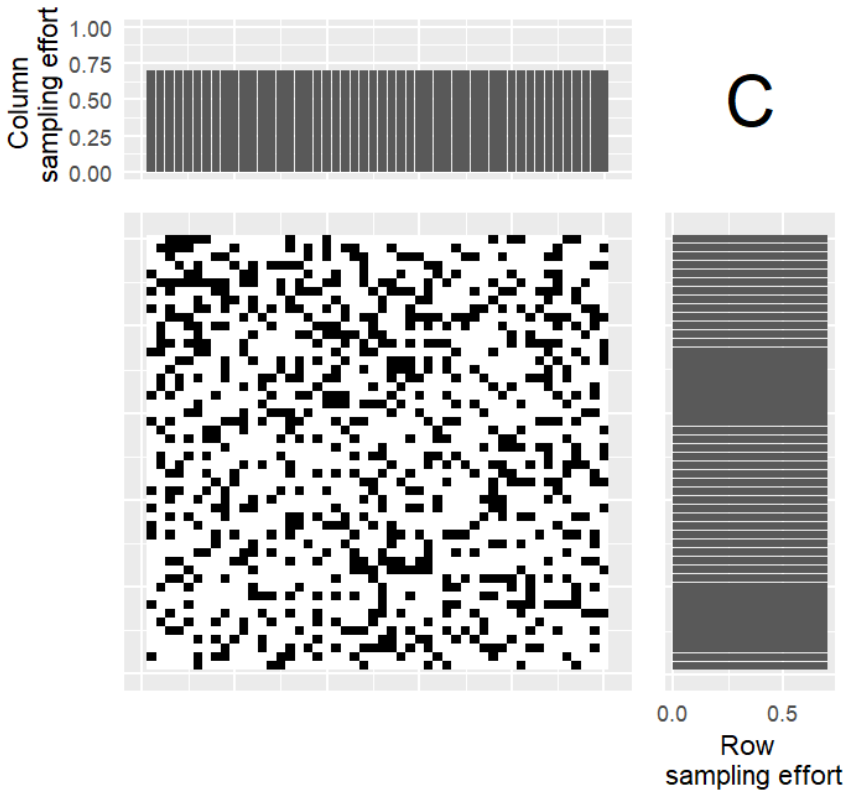
Complete network $M_{i,j}$



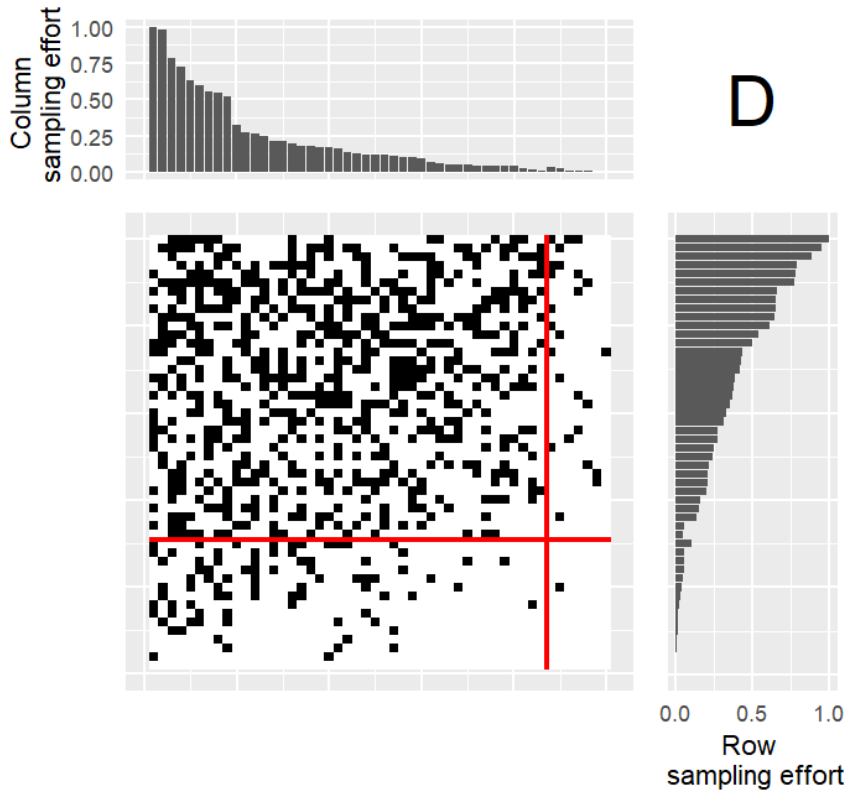
Subsampled network $R_{i,j}$ (70%)

