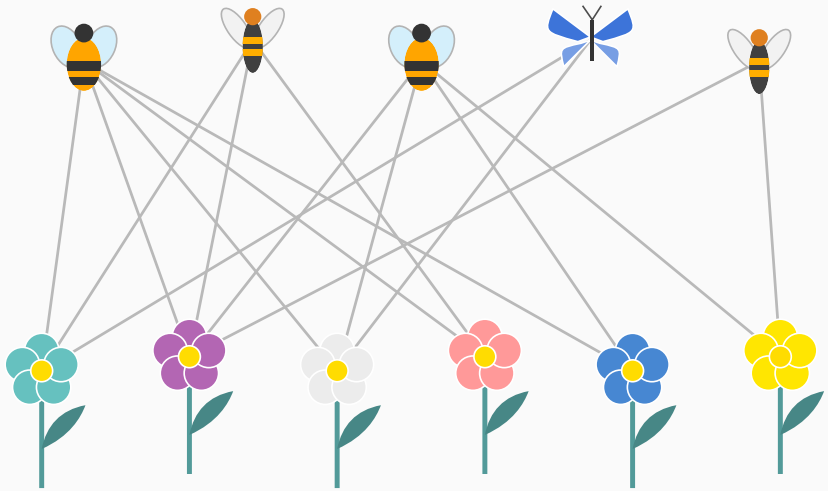
Pollination networks in urban areas

From descriptive indices to stochastic block models

Sophie Donnet, INRAE, MIA Paris-Saclay

Nov. 2026



Who am I?

Sophie Donnet, statistician, INRAE, MIA Paris-Saclay

- Research: statistical models with latent variables, inference (variational, Bayesian) and model selection
- I develop Stochastic Block Models tools for ecological networks
- R packages: `sbm`, `GREMLINS`, `colSBM`

With

Statistics

- P. Barbillon (AgroParisTech, MIA Paris-Saclay)
- L. Lacoste and S.-C. Chabert-Liddell (PhD student, MIA Paris-Saclay)

Ecology

- E. Thébault (CNRS, iEES-Paris)
- S. Kéfi (CNRS, ISEM Montpellier)
- F. Massol (CNRS, iEES-Paris)

Program of the 3 hours

Short version of the CESAB training, adapted to urban biodiversity.

1. Introduction to networks: mathematical objects and their ecological interpretation
2. Introduction to stochastic block models (SBM)
3. Comparing network organizations with SBM

Running example: plant–pollinator networks in 4 UK cities [Baldock et al., 2019].



Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

Conclusion

Why study pollination?

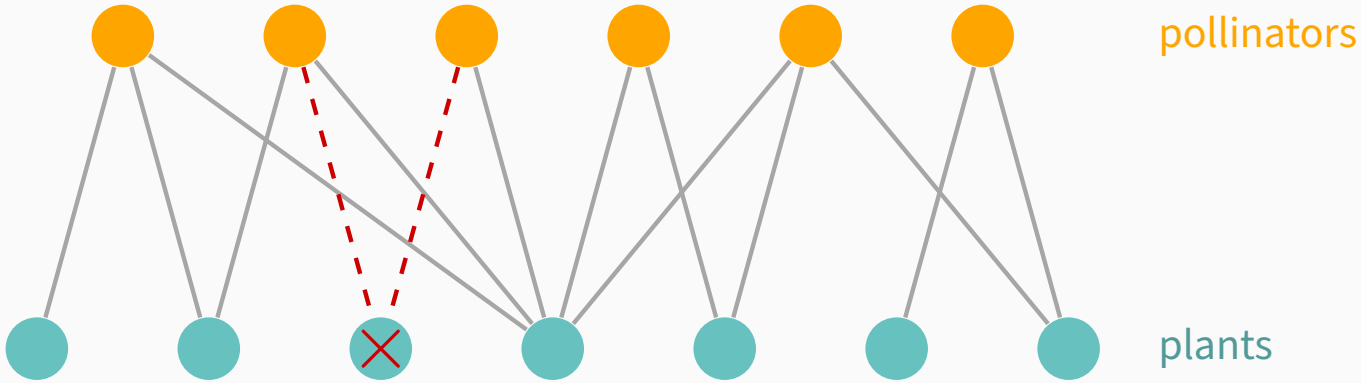


- About 90% of flowering plants are pollinated by animals [Ollerton et al., 2011]
- 75% of the leading crops benefit from animal pollination [Klein et al., 2007]
- Pollinators are declining: habitat loss, pesticides, pathogens, climate [Potts et al., 2010]

→ Understanding what maintains pollination in changing landscapes

Why do we use networks?

- A plant is visited by many insects, an insect visits many plants
- The loss of a species propagates through its interactions
- The structure of the network matters: the organization of the network drives the robustness to species loss
  [Bascompte et al., 2003, Memmott et al., 2004]



→ Study the whole set of interactions, not species one by one

Our running example

UK data



Europe PMC Funders Group
Author Manuscript
Nat Ecol Evol. Author manuscript; available in PMC 2019 July 14.

Published in final edited form as:
Nat Ecol Evol. 2019 March ; 3(3): 363–373. doi:10.1038/s41559-018-0769-y.

A systems approach reveals urban pollinator hotspots and conservation opportunities

Katherine C. R. Baldock[*,1,2], **Mark A. Goddard**[3,4], **Damien M. Hicks**[5], **William E. Kunin**[3], **Nadine Mitschunas**[1,6], **Helen Morse**[1], **Lynne M. Osgathorpe**[1], **Simon G. Potts**[6], **Kirsty M. Robertson**[3], **Anna V. Scott**[6], **Phillip P. A. Staniczenko**[7,8], **Graham N. Stone**[5], **Ian P. Vaughan**[9], and **Jane Memmott**[1,2]

[1]School of Biological Sciences, University of Bristol, 24 Tyndall Avenue, Bristol, BS8 1TQ, UK

[2]Cabot Institute, University of Bristol, Bristol, BS8 1UJ, UK

[3]School of Biology, University of Leeds, Leeds, LS2 9JT, UK

[5]Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, EH9 3JT, UK

[6]Centre for Agri-Environmental Research, School of Agriculture, Policy and Development, University of Reading, Reading, RG6 6AR, UK

[7]National Socio-Environmental Synthesis Center (SESYNC), Annapolis, MD 21401, USA

[9]Cardiff School of Biosciences, Cardiff University, Museum Avenue, Cardiff, CF 10 3AX, UK

Abstract

Urban areas are often perceived to have lower biodiversity than the wider countryside, but a few small-scale studies suggest that some urban land uses can support substantial pollinator populations. We present a large-scale, well-replicated study of floral resources and pollinators in

- Baldock et al. (2019), *Nature Ecology & Evolution*
- Plant–pollinator networks in 4 UK cities
- A reference study: large-scale, well replicated (360 sites)
- 723 citations
- Open data (supplementary material)

Their problematic

UK data

Baldock et al. (2019), *A systems approach reveals urban pollinator hotspots and conservation opportunities*, Nature Ecology & Evolution.

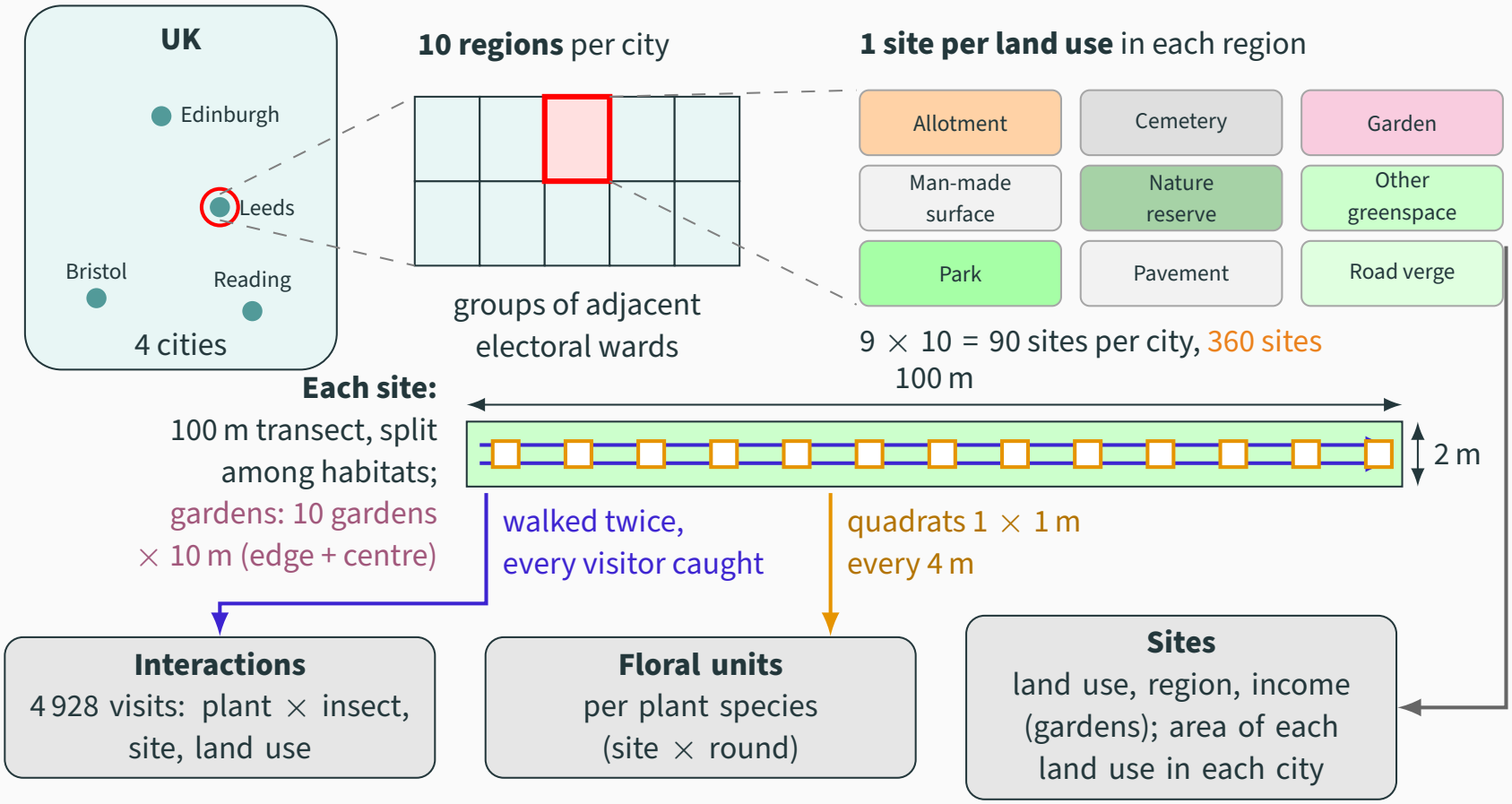
- Cities are a mosaic of land uses: gardens, parks, cemeteries, road verges…
- Which land uses are pollinator hotspots?
- Which management would make the city-scale community more robust to species loss?

What they found

Gardens (by their area) and allotments (by their diversity) are the hotspots.

The protocol

UK data



4 921 visits, 377 plant taxa, 335 insect taxa

From this study to this course

- [Baldock et al., 2019] focus on robustness to species disappearances
- Here: the structure/organization of the networks.

What I propose

- Part 1: networks and descriptors
- Part 2: a probabilistic model, the SBM : definition, inference, interpretation
- Part 3 : comparing networks

→ Illustration on the **UK data**

Motivation

Introduction to networks

- Graphs and matrices
- Node-level descriptors
- Network-level descriptors

Introduction to stochastic block models

Modelling collections of networks

Conclusion

Motivation

Introduction to networks

Graphs and matrices

Node-level descriptors

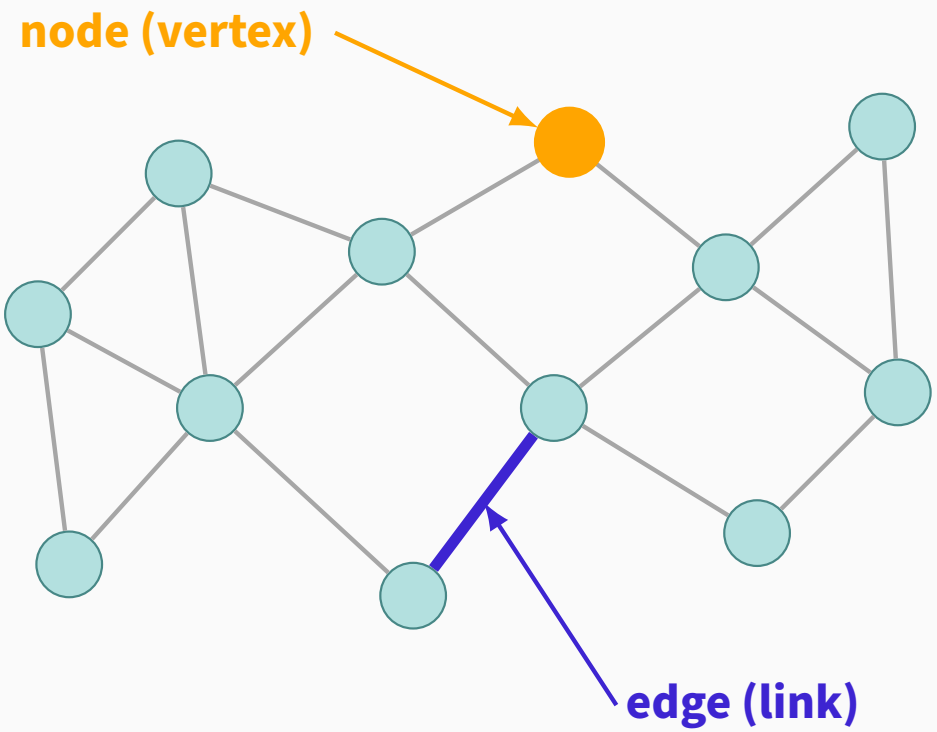
Network-level descriptors

Introduction to stochastic block models

Modelling collections of networks

Conclusion

What is a network?



Nodes	Edges
species	trophic, mutualistic, parasitic interactions
individuals	contacts, social interactions
habitat patches (parks, ponds…)	dispersal of individuals

From a graph to a matrix

A graph $G = (V, E)$:

- $V = \{1, \ldots, n\}$: set of nodes (vertices)
- $E \subset V \times V$: set of edges (links), $(i, j) \in E$ if i and j interact

Adjacency matrix

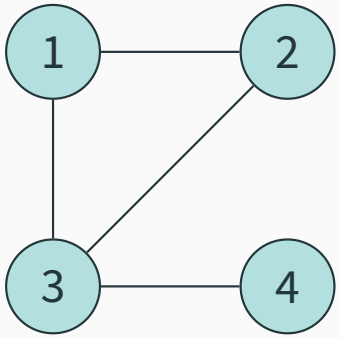
$$Y = (Y_{ij})_{1 \leq i,j \leq n}, \qquad Y_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

- Use of matrix possible because small networks
- In general, $Y_{ii} = 0$: no self interaction.

Undirected and directed networks

Undirected: $Y_{ij} = Y_{ji}$

e.g. co-occurrence, competition

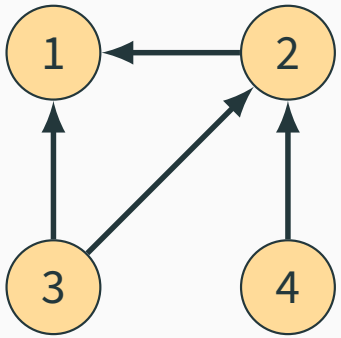


$$Y = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

symmetric

Directed: $Y_{ij} \neq Y_{ji}$

e.g. food web: $Y_{ij} = 1$ if i is eaten by j

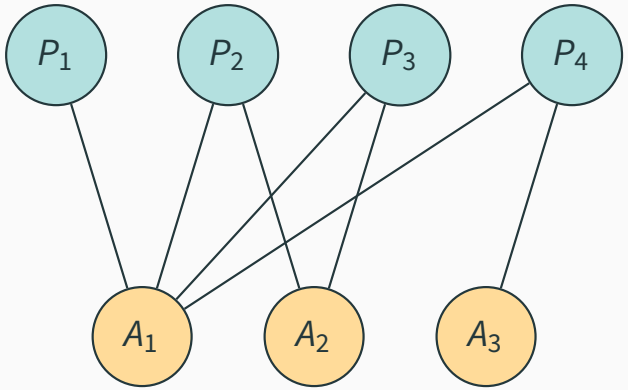


$$Y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

not symmetric

Bipartite networks

Two types of nodes, links only between types: $V = V_1 \cup V_2$, $|V_1| = n$, $|V_2| = m$.



plants (top), pollinators (bottom)

Incidence* matrix Y of size $n \times m$:

$$Y = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \end{array} \begin{array}{c} \begin{array}{ccc} A_1 & A_2 & A_3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \end{array}$$

Rectangular matrix

- Mutualistic (pollination, seed dispersal), host–parasite networks

* "Incidence" is used in ecology, mathematicians use "bi-adjacency matrix"

Weighted (valued) networks

Y_{ij} is no longer $0/1$ but:

- a count: number of interactions between species j and species i, $Y_{ij} \in \mathbb{N}$
- an intensity: biomass flow, interaction strength, $Y_{ij} \in \mathbb{R}_+$
- a probability of dispersal between two patches, $Y_{ij} \in [0, 1]$

Remarks

- Binary network = $\mathbf{1}_{\{Y_{ij}>0\}}$: loses information
- A count mixes interaction preference and abundances / sampling effort: two abundant species are often observed in interaction, even without preference

The Bristol dataset

UK data

One row = one observed visit (4 921 rows)

Where			Who	
City	Land use	Site	Plant (status)	Insect (group)
Bristol	Garden	BR_GDN_01	*Spiraea* sp. (non-native)	*Lasioglossum morio* (bee)

- Plant status: native / non-native
- Insect group: bee, hoverfly, other fly, beetle, …