Simulation and sampling example

- Fitting a LBM model on data doesn't yield the same result



LBM fit on uniformly subsampled $M_{i,j}$



LBM fit on $R_{i,j}$

Leveraging frequencies of observation of interactions

Frequency of interactions $R_{i,j}$ = count of interaction of row species i with column species j.

$R_{i,j} = 0$ can correspond to

- No possible interaction
- Possible interaction but not observed in the sampling process.

Goal: disentangle possible interactions from sampling effects

CoOP-LBM

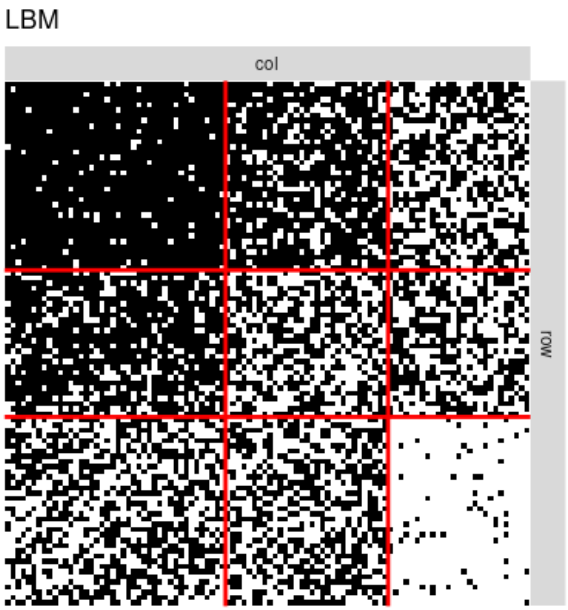


Figure: $M \sim$ LBM $(\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$

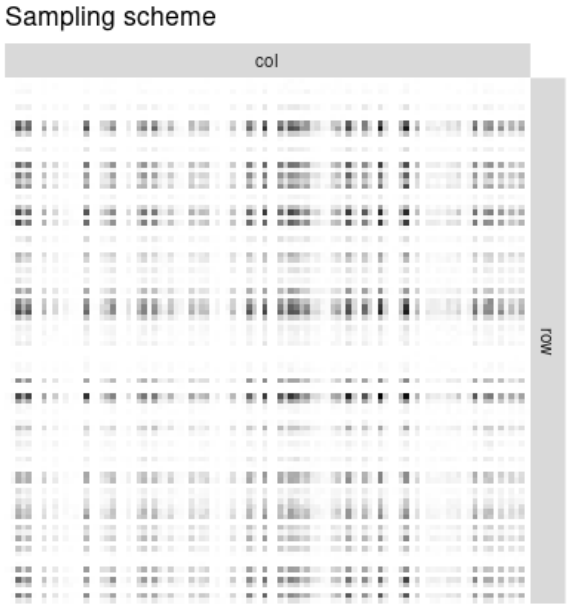


Figure: $N \sim$ Sampling scheme $\mathcal{P}(\lambda_i \mu_j G)$
$\lambda_i, \mu_j \in]0, 1]$, $G > 0$

$R_{i,j} = M_{i,j} \odot N_{i,j}$.

- M and N are supposed independent,
- λ_i and μ_j are related to abundance / activity /...,
- G is an overall sampling intensity.

Adapt inference and model selection

Model and Inference:

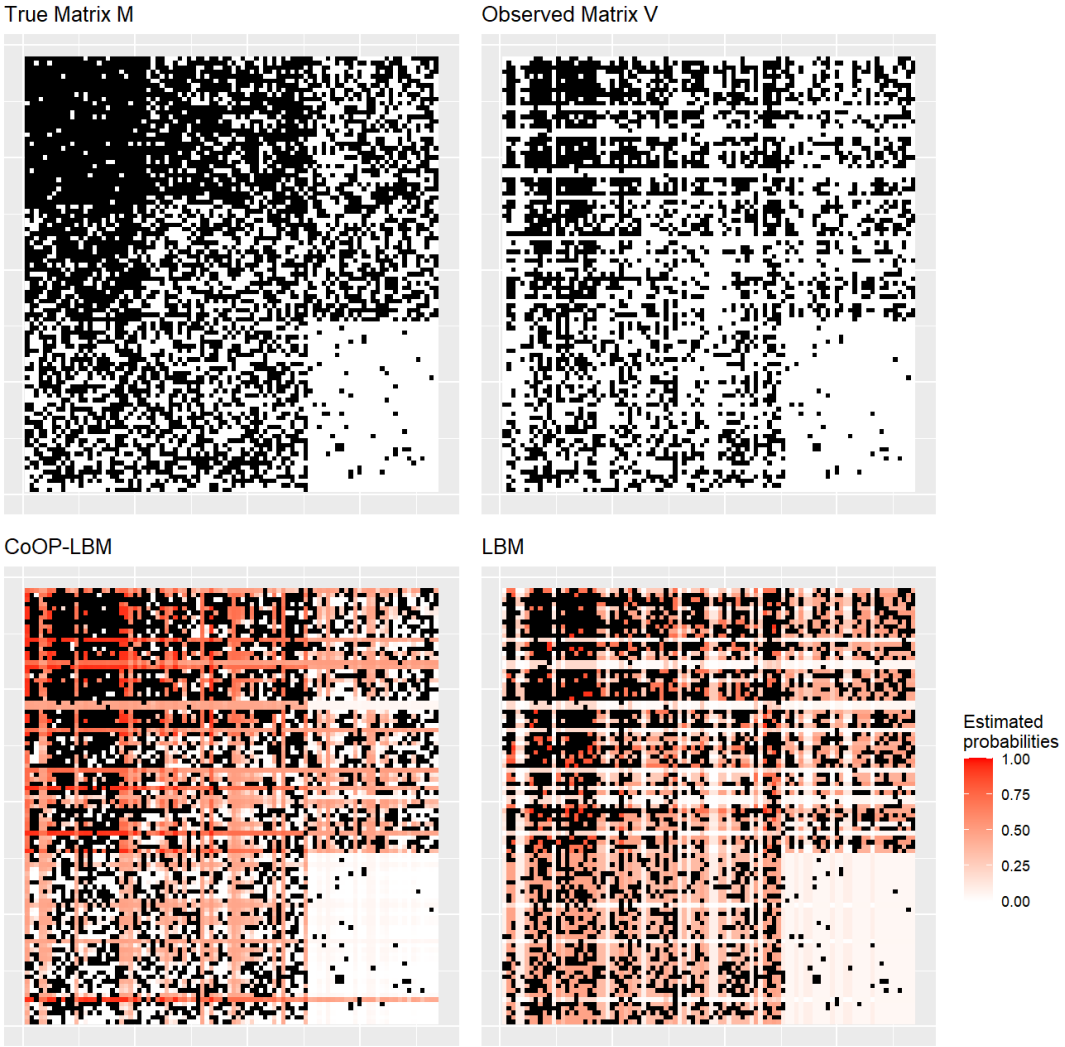
- Identifiability of the model,
- Stochastic version of an EM algorithm for estimation,
  - S-step: simulate latent M matrix,
  - S-step: simulate $\boldsymbol{Z}^1$ and $\boldsymbol{Z}^2$.
- ICL criterion to select the numbers of groups.

Simulation study:

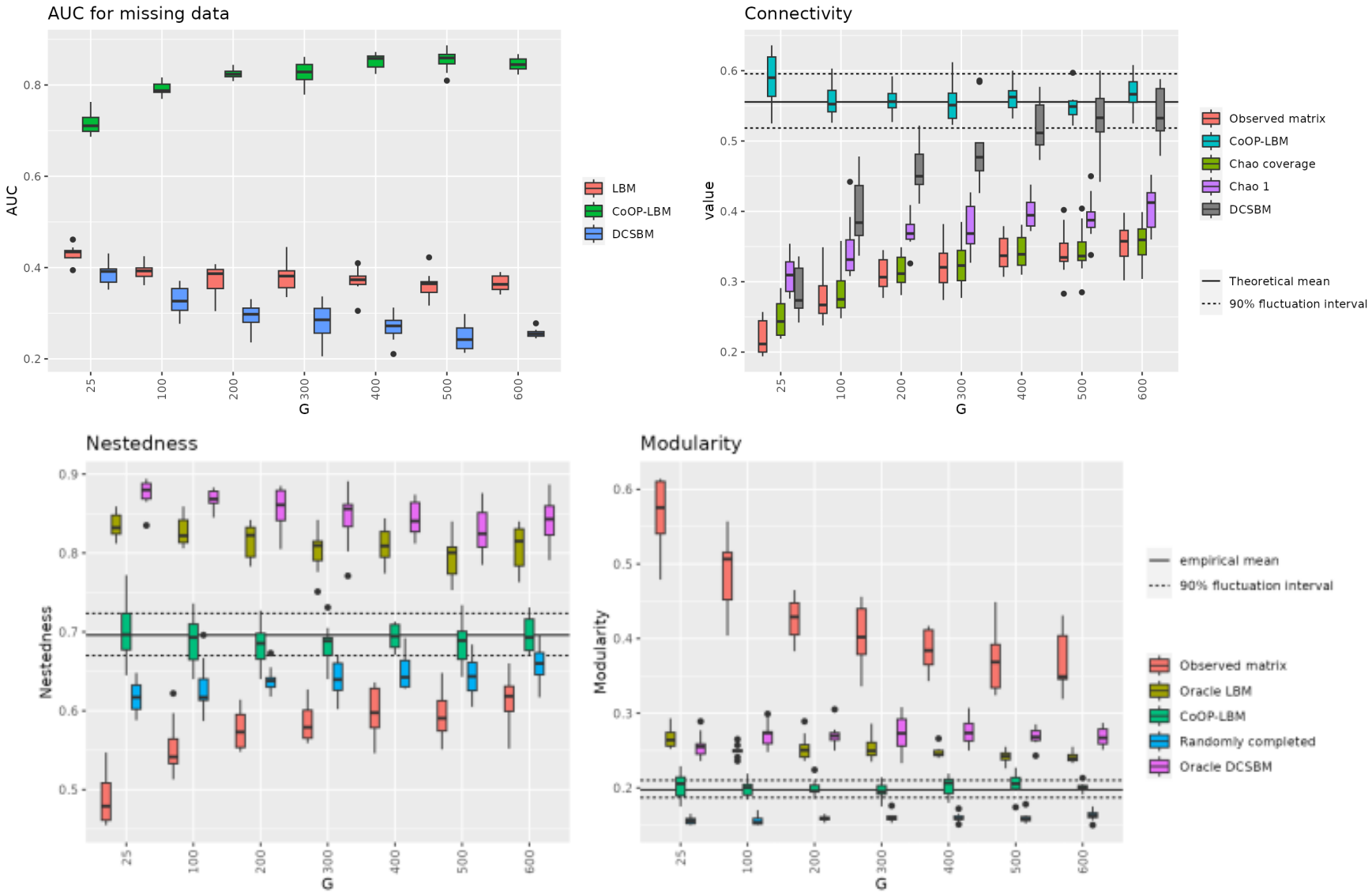
- Comparison with a DCSBM and an LBM: better recovery of missing interactions, better clusterings,
- Correction of usual metrics: connectance, nestedness, modularity,
- Robust to model misspecification: preferences of connection...

Outputs

- estimation and clustering of an LBM model
- estimation of $\mathbb{P}(M_{ij} = 1)$ probabilities that a non-observed interaction is possible