→ Construction of networks

The Bristol dataset

UK data

Raw data: one row = one visit

Site	Plant	Insect
BR_ALT_05	*R. fruticosus*	*A. mellifera*
BR_NR_04	*R. fruticosus*	*A. mellifera*
BR_ALT_07	*R. fruticosus*	*B. pascuorum*
BR_PK_06	*Taraxacum*	*A. mellifera*
⋮		

1 379 visits in Bristol

count each (plant, insect)

sites, land uses, rounds pooled

Count matrix

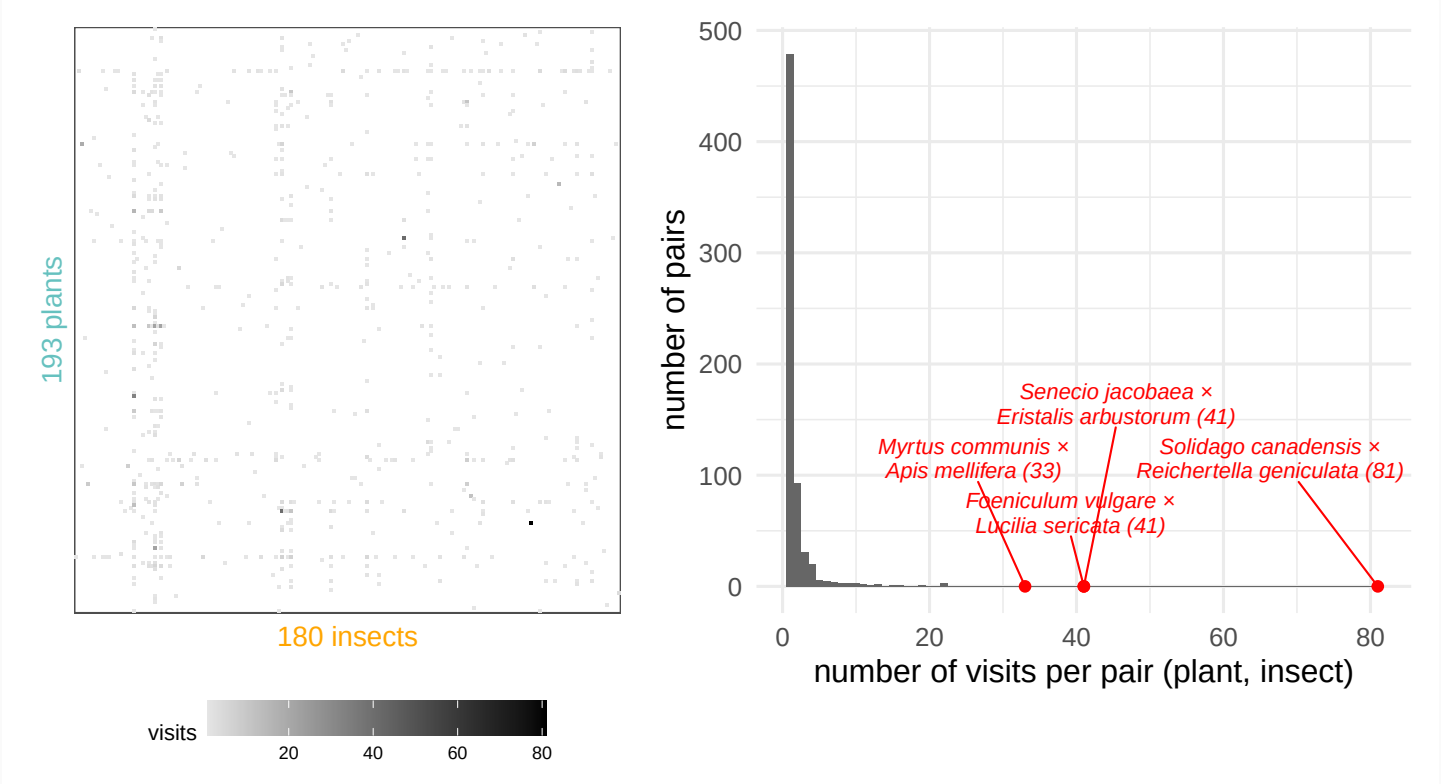
	A. mell.	*B. pasc.*	*E. tenax*
R. fruticosus	**10**	3	0
S. jacobaea	1	1	11
Taraxacum	2	1	7

A count

Y_{ij} = number of visits of insect j on plant i → weighted network

Bristol: the count matrix

UK data



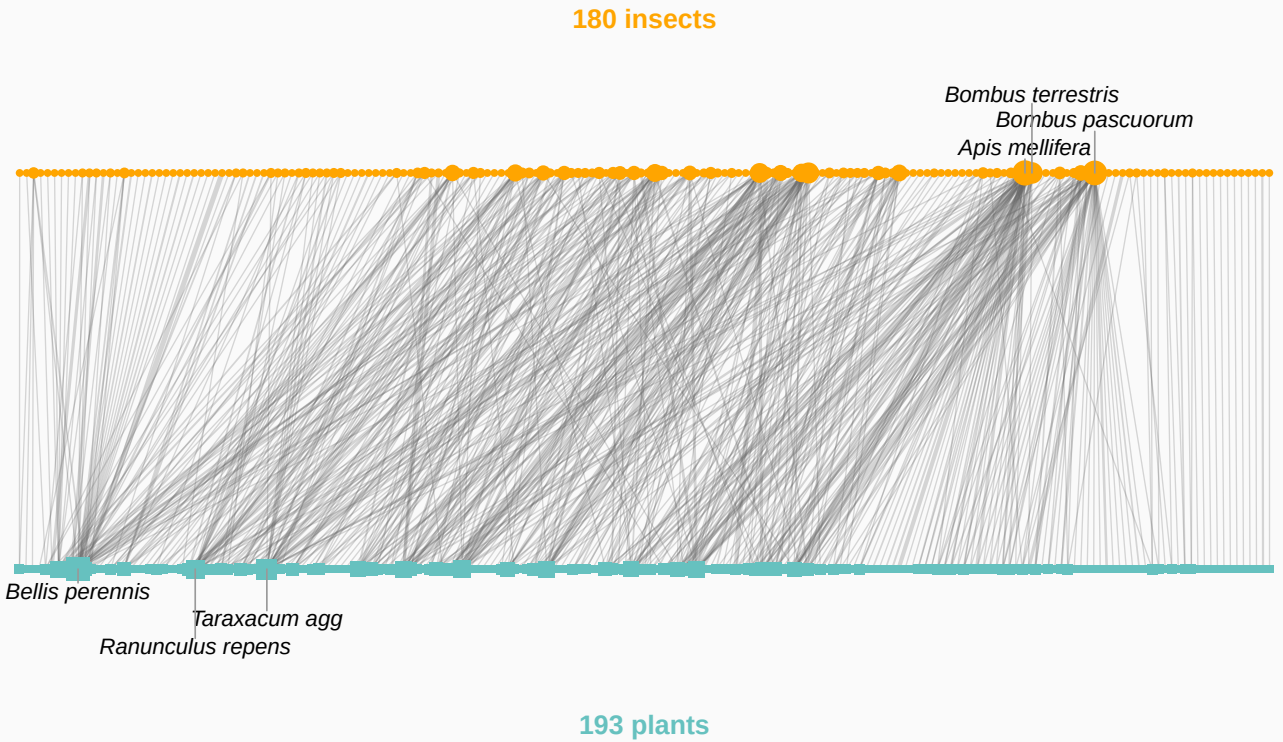
- 72% of the pairs are seen once
- A few large counts crush the scale.
- Largest count: 81 visits of one fly (*Reichertella geniculata*) on *Solidago canadensis*, at a single site





Bristol: the weighted network

UK data



193 plants, 180 insects, 662 links. Node size $\propto$ degree, edge width $\propto$ number of visits.

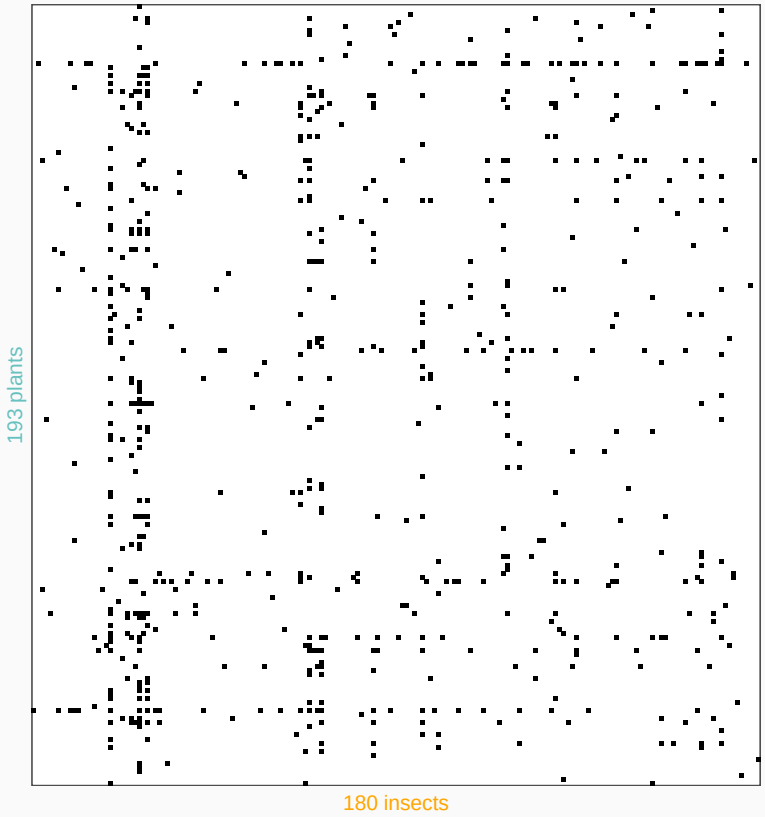
The order of nodes is arbitrary (alphabetic order or given layout).

The Bristol dataset: binarised incidence matrix

UK data

Binarised

Y_{ij} = 1 if at least one visit of insect j on plant i $\rightarrow$ binary network



Warning: only interacting species

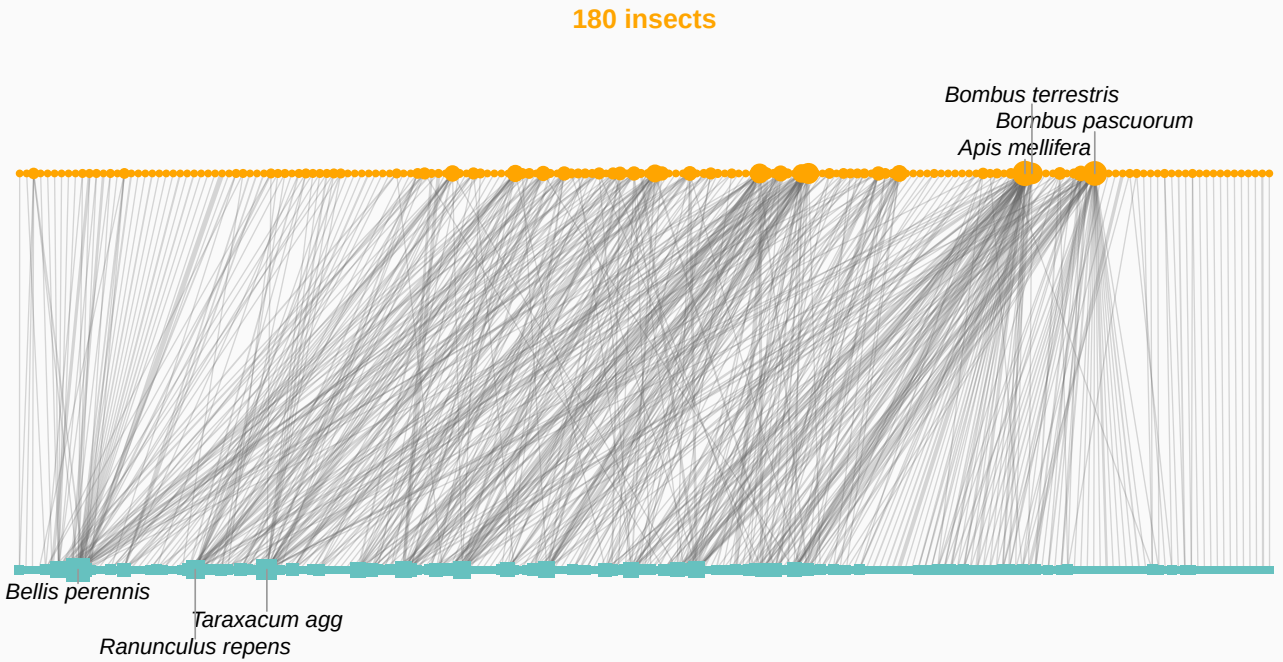
- Every insect in the network was seen on a flower
- Every plant in the network was visited

- No inventory of the insects present
- No complete inventory of the flowers present (quadrats only)

→ No row or column full of 0.

→ A zero means "not observed" not "impossible".

Descriptors



- All the nodes do not interact the same: can we describe them?
  → node-level descriptors
- Is there a particular organization of the matrix : can we describe it?
  → network-level descriptors

Motivation

Introduction to networks

- Graphs and matrices
- Node-level descriptors
- Network-level descriptors

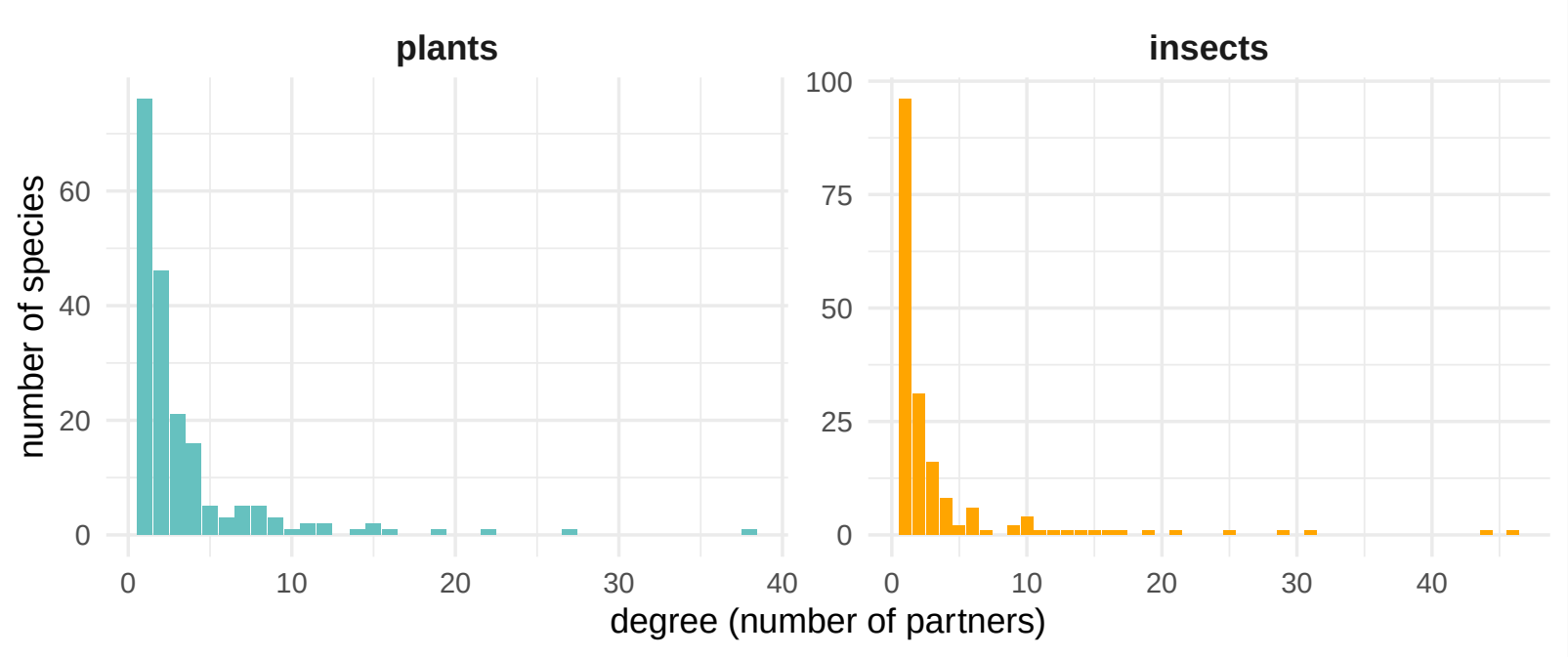
Introduction to stochastic block models

Modelling collections of networks

Conclusion

Degree = number of partners

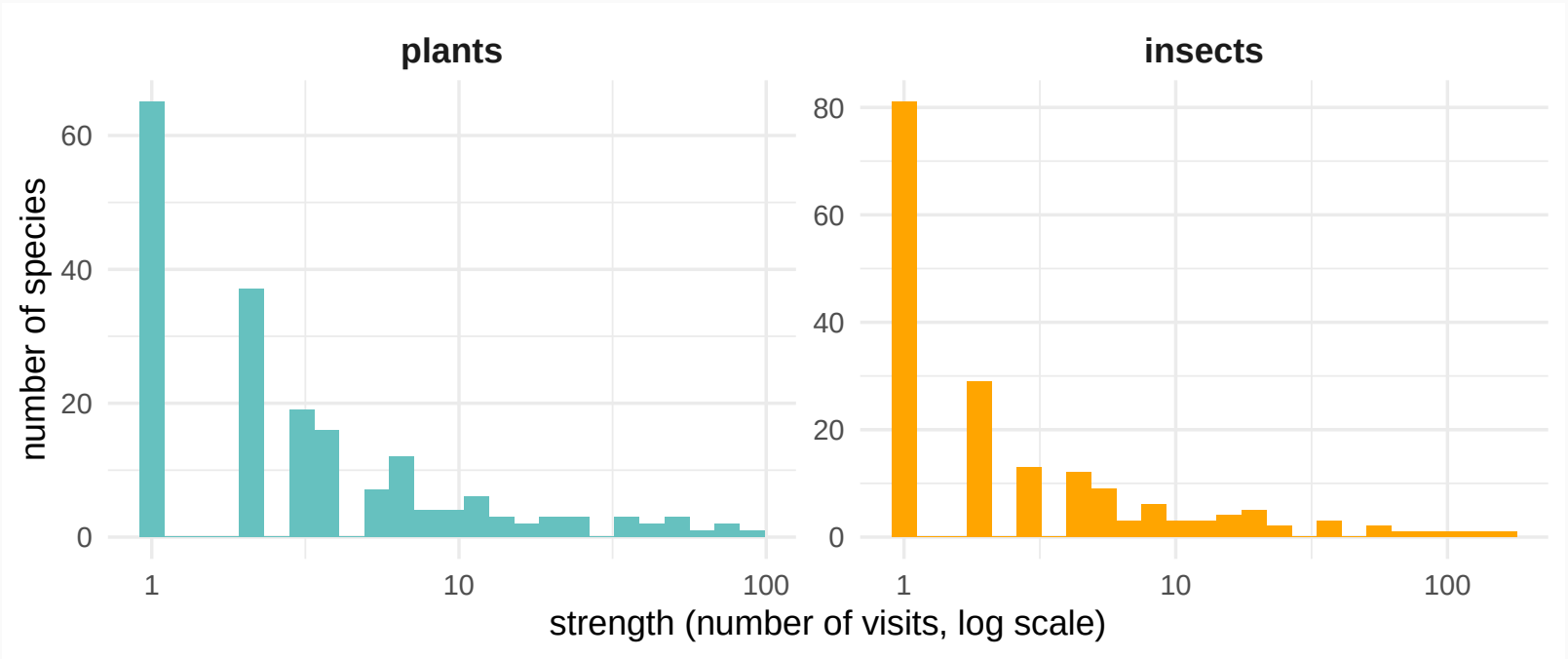
$$d_i = \sum_{j=1}^{m} \mathbf{1}_{\{Y_{ij}>0\}} \quad \text{(plant } i\text{)}, \qquad d_j = \sum_{i=1}^{n} \mathbf{1}_{\{Y_{ij}>0\}} \quad \text{(pollinator } j\text{)}$$