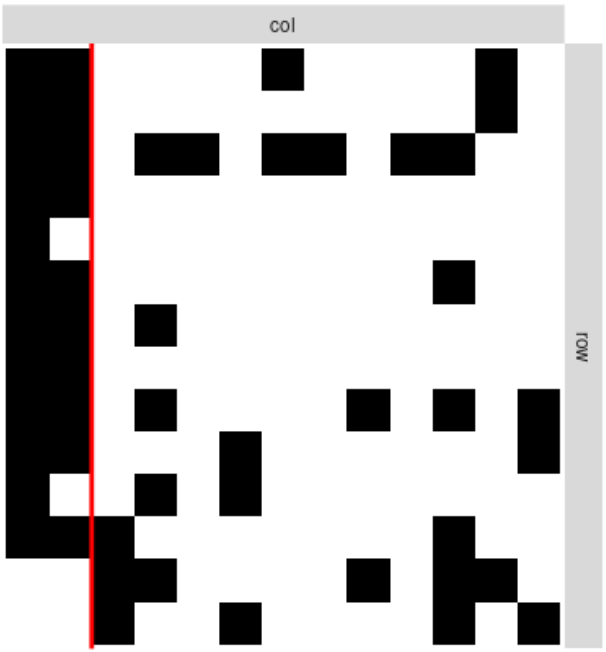


Recovery of unobservef interactions and correction of metrics

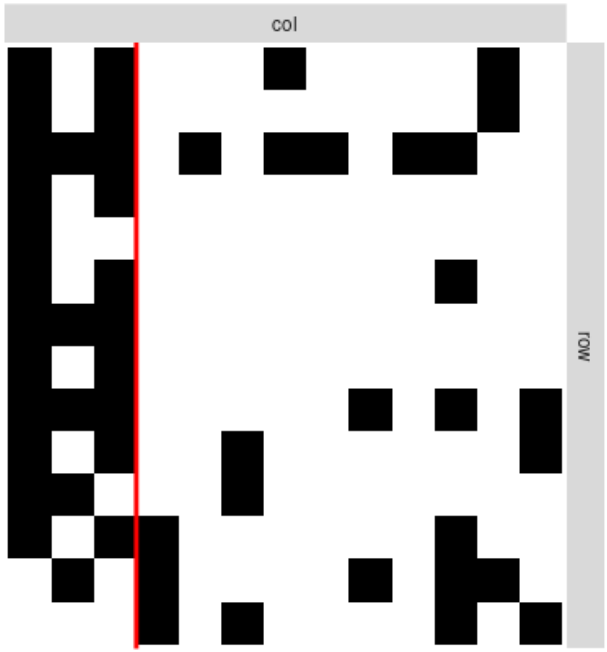


Model fit

- [Olesen et al., 2002]: Invasion of pollination networks on oceanic islands: importance of invader complexes and endemic super generalists
- 14 species of plants, 13 species of insects.
- 1395 interactions observed, very well sampled interaction matrix.



LBM



CoOP-LBM

Only difference is observed for the insect species Lycaenidae pirithous

Estimated coverage

Boraginaceae argentea	0.8493672
Asparagaceae concinna	0.8007462
Araliaceae mauritiana	0.9998364
Malvaceae tiliaceus	0.8137610
Convolvulaceae macrantha	0.6746022
Fabaceae leucocephala	0.8028505
Rubiaceae citrifolia	0.9993632
Passifloraceae suberosa	0.7164068
Lythraceae acidula	0.9952563
Goodeniaceae sericea	0.8680551
Surianaceae maritima	0.9703093
Malvaceae populnea	0.8608773
Verbenaceae jamaicencis	0.9348724
Turneraceae angustifolia	0.7207160

Apidae mellifera	0.9997752
Hesperiidae borbonica	0.9227533
Lycaenidae pirithous	0.4642996
Muscidae sp.	0.8855992
Megachilidae sp.	0.9971414
Muscidae domestica	0.9371763
Syrphidae obesa	0.8847282
Nymphalidae phalantha	1.0000000
Gekkonidae ornata	0.9452383
Cetoniidae aurichalcea	0.8761673
Stratiomyidae sp.	0.9988445
Syrphidae sp.	0.9876197
Apidae fenestrata	0.9964598

Lycaenidae pirithous has been observed only 7 times on 5 different flowers.