- Highly skewed: 53% of the insects have a single partner
- d_j small: specialist; d_j large: generalist

Strength = number of interactions

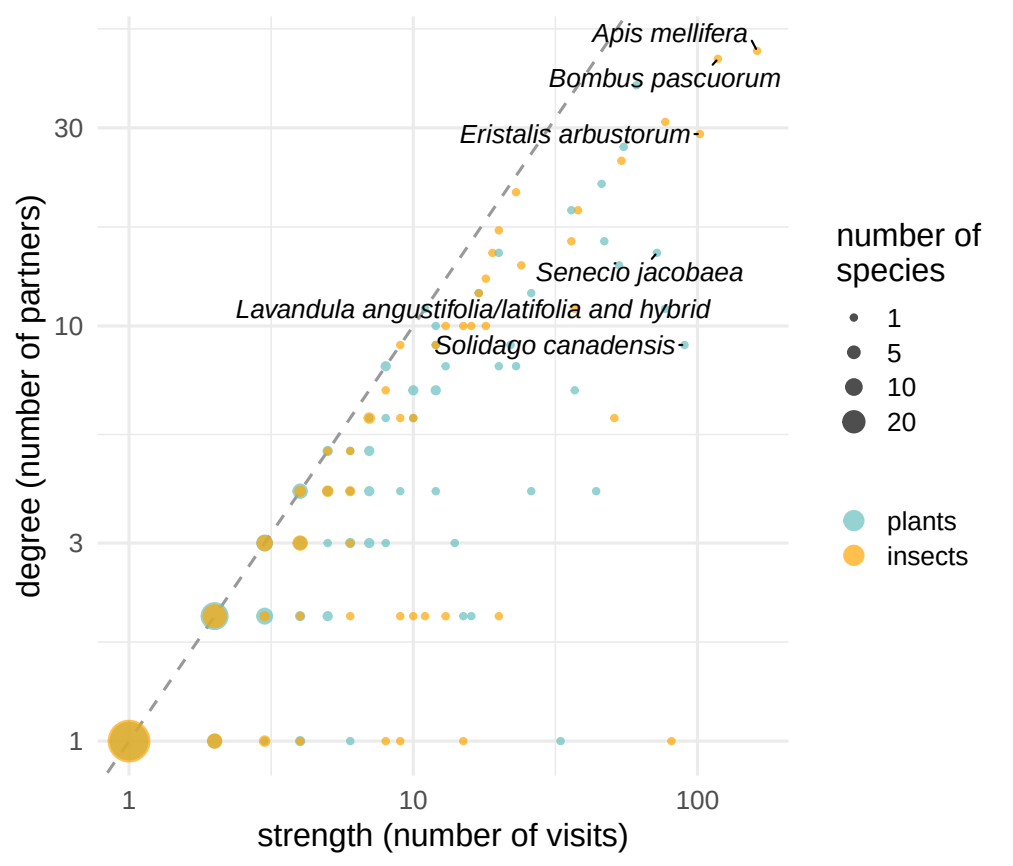
$$s_i = \sum_{j=1}^{m} Y_{ij} \quad \text{(plant } i\text{)}, \qquad s_j = \sum_{i=1}^{n} Y_{ij} \quad \text{(pollinator } j\text{)}$$



- From 1 visit to 162 (*Apis mellifera*): note the log scale
- Strength ≈ how much a species is seen : naturally linked to the abundance.

Degree vs strength

UK data



Apis mellifera (know to be an abundant taxa) is seen with many (various) partners

Specialisation index d'

The strength ignores how interactions are distributed and what is available. [Blüthgen et al., 2006]

- For plant i: $p_{ij} = Y_{ij}/s_i$, proportion of its visits made by insect j
- For each insect j: $q_j = s_j / \sum_{j'} s_{j'}$, its share of all the visits (availability)

Specialisation index d'_i: comparison of distributions

$$\delta_i = \mathrm{KL}(p_{i\cdot} \,\|\, q) = \sum_{j=1}^{m} p_{ij} \log \frac{p_{ij}}{q_j}$$

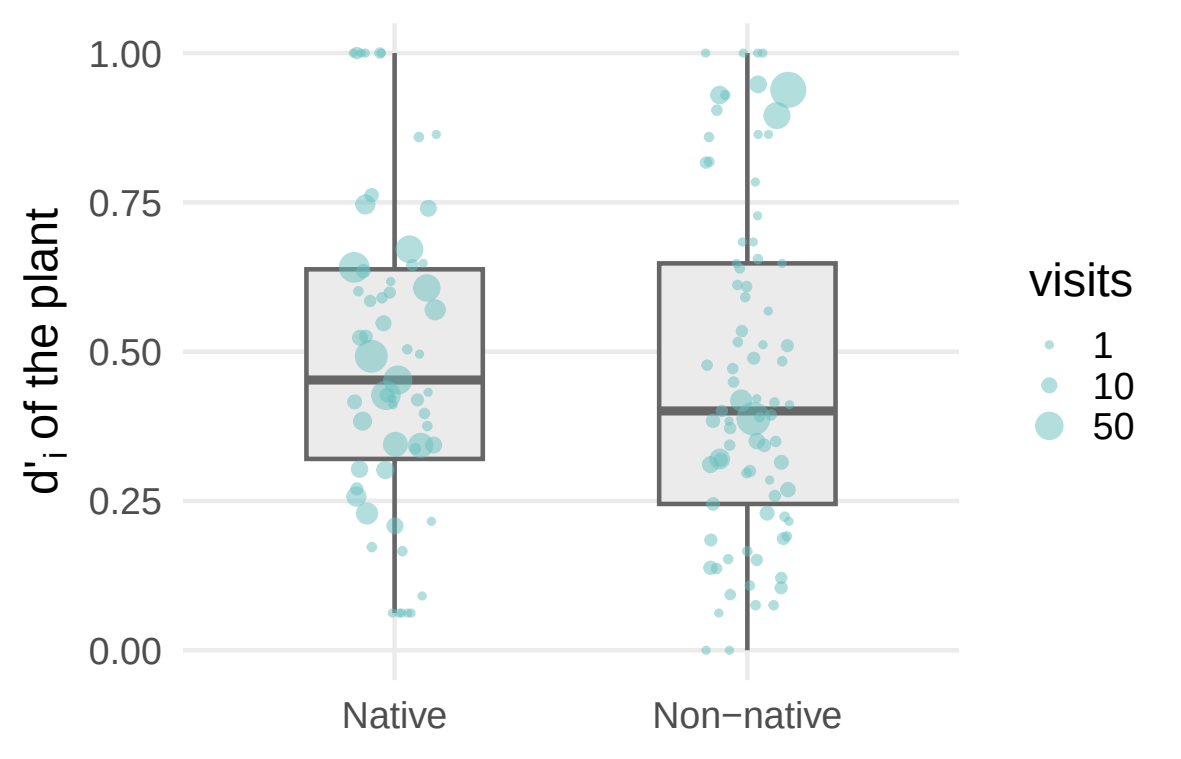
$$\boxed{d'_i = \frac{\delta_i - \delta_{\min}}{\delta_{\max} - \delta_{\min}} \in [0, 1]}$$

- $d'_i = 0$: visited in proportion to availability (generalist)
- $d'_i = 1$: deviates as much as possible (specialist)

Specialisation index d'

UK data

d'_i of each plant i (with respect to the insects) for the **Bristol count network**



→ Native and non-native plants: same specialisation (medians 0.45 and 0.40)

Paths, distances and centralities

- Path from i to j: sequence of edges $i = u_0 \to u_1 \to \cdots \to u_k = j$
- Distance $\delta(i,j)$: length of the shortest path; connected component: maximal set of nodes linked by paths

Centralities

- Closeness: $c_i = \left(\sum_j \delta(i,j) \right)^{-1}$
- Betweenness [Freeman, 1977]: $b_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$
  - σ_{jk} = number of shortest paths from j to k
  - $\sigma_{jk}(i)$ = those going through i

- Interaction network: connector species linking otherwise separated groups

Motivation

Introduction to networks

- Graphs and matrices
- Node-level descriptors
- Network-level descriptors

Introduction to stochastic block models

Modelling collections of networks

Conclusion

Connectance

Connectance (density)

$$C = \frac{L}{n\,m}, \qquad L = \sum_{i,j} \mathbf{1}_{\{Y_{ij}>0\}}$$

(undirected unipartite: $C = \frac{2L}{n(n-1)}$). Proportion of realised interactions.

UK data

- Leeds (155×126): $C = 0.029$
- Bristol (193×180): $C = 0.019$

→ In general, people observe that $C \searrow$ as the size of the network $\nearrow$

→ Comparing C between networks of different size is not straightforward

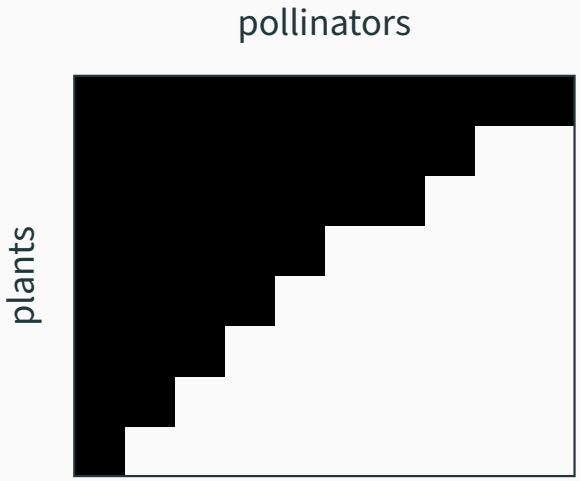
Interaction evenness

Interaction evenness

$$E = \frac{-\sum_{(i,j):Y_{ij}>0} p_{ij} \log p_{ij}}{\log L}, \qquad p_{ij} = \frac{Y_{ij}}{\sum_{i',j'} Y_{i'j'}}$$

- Shannon entropy
- $E = 1$: all realized links have the same frequency.

Nestedness



Perfectly nested matrix (sorted by degree)

- Partners of specialists $\subset$ partners of generalists [Bascompte et al., 2003]
- Specialists interact with generalists
- Ecological interpretation: linked to robustness to species loss

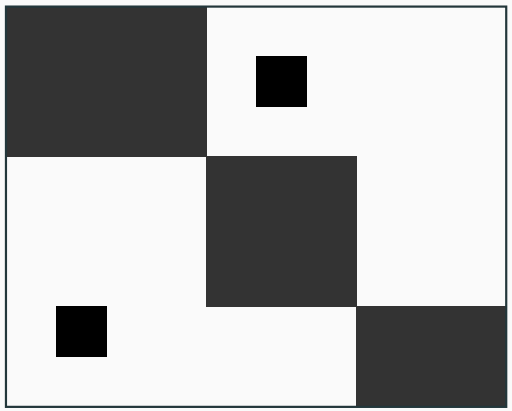
NODF [Almeida-Neto et al., 2008]

Sort the matrix by $\nearrow$ degrees in columns and $\searrow$ in rows:
For any pairs of rows: i, i'

$$N_{ii'} = \begin{cases} 0 & \text{if } d_i \leq d_{i'} \\ 100 \times \frac{\#\{j : Y_{ij} > 0, Y_{i'j} > 0\}}{d_{i'}} & \text{if } d_i > d_{i'} \end{cases}$$

NODF = mean of N over all pairs of rows and all pairs of columns $\in [0, 100]$

Modularity



Modular matrix

- Groups of species interacting more within than between groups (compartments)
- Interpretation: phenology, morphological matching (tongue length / corolla depth), habitat

Calculation by testing all the possible partitions of the nodes.

$\rightarrow$ Computationally intensive

Descriptors per city

UK data

City	*n* plants	*m* insects	*L*	visits	*C*	*E*	NODF
Bristol	193	180	662	1379	0.019	0.90	10.1
Edinburgh	137	129	425	1507	0.024	0.73	12.9
Leeds	155	126	573	1079	0.029	0.94	13.3
Reading	156	167	574	956	0.022	0.95	12.6

- *C*: connectance, *E*: interaction evenness, NODF: nestedness

→ Are the differences between cities significant? Are they due to the ecology or to the size of the networks?

The need for null models

- An index alone has no reference value: is NODF = 20 high?
- Most indices depend on n, m and C

→ compare the observed value $T(Y^{obs})$ with its distribution under a random model keeping some features of the data

Principle

1. Simulate $Y^{(1)}, \ldots, Y^{(B)}$ under the null model $\mathcal{M}_0$
2. Compute $T(Y^{(b)})$
3. $z = \frac{T(Y^{obs}) - \overline{T}}{\mathrm{sd}(T)}$, $\quad p = \frac{1 + \#\{b : T(Y^{(b)}) \geq T(Y^{obs})\}}{B+1}$

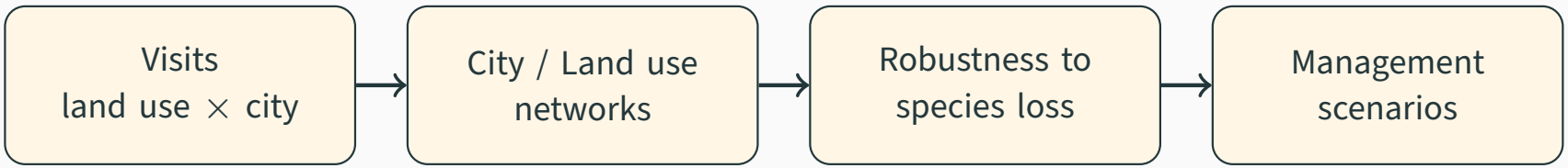
The conclusion depends on what $\mathcal{M}_0$ keeps.

→ If you want to know more: apply to the **CESAB training**!

What [Baldock et al., 2019] do?

UK data

The network is used to measure robustness



Bayesian network models: pollinator dispersal and resource switching between land uses.

→ Here: the structure of the networks itself.

Summary of Part 1

- A network = a matrix Y (binary / weighted, square / rectangular, symmetric or not)
- Node descriptors: degree, strength, specialisation, centralities
- Network descriptors: connectance, evenness, nestedness, modularity
- Each index looks for a specific pattern, chosen a priori
- Indices depend on size $\rightarrow$ null models to assess the significativity

What we would like

A probabilistic model adapted to our observations Y

- flexible enough to reproduce heterogeneity (generalists / specialists),
- without choosing the structure a priori,
- interpretable in terms of species roles,
- with the statistical toolbox: estimation, model selection, comparison.

Motivation

Introduction to networks

Introduction to stochastic block models

- Stochastic Block Model
- Bipartite stochastic block model
- Estimation of the parameters
- Selection of the numbers of blocks
- Application to UK dataset

Modelling collections of networks

Conclusion

Probabilistic approach

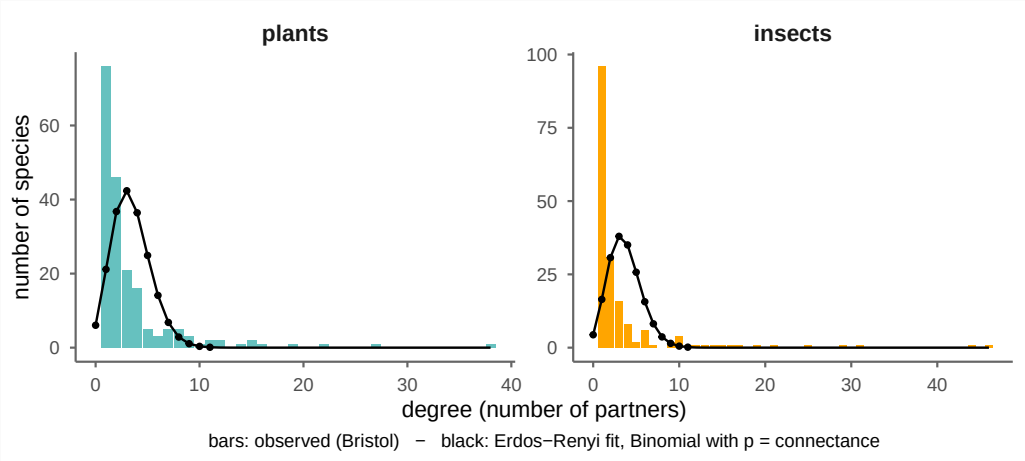
- Context: the matrix Y is the realisation of a stochastic process
- Aim: propose a stochastic process able to mimic the heterogeneity of the connections
- Advantage: benefit from the statistical tools (estimation, tests, model selection…)

The simplest model: [Erdős and Rényi, 1959]

$$\forall i, j, \quad Y_{ij} \overset{i.i.d.}{\sim} \mathcal{B}ern(p)$$

- One parameter p, estimated by the connectance: $\hat{p} = C$
- All species are equivalent: $d_i \sim \mathcal{B}in(m, p), \quad d_j \sim \mathcal{B}in(n, p)$

Bristol $\hat{p} = 0.019$



→ ER misses the heterogeneity of the species: *Apis*: 46 partners, many species with very few interactions.

Motivation

Introduction to networks

Introduction to stochastic block models

- Stochastic Block Model
- Bipartite stochastic block model
- Estimation of the parameters
- Selection of the numbers of blocks
- Application to UK dataset

Modelling collections of networks

Conclusion

Stochastic Block Model

[Nowicki and Snijders, 2001] Let (Y_{ij}) be an adjacency matrix (n nodes).

Latent variables

- The nodes $i = 1, \ldots, n$ are partitioned into K clusters
- $Z_i = k$ if node i belongs to cluster (block) k
- Z_i independent variables

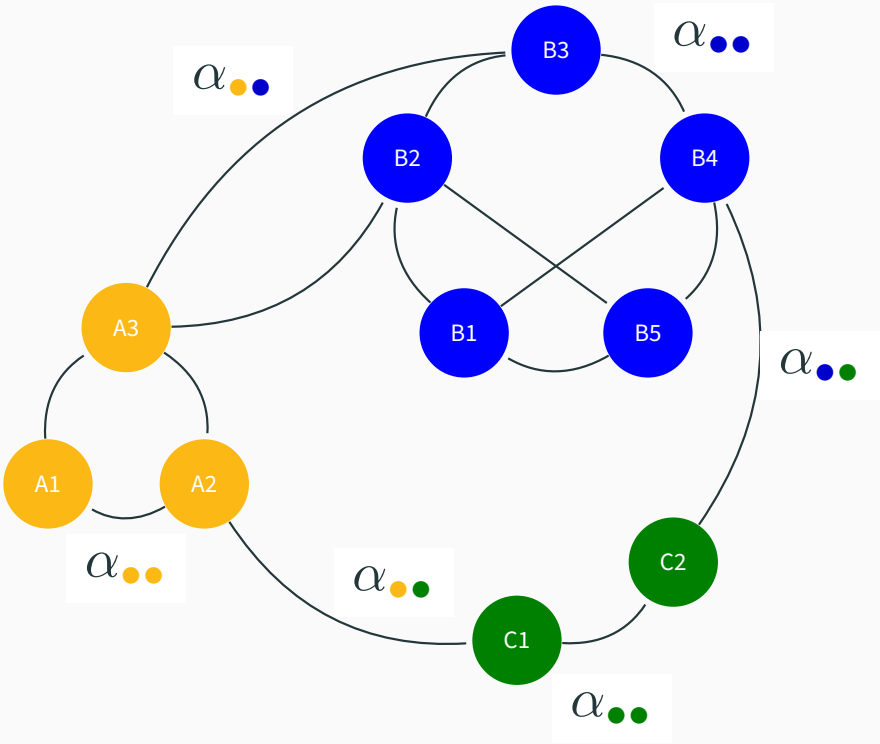
$$\mathbb{P}(Z_i = k) = \pi_k$$

Conditionally to $(Z_i)_{i=1,\ldots,n}$...

(Y_{ij}) independent and

$$Y_{ij}|Z_i, Z_j \sim \mathcal{B}ern(\alpha_{Z_i,Z_j}) \quad \Leftrightarrow \quad P(Y_{ij} = 1|Z_i = k, Z_j = \ell) = \alpha_{k\ell}$$

Stochastic Block Model : illustration



Parameters

Let n nodes divided into 3 clusters

- $\mathcal{K} = \{\bullet, \bullet, \bullet\}$ clusters
- $\pi_\bullet = \mathbb{P}(i \in \bullet), \bullet \in \mathcal{K}, i = 1, \ldots, n$
- $\alpha_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

$$Z_i = \mathbf{1}_{\{i \in \bullet\}} \sim^{\text{iid}} \mathcal{M}(1, \pi), \quad \forall \bullet \in \mathcal{K},$$

$$Y_{ij} \mid \{i \in \bullet, j \in \bullet\} \sim^{\text{ind}} \mathcal{B}(\alpha_{\bullet\bullet})$$

Motivation

Introduction to networks

Introduction to stochastic block models

- Stochastic Block Model
- Bipartite stochastic block model
- Estimation of the parameters
- Selection of the numbers of blocks
- Application to UK dataset

Modelling collections of networks

Conclusion

Bipartite SBM (Latent Block Model)

Y: $n \times m$ incidence matrix, plants in rows, pollinators in columns.

Latent variables: bi-clustering

- Plants $i = 1, \dots, n$ in K groups: $\mathbb{P}(Z_i = k) = \pi_k$
- Pollinators $j = 1, \dots, m$ in L groups: $\mathbb{P}(W_j = \ell) = \rho_\ell$
- $(Z_i)_i$, $(W_j)_j$ independent

Conditionally to $(Z_i)_{i=1,\dots,n}$, $(W_j)_{j=1,\dots,m}$

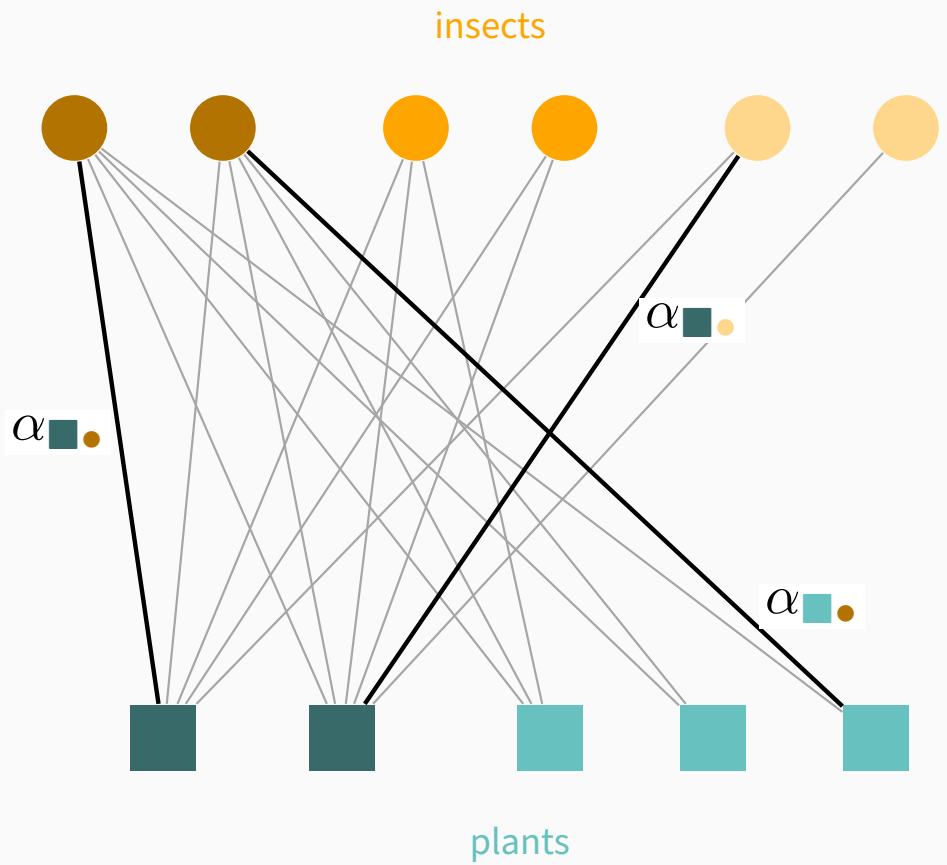
(Y_{ij}) independent and

$$Y_{ij} | Z_i = k, W_j = \ell \sim \mathcal{B}ern(\alpha_{k\ell})$$

Parameters: $\theta = (\pi, \boldsymbol{\rho}, \boldsymbol{\alpha})$, $\quad K - 1 + L - 1 + KL$ parameters.

[Govaert and Nadif, 2008]

Bipartite SBM: illustration



Parameters

- Plant groups $\{\blacksquare, \blacksquare\}$, $\pi_k = \mathbb{P}(i \in k)$
- Insect groups $\{\bullet, \bullet, \bullet\}$, $\rho_\ell = \mathbb{P}(j \in \ell)$
- $\alpha_{\blacksquare\bullet} = \mathbb{P}(i \leftrightarrow j \mid i \in \blacksquare, j \in \bullet)$

- Links only between a plant and an insect: α is $K \times L$, not symmetric

Ecological interpretation

- A block = a group of species with the same interaction profile (stochastic equivalence): an ecological role, not a taxonomic group
- $\alpha_{k\ell}$ = probability that a plant of group k and a pollinator of group ℓ interact
- π_k, ρ_ℓ = proportions of species playing each role

Link with the descriptors of Part 1

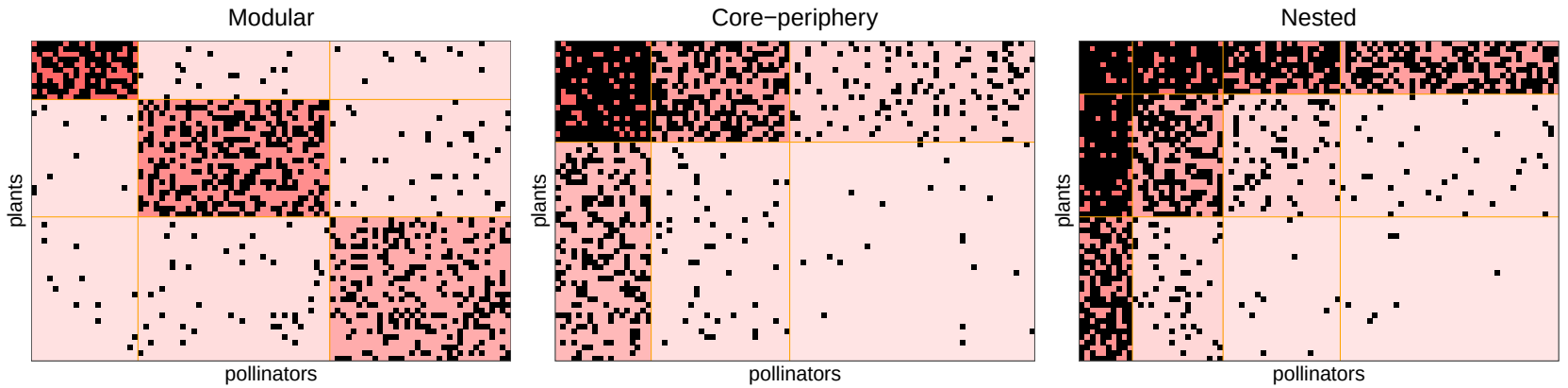
Degrees: one distribution for each type of nodes

- Plant i of group k: $d_i \mid Z_i = k \sim \mathcal{B}in\left(m, \sum_\ell \rho_\ell \, \alpha_{k\ell}\right)$
- Insect j of group ℓ: $d_j \mid W_j = \ell \sim \mathcal{B}in\left(n, \sum_k \pi_k \, \alpha_{k\ell}\right)$

$\rightarrow$ A mixture of binomials on each side: generalists and specialists

- Expected connectance: $\mathbb{E}[C] = \sum_{k,\ell} \pi_k \, \rho_\ell \, \alpha_{k\ell}$
- $K = L = 1$: Erdős–Rényi

One model, many structures



$$\alpha = \begin{pmatrix} .60 & .05 & .02 \\ .03 & .40 & .05 \\ .02 & .04 & .25 \end{pmatrix} \qquad \alpha = \begin{pmatrix} .85 & .45 & .15 \\ .30 & .05 & .01 \end{pmatrix} \qquad \alpha = \begin{pmatrix} .95 & .85 & .70 & .50 \\ .85 & .45 & .15 & .03 \\ .60 & .10 & .02 & .01 \end{pmatrix}$$

$n = 60$ plants, $m = 90$ pollinators. α is $K \times L$: neither square nor symmetric.

$\rightarrow$ No a priori on the structure: modularity, nestedness, core-periphery are particular cases of α.

Valued networks: counts of visits

Keep the same latent structure, change the distribution of Y_{ij}:

- Poisson: $Y_{ij} \mid Z_i = k, W_j = \ell \sim \mathcal{P}(\lambda_{k\ell})$ (number of visits)
- Gaussian, zero-inflated… [Mariadassou et al., 2010]

What if we have covariates?

Covariates on nodes / species: x_i for plant i, x_j for insect j

- plants: floral abundance, native / non-native, traits
- insects: abundance, group, traits (body size)

A vector of covariates x_{ij} on each pair (plant i, insect j):

- trait matching: tongue length vs corolla depth
- phenological overlap

SBM with covariates

Covariates drive the memberships

$$\mathbb{P}(Z_i = k \mid x_i) = \frac{e^{\gamma_k^\top x_i}}{\sum_{k'} e^{\gamma_{k'}^\top x_i}} \qquad \mathbb{P}(W_j = \ell \mid x_j) = \frac{e^{\delta_\ell^\top x_j}}{\sum_{\ell'} e^{\delta_{\ell'}^\top x_j}}$$

$$Y_{ij} \mid Z_i = k, W_j = \ell \sim \mathcal{P}(\alpha_{k\ell})$$

$\rightarrow$ blocks explained by the covariates

Covariates drive the interactions

$$Y_{ij} \mid Z_i = k, W_j = \ell \sim \mathcal{B}\left(\text{logit}^{-1}(\alpha_{k\ell} + \beta^\top x_{ij})\right)$$

$\rightarrow$ blocks = structure not explained by the covariates

Motivation

Introduction to networks

Introduction to stochastic block models

- Stochastic Block Model
- Bipartite stochastic block model
- Estimation of the parameters
- Selection of the numbers of blocks
- Application to UK dataset

Modelling collections of networks

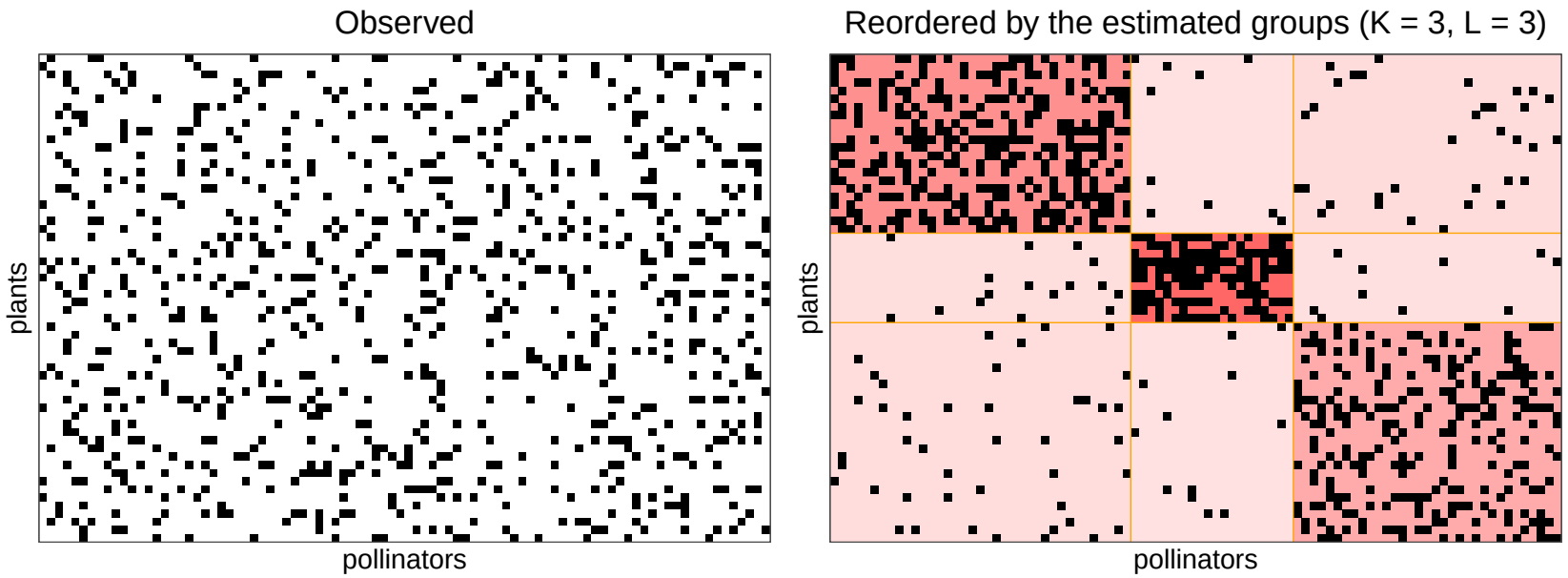
Conclusion

Statistical inference

From the observed Y, we want:

1. the numbers of groups K and L
2. the parameters $\widehat{\boldsymbol{\theta}} = (\widehat{\pi}, \widehat{\boldsymbol{\rho}}, \widehat{\boldsymbol{\alpha}})$
3. the clustering $\widehat{Z}_i$, $\widehat{W}_j$: which species play the same role?
4. Which is this role? $\rightarrow$ interpretation of $\widehat{\boldsymbol{\alpha}}$

What inference does



Simulated modular network (60 plants $\times$ 90 pollinators): the same matrix, rows and columns in the observed order (left) and grouped by the estimated $\widehat{Z}_i$, $\widehat{W}_j$ (right)

$\rightarrow$ The structure is hidden in the observed order: inference recovers it

Complete and marginal likelihoods

Complete likelihood: Y and the latent groups $(\mathbf{Z}, \mathbf{W})$

$$\begin{aligned}\ell_c(Y, \mathbf{Z}, \mathbf{W}; \theta) &= p(Y \mid \mathbf{Z}, \mathbf{W}; \alpha)\, p(\mathbf{Z}; \pi)\, p(\mathbf{W}; \rho) \\ &= \prod_{i,j} \alpha_{Z_i W_j}^{Y_{ij}} (1 - \alpha_{Z_i W_j})^{1 - Y_{ij}} \prod_i \pi_{Z_i} \prod_j \rho_{W_j}\end{aligned}$$