return to the first example

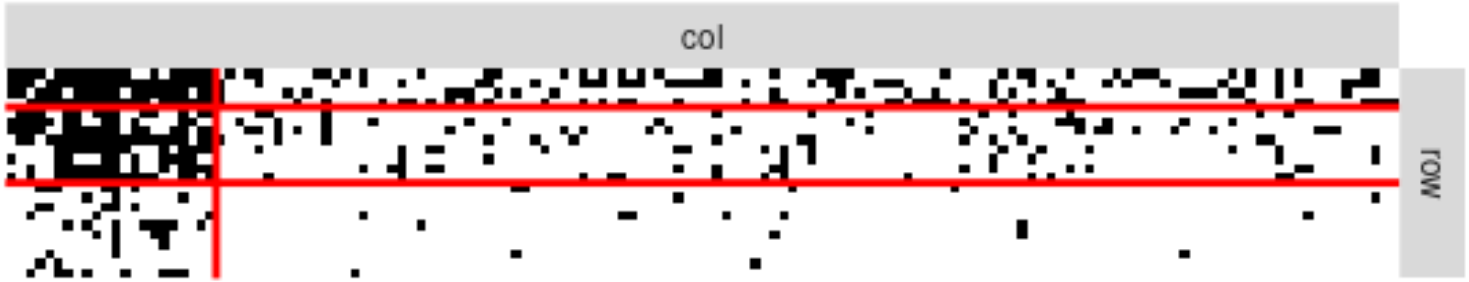


Figure: LBM fit on [Lara-Romero et al., 2019]

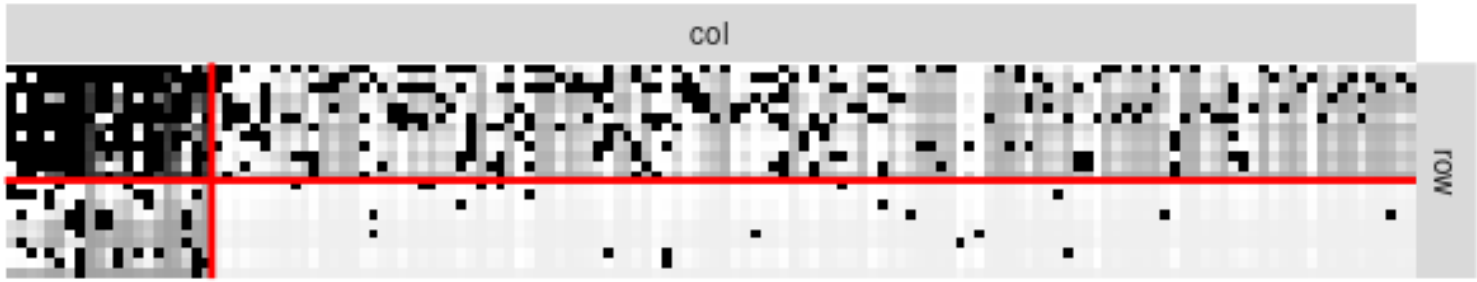


Figure: CoOP-LBM fit on [Lara-Romero et al., 2019]

More in Anakok et al.

paper:
Anakok, Emre, Pierre Barbillon, Colin Fontaine, and Elisa Thebault. 2025. "Disentangling the Structure of Ecological Bipartite Networks from Observation Processes." The Annals of Applied Statistics.

Simulation studies with reproducible code within the `R` package:
`https://github.com/AnakokEmre`.

On 70 networks collected in [Doré et al., 2021]

- Clear differences in the inferred structures with the COOP-LBM in comparison with DCSBM and LBM: LBM and DCSBM tend to create more clusters...
- Correction of classical metrics.

Biernacki, C., Celeux, G., and Govaert, G. (2000).

Assessing a mixture model for clustering with the integrated completed likelihood.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(7):719–725.

Daudin, J.-J., Picard, F., and Robin, S. (2008).

A mixture model for random graphs.
Statistics and computing, 18(2):173–183.

Doré, M., Fontaine, C., and Thébault, E. (2021).

Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale.
Global Change Biology, 27(6):1266–1280.

Govaert, G. and Nadif, M. (2010).

Latent block model for contingency table.
Communications in Statistics—Theory and Methods, 39(3):416–425.

Lara-Romero, C., Seguí, J., Pérez-Delgado, A., Nogales, M., and Traveset, A. (2019).

Beta diversity and specialization in plant–pollinator networks along an elevational gradient.
Journal of Biogeography, 46(7):1598–1610.

Nowicki, K. and Snijders, T. A. B. (2001).

Estimation and prediction for stochastic blockstructures.
Journal of the American Statistical Association, 96(455):1077–1087.

Olesen, J. M., Eskildsen, L. I., and Venkatasamy, S. (2002).

Invasion of pollination networks on oceanic islands: importance of invader complexes and endemic super generalists.
Diversity and Distributions, 8(3):181–192.