Marginal (observed) likelihood: Y only

$$\log \ell(Y; \theta) = \log \sum_{\mathbf{Z}, \mathbf{W}} \ell_c(Y, \mathbf{Z}, \mathbf{W}; \theta)$$

- $K^n L^m$ terms in the sum: impossible to compute
- EM needs $p(\mathbf{Z}, \mathbf{W} \mid Y; \theta)$, but given Y the groups are not independent $\rightarrow$ not tractable either

A lower bound of the likelihood (ELBO)

For any distribution $\mathcal{R}$ on $(\mathbf{Z}, \mathbf{W})$:

Central identity

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{\theta}, \mathcal{R}) &= \log \ell(Y; \boldsymbol{\theta}) - \mathrm{KL}\left[\mathcal{R} \,\|\, p(\cdot \mid Y; \boldsymbol{\theta})\right] \;\leq\; \log \ell(Y; \boldsymbol{\theta}) \\ &= \mathbb{E}_{\mathcal{R}}\left[\log \ell_c(Y, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta})\right] + \mathcal{H}(\mathcal{R})\end{aligned}$$

▸ Proof

- $\mathcal{H}(\mathcal{R}) = -\mathbb{E}_{\mathcal{R}}[\log \mathcal{R}(\mathbf{Z}, \mathbf{W})]$: entropy
- Equality iff $\mathcal{R} = p(\cdot \mid Y; \boldsymbol{\theta})$
- Second line: expectation of the complete log-likelihood
  $\rightarrow$ computable if $\mathcal{R}$ allows it

$\rightarrow$ Maximise ELBO instead of $\log \ell$

Variational EM

Choose $\mathcal{R}$ such that $\mathbb{E}_{\mathcal{R}}\left[\log \ell_c(Y, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta})\right]$ can be evaluated

Mean field: groups independent under $\mathcal{R}$

$$\mathcal{R}_{\tau,\boldsymbol{\eta}}(\mathbf{Z}, \mathbf{W}) = \prod_i \mathcal{M}(Z_i; \tau_i) \prod_j \mathcal{M}(W_j; \eta_j), \qquad \tau_{ik} \approx \mathbb{P}(Z_i = k \mid Y)$$

At iteration t:

- VE-step: $(\tau, \boldsymbol{\eta})^{(t)} = \arg\max_{\tau,\boldsymbol{\eta}} \mathcal{I}(\boldsymbol{\theta}^{(t-1)}, \mathcal{R}_{\tau,\boldsymbol{\eta}})$ ▸ Details
- M-step: $\boldsymbol{\theta}^{(t)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{R}^{(t)}}[\log \ell_c(Y, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta})]$ ▸ Details

Note that VE-step $\Leftrightarrow$ minimising $\mathrm{KL}[\mathcal{R} \,\|\, p(\cdot \mid Y; \boldsymbol{\theta})]$

Fast, but depends on the initialisation [Daudin et al., 2008, Leger, 2016].

Motivation

Introduction to networks

Introduction to stochastic block models

- Stochastic Block Model
- Bipartite stochastic block model
- Estimation of the parameters
- Selection of the numbers of blocks
- Application to UK dataset

Modelling collections of networks

Conclusion

Model selection: principle

- One model $\mathcal{M}_{K,L}$ per pair (K, L) of numbers of groups
- More groups $\rightarrow$ the maximised likelihood always increases
  $\rightarrow$ overfitting

Penalised likelihood

$$\mathrm{Crit}(\mathcal{M}) = \underbrace{\log \ell(Y; \hat{\theta}_{\mathcal{M}})}_{\text{fit}} - \underbrace{\mathrm{pen}(\mathcal{M})}_{\text{complexity}}, \qquad \hat{\mathcal{M}} = \arg\max_{\mathcal{M}} \mathrm{Crit}(\mathcal{M})$$

BIC [Schwarz, 1978]

$$\mathrm{BIC}(\mathcal{M}) = \log \ell(Y; \hat{\theta}_{\mathcal{M}}) - \frac{\dim(\theta_{\mathcal{M}})}{2} \log(\text{nb. of observations})$$

$\rightarrow$ But $\log \ell(Y; \theta)$ is not computable in the SBM...

BIC on the complete data

If $(\mathbf{Z}, \mathbf{W})$ were known: BIC on the complete likelihood, each parameter penalised by its own number of observations

$$\log \ell_c(Y, \mathbf{Z}, \mathbf{W}; \hat{\theta}) - \frac{1}{2}\Big\{ \underbrace{(K-1)\log n}_{\pi\text{: } n \text{ plants}} + \underbrace{(L-1)\log m}_{\boldsymbol{\rho}\text{: } m \text{ pollinators}} + \underbrace{KL \log(nm)}_{\boldsymbol{\alpha}\text{: } nm \text{ pairs}} \Big\}$$

$(\mathbf{Z}, \mathbf{W})$ unknown: two ways to get rid of them

MAP: plug-in

$\hat{Z}_i = \arg\max_k \hat{\tau}_{ik}$, $\hat{W}_j = \arg\max_\ell \hat{\eta}_{j\ell}$

$$\log \ell_c(Y, \hat{\mathbf{Z}}, \hat{\mathbf{W}}; \hat{\theta}) - \text{pen}$$

Integration

Expectation under $\mathcal{R}_{\hat{\tau},\hat{\boldsymbol{\eta}}} \approx p(\cdot \mid Y; \hat{\theta})$

$$\mathbb{E}_{\mathcal{R}}\big[\log \ell_c(Y, \mathbf{Z}, \mathbf{W}; \hat{\theta})\big] - \text{pen}$$

$\rightarrow$ ICL: Integrated Completed Likelihood [Biernacki et al., 2000]. Package `sbm`: integration.

ICL versus BIC

With the same penalty, and since $\log \ell_c = \log \ell + \log p(\mathbf{Z}, \mathbf{W} \mid Y)$:

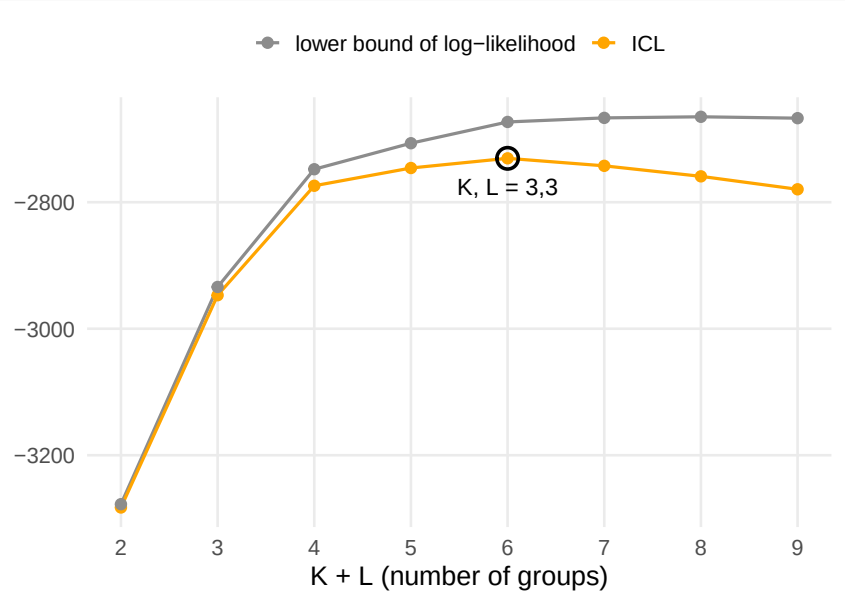
$$\mathrm{ICL}(\mathcal{M}) \approx \mathrm{BIC}(\mathcal{M}) - \underbrace{\mathcal{H}\big(p(\cdot \mid Y; \hat{\boldsymbol{\theta}})\big)}_{\text{entropy} \geq 0}$$

- Entropy small when the groups are well separated ($\tau_{ik} \approx 0$ or 1)

→ ICL favours clear clusterings: fewer groups than BIC

- Models explored by split / merge of groups

Bristol binary network



In practice: R package `sbm`

```
library(sbm)
## B: binary plants x insects matrix (Bristol)
fit <- estimateBipartiteSBM(netMat = B, model = "bernoulli",
         dimLabels = c(row = "plants", col = "insects"))
fit$nbBlocks # K and L selected by ICL
fit$blockProp # pi and rho
fit$connectParam$mean # alpha
fit$memberships # groups of each species
plot(fit, type = "data") # reordered matrix
plot(fit, type = "expected") # fitted probabilities
```

- `model = "poisson"` for counts, `covariates = list(...)` for covariates

https://grosssbm.github.io/sbm/

Motivation

Introduction to networks

Introduction to stochastic block models

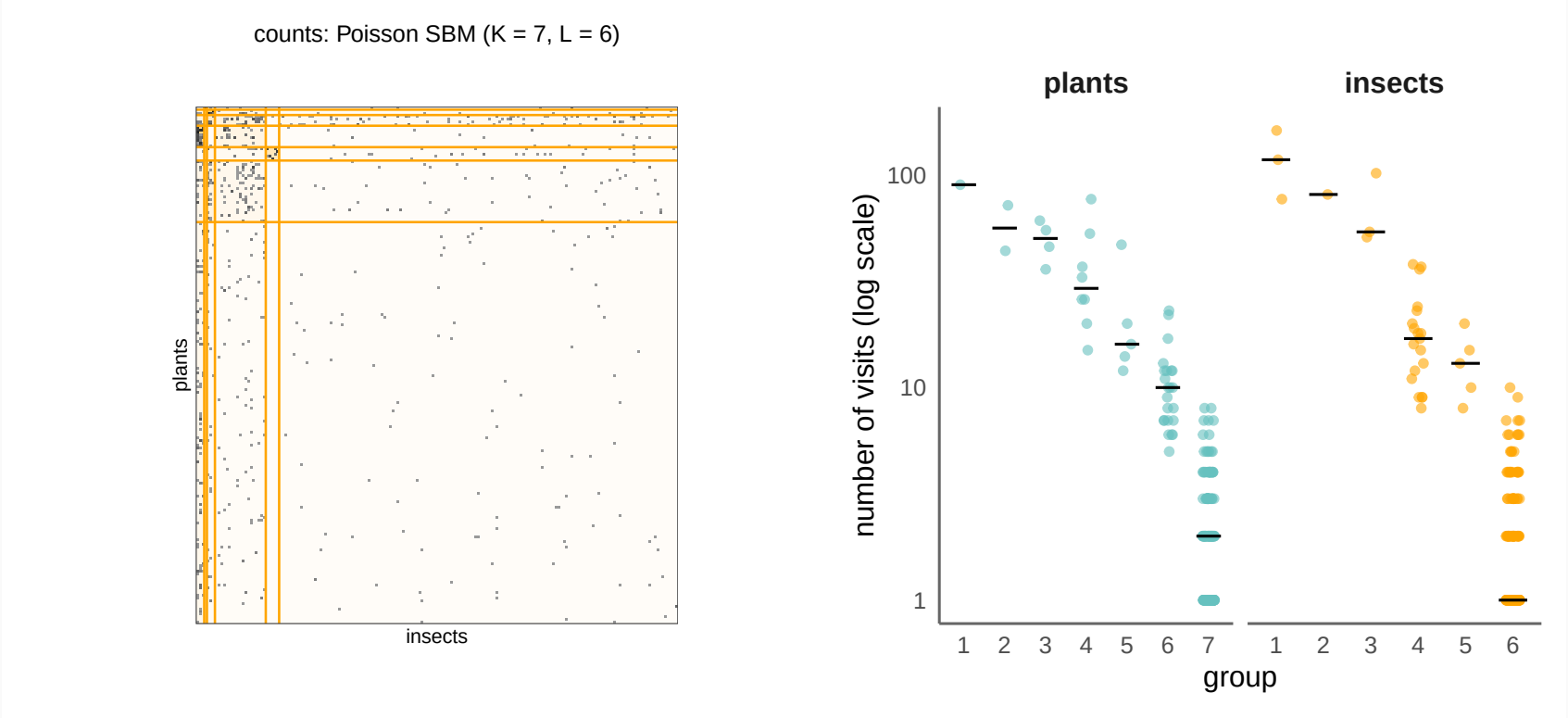
- Stochastic Block Model
- Bipartite stochastic block model
- Estimation of the parameters
- Selection of the numbers of blocks
- Application to **UK dataset**

Modelling collections of networks

Conclusion

Bristol: SBM on the counts

UK data

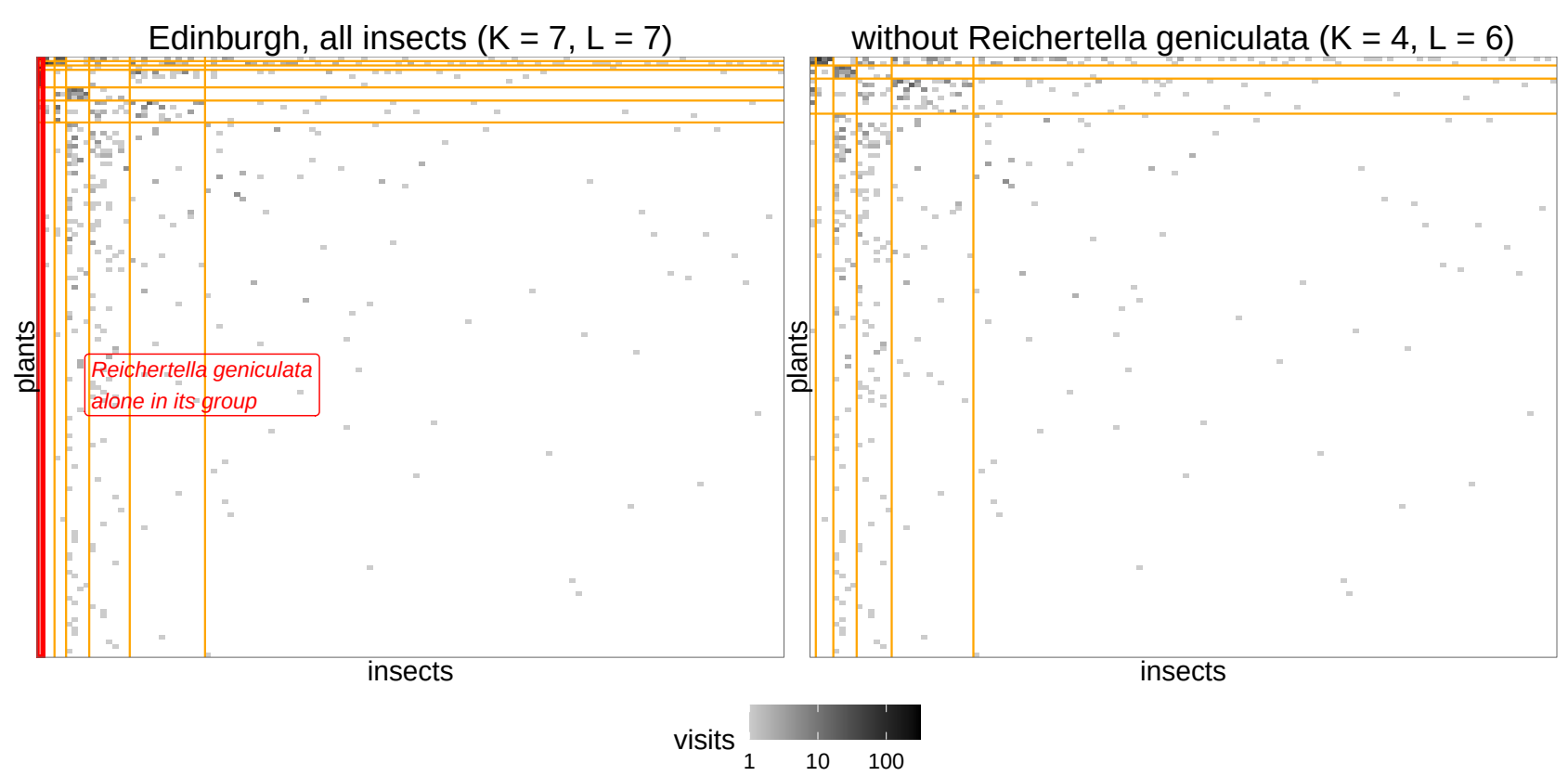


Poisson SBM, Y_{ij} = number of visits. Right: number of visits of each species, by group.

→ The groups mostly sort species by strengh.

Edinburgh: one insect drives the counts

UK data



Reichertella geniculata: 300 visits on *Taraxacum* at one site, 162 on *Cirsium arvense* at another.

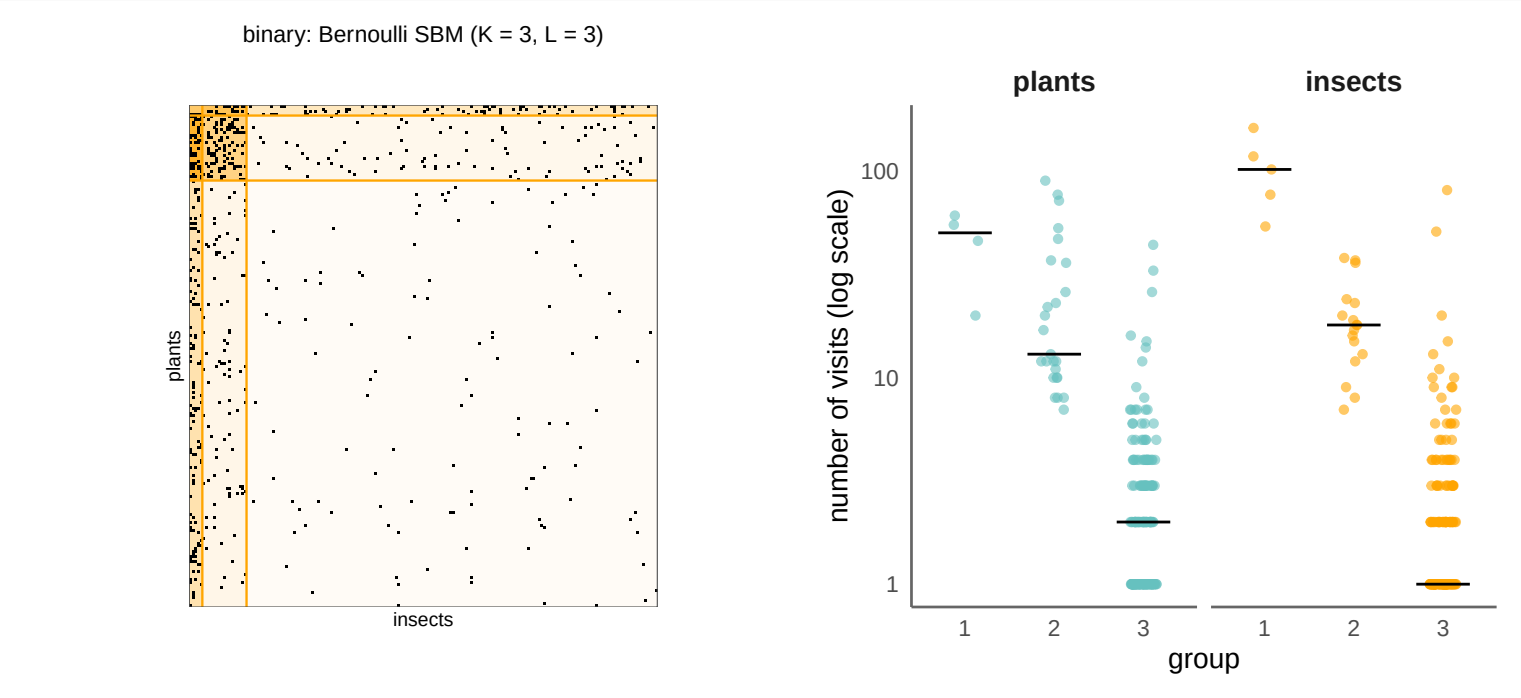
→ Poisson SBM: 7×7 groups with this fly, 4×6 without.

→ Counts reflect local aggregations: prefer the binary network.

Bristol: SBM on the binary network

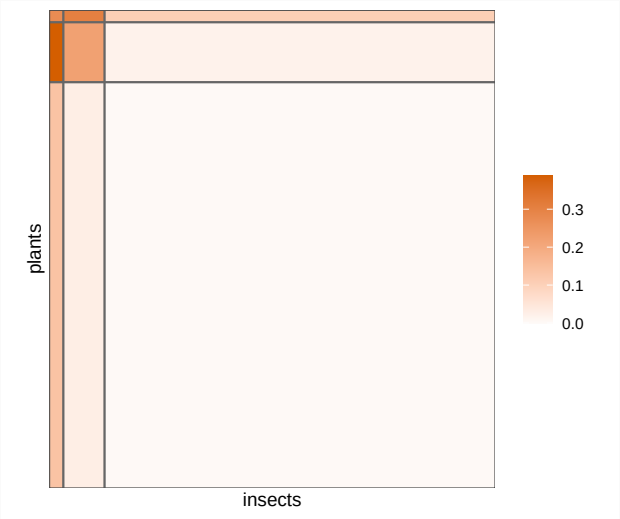
UK data

Counts mix who with whom and abundance → switch to presence / absence



Bristol: interpretation

UK data

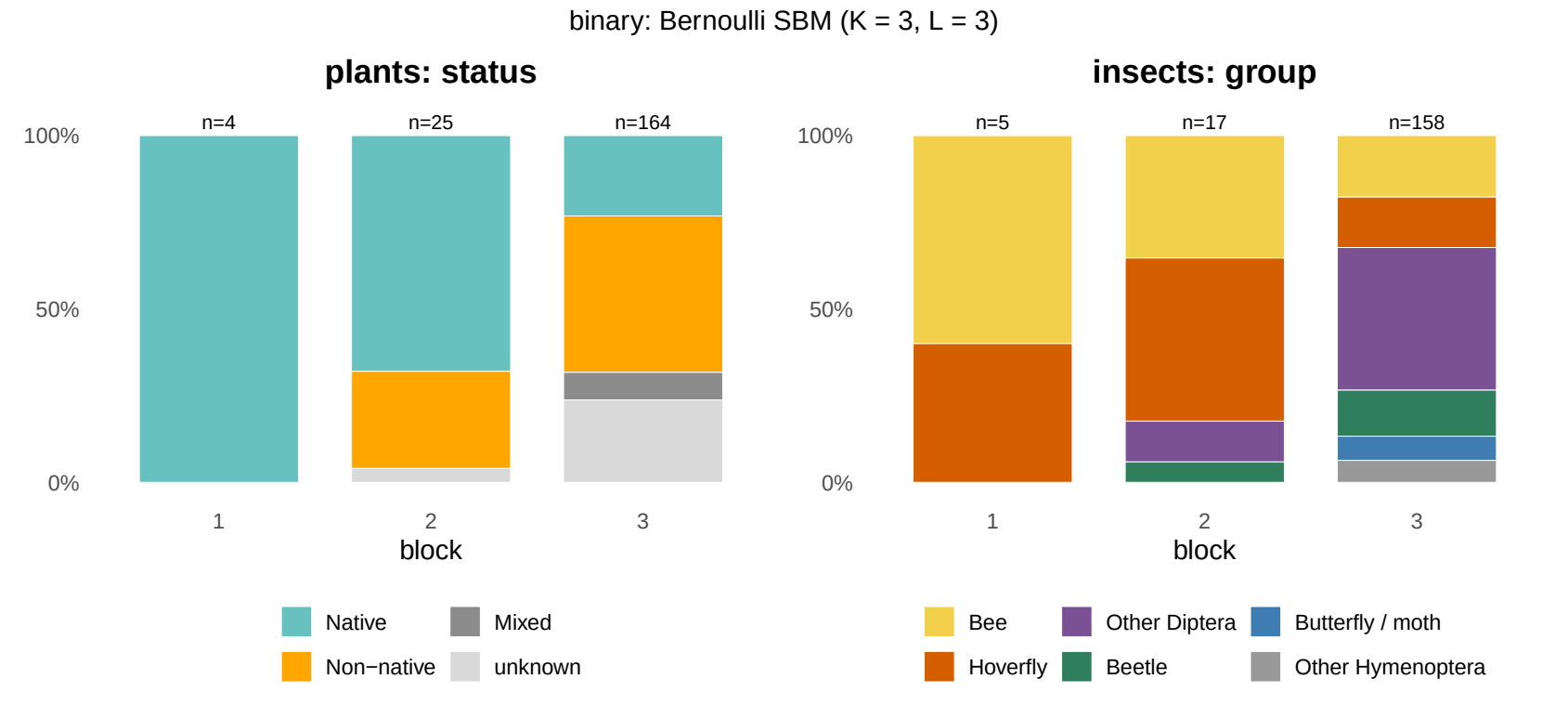




- Core–periphery structure rather than modules (3 × 3 groups)
- 5 generalist insects (*Apis mellifera*, *Bombus pascuorum*, *B. terrestris*, *Eristalis*) visiting most plant groups
- 4 hub plants: *Bellis*, *Taraxacum*, *Ranunculus repens*, *Heracleum*, visited even by the less observed insects
- Large peripheral groups: 164 plants, 158 insects with few links (specialists or under-observed)

Bristol: composition of the groups

UK data



- Core plants: native.
- *Bellis, Taraxacum, Ranunculus repens, Heracleum*: Common European weeds and wildflowers.
- Non-native plants: mostly in the periphery
- Core insects: bees and hoverflies

One SBM per city

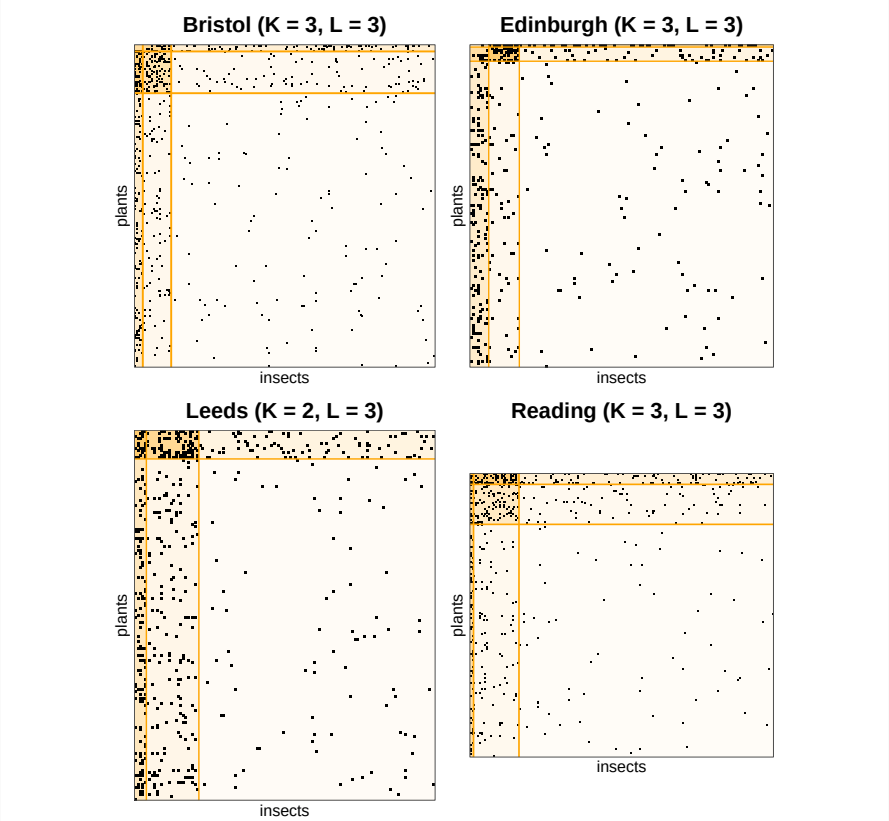
UK data

	Bristol	Edinburgh	Leeds	Reading
Bernoulli	3×3	3×3	2×3	3×3
Poisson	7×6	7×7	5×6	5×6

Each city is fitted separately: the groups (and their numbers) are not comparable from one city to the other.

One SBM per city: binary networks

UK data



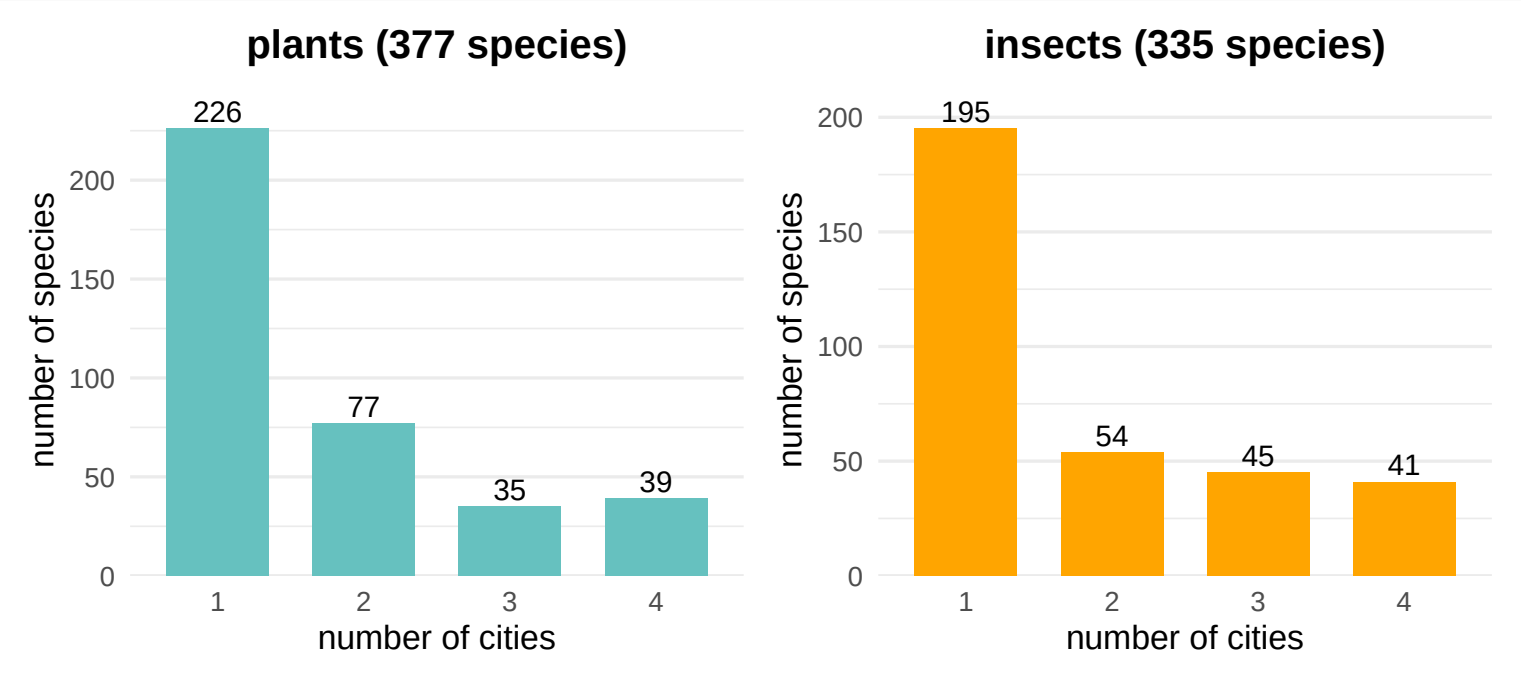
Background: fitted $\alpha_{k\ell}$ (same scale for the 4 cities)

→ The same core–periphery structure in the 4 cities?

→ How can we compare them, since they do not share their species?

Most species are seen in one city

UK data



- In the 4 cities: only 39 plants and 41 insects
- Groups cannot be matched species by species

Towards collections of networks

- Same type of interaction, several cities / land uses
- Questions: do the networks share the same structure? Does a species keep its role from one city to another? Which land uses differ?

→ Next section.

Stochastic block models in a nutshell

- Generative model: easy to simulate, natural null model
- Flexible: no a priori on the structure; modularity, nestedness, core-periphery are particular cases
- Linked to the classical descriptors (degrees, connectance)
- Interpretable: groups of species with the same role
- Extensions: valued edges, covariates on nodes or pairs of nodes.

References: review on SBM [Lee and Wilkinson, 2019], other models for heterogeneity [Matias and Robin, 2014]

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

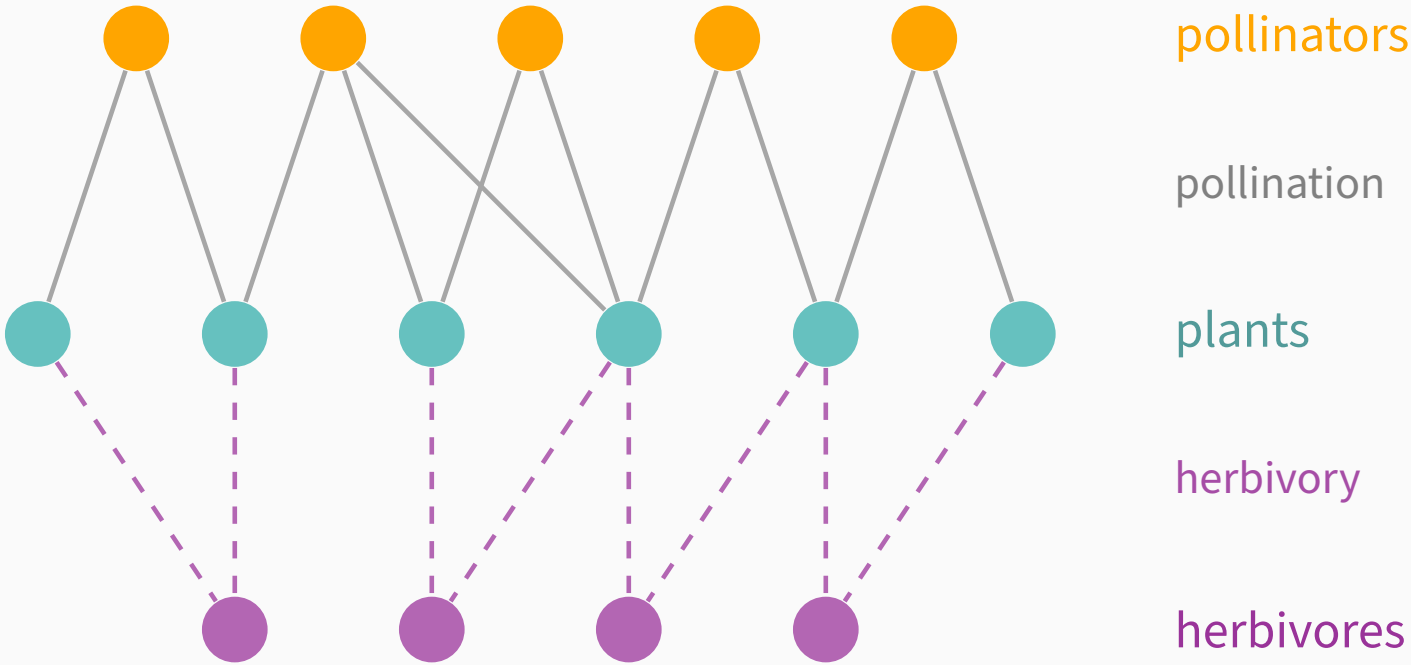
When several networks coexist

Need for the analysis of ecological systems to include multiple layers of complexity:

[Pilosof et al., 2017]

Multipartite networks

More than two types of nodes

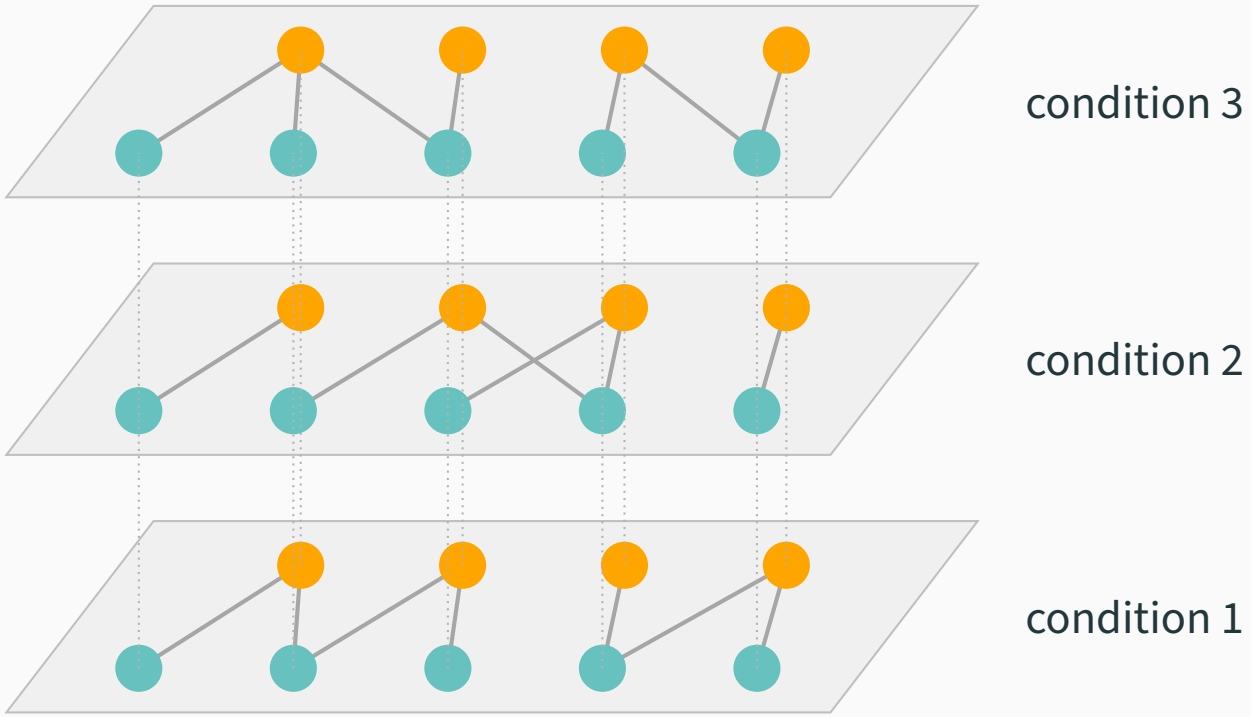


→ The plants take part in both interactions

[Bar-Hen et al., 2022], [Dáttilo et al., 2016]

Multiplex networks

The same plants and the same insects in several conditions: $Y^{(1)}, \ldots, Y^{(M)}$

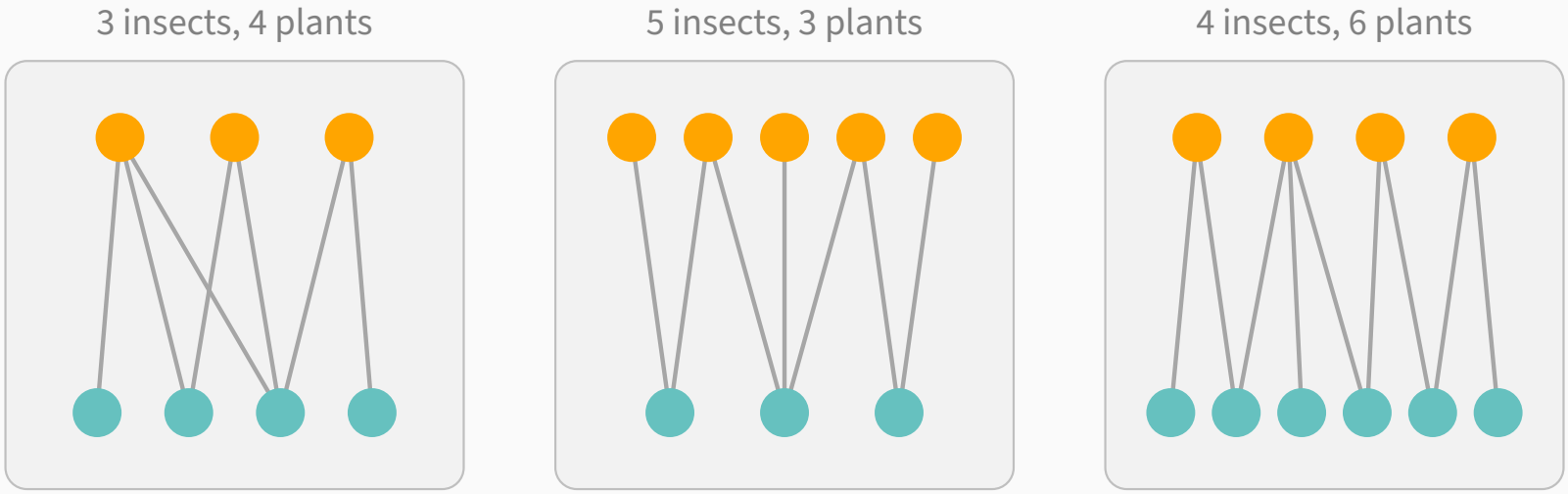


$\rightarrow$ e.g. several years, seasons, sampling rounds, several types of relations

[Kéfi et al., 2016]

Multilayer networks: collections

The same type of interaction, not the same nodes



→ e.g. the 4 cities: different species, different sizes

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

A collection of bipartite networks

S networks $Y^{(1)}, \ldots, Y^{(S)} \rightarrow$ network s: n_s plants $\times$ m_s insects

Questions

- Do the networks share a common structure?
- Which species play the same role across the networks?
- Are some groups specific to some networks?

$\rightarrow$ One SBM per network: the groups are not comparable

First model: iid-colBiSBM

iid-colBiSBM

$$Y^{(s)} \sim_{\text{ind}} \text{biSBM}_{n_s,m_s}(K, L, \pi, \boldsymbol{\rho}, \boldsymbol{\alpha}), \quad s = 1, \ldots, S$$

- The same $\boldsymbol{\alpha}$ for all the networks

$\rightarrow$ A group = the same role in all the networks

- But the same proportions π, $\boldsymbol{\rho}$ everywhere: too restrictive?

First proposed for collections of simple (unipartite) networks: colSBM [Chabert-Liddell et al., 2024]

Free block proportions: $\pi\rho$-colBiSBM

$\pi\rho$-colBiSBM

$$Y^{(s)} \sim_{\text{ind}} \text{biSBM}_{n_s,m_s}(K, L, \pi^{(s)}, \boldsymbol{\rho}^{(s)}, \boldsymbol{\alpha})$$

- Common connection structure $\boldsymbol{\alpha}$
- Proportions of the groups specific to each network
- $\pi_k^{(s)} = 0$ allowed: group k absent from network s

$\pi\rho$-colBiSBM: different proportions

$S = 2$ networks, $K = 3$ groups of plants, $L = 4$ groups of insects

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \end{pmatrix}$$

Plants	Insects
$\pi^{(1)} = [.25, .25, .5]$	$\boldsymbol{\rho}^{(1)} = [.1, .3, .4, .2]$
$\pi^{(2)} = [.2, .5, .3]$	$\boldsymbol{\rho}^{(2)} = [.3, .1, .4, .2]$

$\rightarrow$ Same roles, in different proportions

$\pi\rho$-colBiSBM: nested structures

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \end{pmatrix}$$

Plants	Insects
$\pi^{(1)} = [.25, .25, .5]$	$\boldsymbol{\rho}^{(1)} = [.1, .3, .4, .2]$
$\pi^{(2)} = [.5, .5, 0]$	$\boldsymbol{\rho}^{(2)} = [.4, 0, .4, .2]$

- Network 1 has all the groups
- Plant group 3 and insect group 2 are absent from network 2

$\pi\rho$-colBiSBM: partially nested structures

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \end{pmatrix}$$

Plants	Insects
$\pi^{(1)} = [.25, .25, .5]$	$\boldsymbol{\rho}^{(1)} = [.5, .3, 0, .2]$
$\pi^{(2)} = [.25, .25, .5]$	$\boldsymbol{\rho}^{(2)} = [.1, .3, .4, 0]$

- Shared: the 3 plant groups and insect groups 1, 2
- Insect group 4: only in network 1; group 3: only in network 2

Summary of the models

$$Y^{(s)} \sim_{\text{ind}} \text{biSBM}_{n_s,m_s}(K_s, L_s, \pi^{(s)}, \boldsymbol{\rho}^{(s)}, \boldsymbol{\alpha}^{(s)})$$

Model	Proportions	Connections	Nb of parameters
iid-colBiSBM	$\pi, \boldsymbol{\rho}$	$\boldsymbol{\alpha}$	$(K-1)+(L-1)+KL$
π-colBiSBM	$\pi^{(s)}, \boldsymbol{\rho}$	$\boldsymbol{\alpha}$	$S(K-1)+L-1+KL$
ρ-colBiSBM	$\pi, \boldsymbol{\rho}^{(s)}$	$\boldsymbol{\alpha}$	$(K-1)+S(L-1)+KL$
$\pi\rho$-colBiSBM	$\pi^{(s)}, \boldsymbol{\rho}^{(s)}$	$\boldsymbol{\alpha}$	$\leq S(K-1+L-1)+KL$
sep-BiSBM	$\pi^{(s)}, \boldsymbol{\rho}^{(s)}$	$\boldsymbol{\alpha}^{(s)}$	$\sum_s (K_s-1)+(L_s-1)+K_sL_s$

$\rightarrow$ sep-BiSBM = one SBM per network

A note on the connectance

- In iid-coBilSBM, all the networks are assumed to have the same expected connectance:

$$\mathbb{E}[C] = \sum_{k,\ell} \rho_\ell \pi_k \alpha_{k\ell}$$

  $\rightarrow$ If the networks of the collection do not have $\approx$ same C: model not adapted

- $\pi\rho$-colBiSBM relaxes that but not designed for that purpose.

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

Fitting a colBiSBM

For each model (iid, $\pi\rho$):

- Fixed (K, L): variational EM, as for one network
- Choice of (K, L): ICL adapted to each model

R package `colSBM`

```
estimate_colBiSBM(netlist, colsbm_model = "iid")
colsbm_model: "iid", "pi", "rho", "pirho"
```

Simple networks [Chabert-Liddell et al., 2024], bipartite networks [Lacoste et al., 2025]

A modified model selection criterion

- ICL favours clear clusterings (small entropy)
- Here: relax this property

Replace the fit term $\mathbb{E}_{\hat{p}}[\log \ell(Y, Z, W; \hat{\theta})]$ by the ELBO:

$$\mathbb{E}_{\hat{p}}\left[\log \ell(Y, Z, W; \hat{\theta})\right] + \mathcal{H}(\hat{p})$$

$\hat{p}$: variational approximation of $p(Z, W \mid Y)$

Two questions = two model choices

Is there a common structure?

$$\mathrm{ICL}^{\text{colBiSBM}}(\hat{K}, \hat{L}) \quad \text{vs} \quad \sum_{s=1}^{S} \mathrm{ICL}^{\text{SBM}}(Y^{(s)}; \hat{K}_s, \hat{L}_s)$$

Do we need free proportions?

$$\mathrm{ICL}^{\text{iid}}(\hat{K}, \hat{L}) \quad \text{vs} \quad \mathrm{ICL}^{\pi\rho}(\hat{K}, \hat{L})$$

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

Going further: clustering the networks

What if the S networks do not all share the same structure?

Partition of the networks

Find groups $\mathcal{S}_1, \ldots, \mathcal{S}_G$ of $\{1, \ldots, S\}$ such that

$$\forall s \in \mathcal{S}_g, \quad Y^{(s)} \sim \text{biSBM}_{n_s, m_s}(K_g, L_g, \pi^{(s)}, \boldsymbol{\rho}^{(s)}, \boldsymbol{\alpha}^g)$$

- One $\pi\rho$-colBiSBM per group of networks

$\rightarrow$ e.g. groups of land uses with the same structure

Scoring a partition

Score of a partition $\mathcal{G} = (\mathcal{S}_1, \ldots, \mathcal{S}_G)$:

$$\mathrm{Sc}(\mathcal{G}) = \sum_{g=1}^{G} \mathrm{ICL}\left((Y^{(s)})_{s \in \mathcal{S}_g}; \hat{K}_g, \hat{L}_g\right)$$

Best partition:

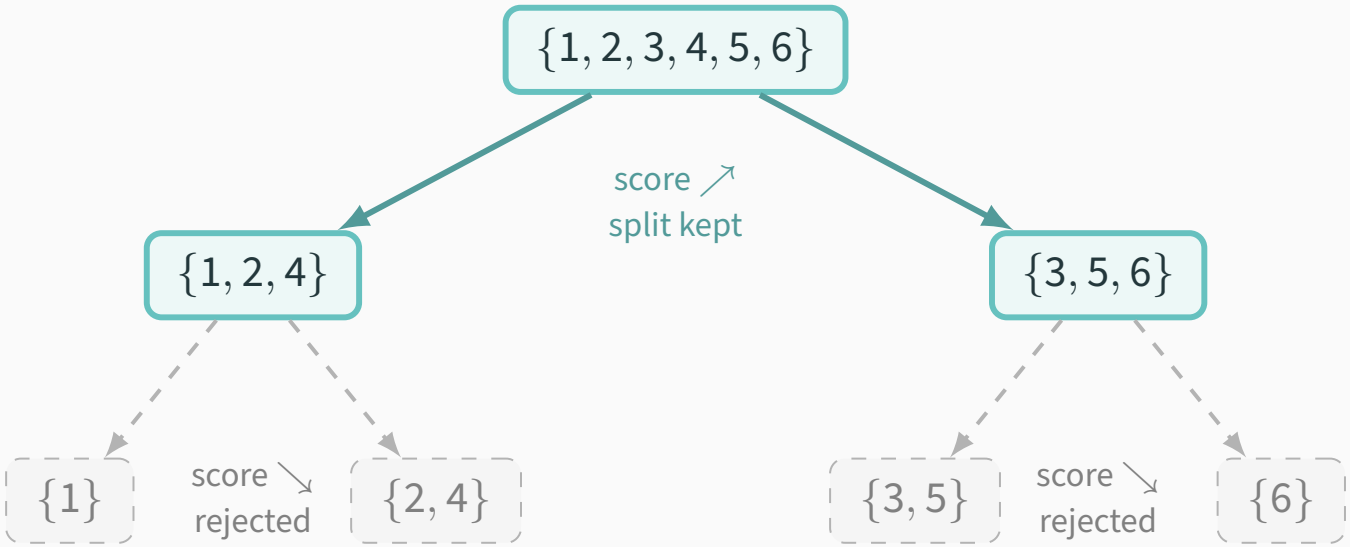
$$\mathcal{G}^\star = \arg\max_{\mathcal{G}} \mathrm{Sc}(\mathcal{G})$$

- $G = 1$: one colBiSBM for all $G = S$: one SBM per network

$\rightarrow$ The same criterion as for the two questions

Finding the best partition

Too many partitions $\rightarrow$ divisive algorithm



- Split a group in 2; keep the split if the score increases
- Package `colSBM`: `clusterize_bipartite_networks()`

[Lacoste et al., 2025], see also [Rebafka, 2024]

Motivation

Introduction to networks

Introduction to stochastic block models

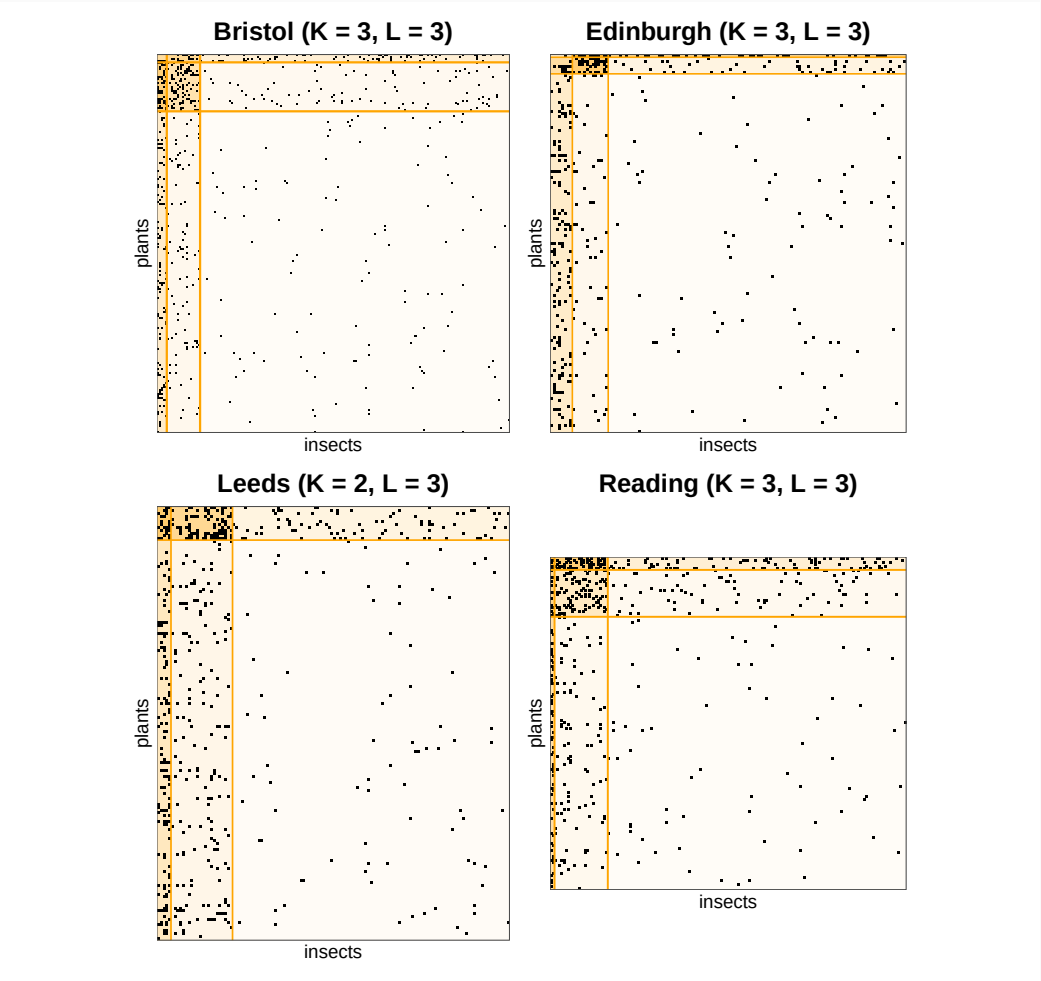
Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

Back to the 4 cities

UK data



→ One SBM per city: the same core–periphery structure?

But the groups do not match

UK data

	Bristol	Edinburgh	Leeds	Reading
$K \times L$	3×3	3×3	2×3	3×3

- Not the same number of groups (Leeds: 2 plant groups)
- Group k of Bristol $\neq$ group k of Leeds: labels are arbitrary
- The $\alpha_{k\ell}$ are estimated city by city

$\rightarrow$ Is it really the same structure?

Joint or separate models?

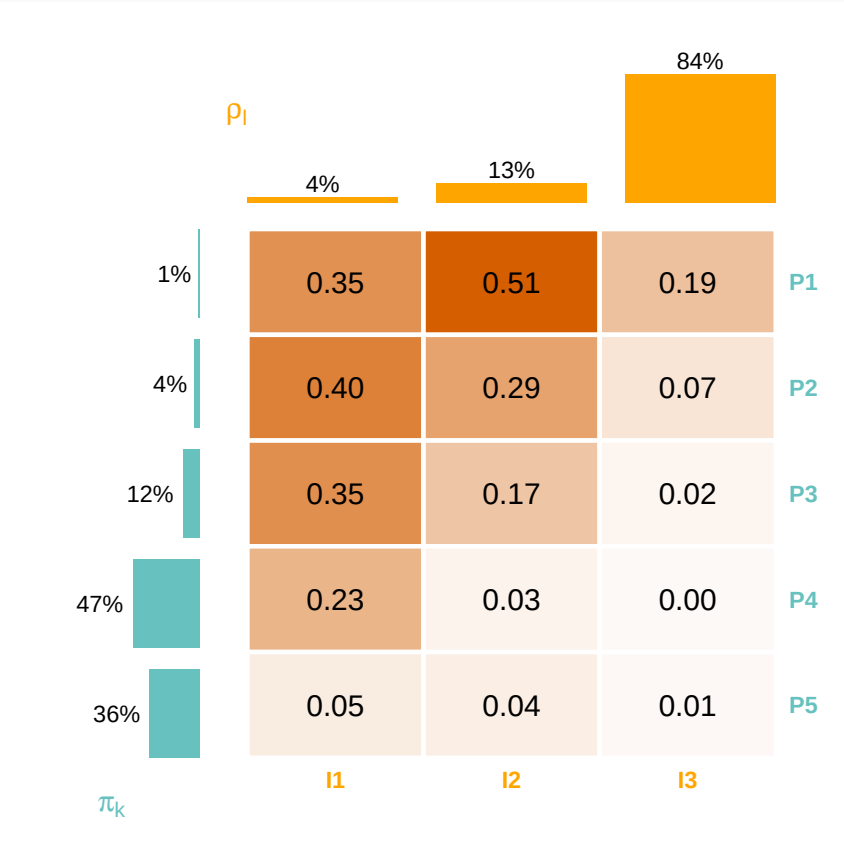
UK data

Model	$K \times L$	ICL
sep-BiSBM (one per city)	2–3 $\times$ 3	-8780
π-colBiSBM	5 $\times$ 3	-8677
ρ-colBiSBM	5 $\times$ 3	-8665
iid-colBiSBM	**5 $\times$ 3**	$\mathbf{-8651}$

- Joint model preferred to separate SBMs
- iid: the same structure, in the same proportions

The common structure

UK data

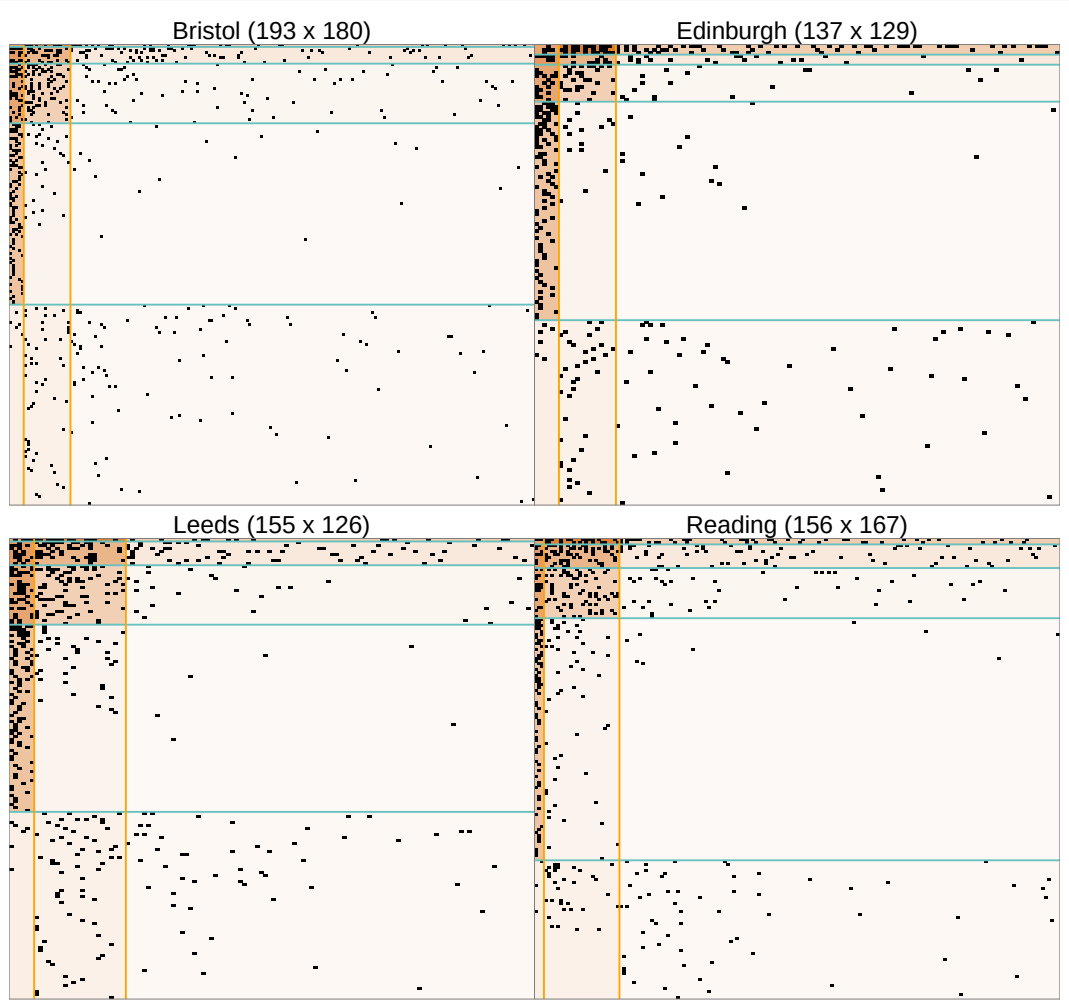


- $\hat{\alpha}_{k\ell}$ shared by the 4 cities

  (bars: proportions $\hat{\pi}_k$, $\hat{\rho}_\ell$)
- Core: P1, P2 $\times$ I1, I2
- Periphery: P4, P5 and I3 (84% of the insects)

The same groups in every city

UK data



Rows / columns ordered by the joint groups. Background: $\hat{\alpha}_{k\ell}$

Joint modelling: more statistical power

UK data

- Separately: 2–3 $\times$ 3 groups per city
- Jointly: **5** $\times$ **3** groups

- P1: 1% of the plants = 1 to 3 species per city

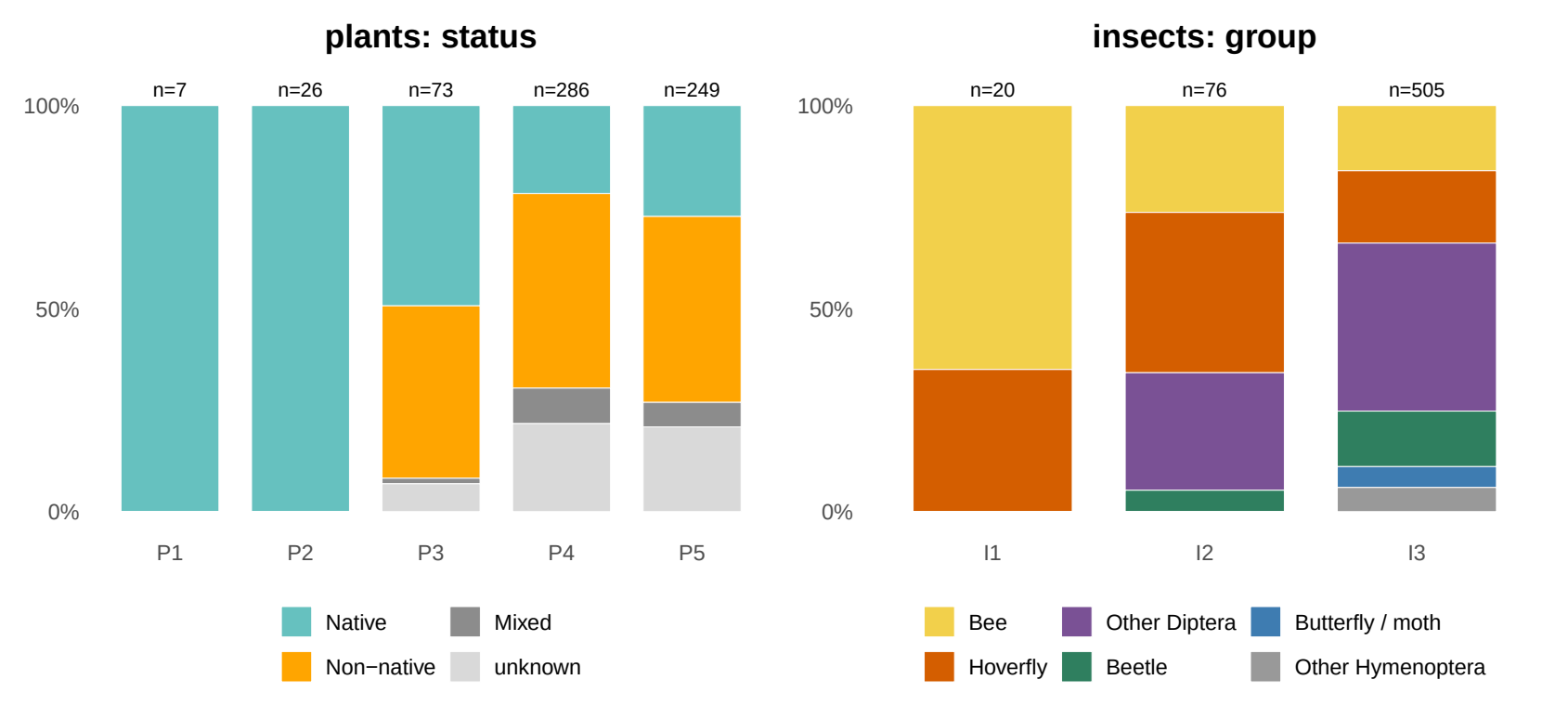
$\rightarrow$ No city alone can isolate it

Borrowing strength

The same $\alpha_{k\ell}$ estimated on the 4 networks $\rightarrow$ a finer structure

The 4 cities: composition of the groups

UK data



Species × city: a species seen in 2 cities counts twice

The same composition as in Bristol?

UK data

Yes

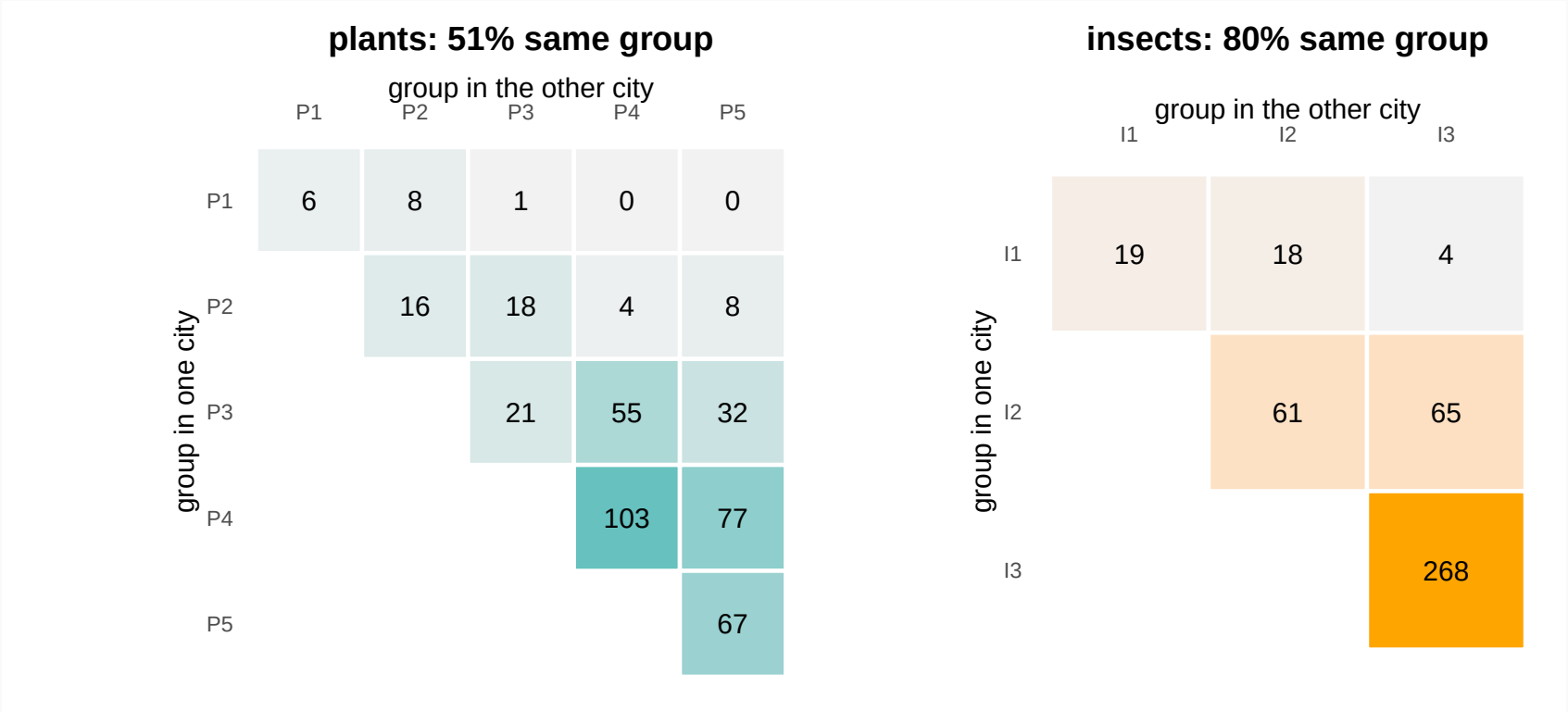
- Core plants P1, P2: 100% native, in the 4 cities
- Periphery P4, P5: 40–60% non-native in every city
- Core insects I1: bees (65%) and hoverflies (35%)

Finer than in Bristol

- Bristol core plants (4) $\rightarrow$ split into P1 (1) and P2 (3)
- Bristol plant periphery (164) $\rightarrow$ split into P4 (76) and P5 (84)
- Insect groups: the same as in Bristol

Does a species keep its group across cities?

UK data



Each species seen in 2 cities, for every pair of cities: its group in each city

→ Insects keep their group more than plants. Moves: mostly to a neighbouring group

The insects that change group

UK data

	Bristol	Edinburgh	Leeds	Reading
Platycheirus albimanus	I2	I1	I1	I2
Helophilus pendulus	I2	I1	I2	I2
Episyrphus balteatus	I2	I1	I2	I2
Bombus lapidarius	I2	I2	I1	I2
Eristalis tenax	I1	I3	I2	I2
Eristalis arbustorum	I1	I3	I2	I2
Paregle audacula	I3	I2	I2	I2
Delia sp	I3	I2	I2	I2
Melanostoma mellinum	I2	I3	I1	I3
Eristalis pertinax	I2	I2	I2	I3
Lucilia sericata	I3		I2	I2
Scathophaga stercoraria	I3	I2	I2	I3
Rhinophora lepida	I3			I2
Reichertella geniculata	I3	I2		
Hydrellia sp	I3	I2	I3	I2
Coenosia sp	I3	I2	I3	I2
Eupeodes luniger	I3	I3	I2	I2
Bombus hortorum	I2	I3	I2	I3
Eupeodes corollae		I3	I2	I3
Oedemera lurida	I3		I3	I2
Lasioglossum morio	I3		I3	I2
Bombus vestalis	I3		I2	I3
Sciaridae sp	I3	I2	I3	I3
Adia cinerella	I3	I3	I2	I3
Syrphus ribesii	I3	I3	I2	I3
Rhagonycha fulva	I3	I2	I3	I3
Osmia bicornis	I3	I3	I3	I2
Lasioglossum calceatum	I2	I3	I3	I3

- 28 of the 140 insects seen in several cities
- Top: hoverflies and *B. lapidarius*, between I1 and I2
- Bottom: flies, solitary bees, beetles, between I2 and I3

→ A species is a little more or less connected depending on the city

Ecological messages

- Highly connected blocks are made of the same species everywhere.
  - → P1 contains only species observed in the 4 cities,
  - → P2 mostly (77 %): Bellis, Taraxacum, Ranunculus repens, Senecio jacobaea, Rubus fruticosus.
  - → Idem for the insect core I1
- Least connected blocks (P4, P5, I3) mostly made of species seen in 1 or 2 cities: each city has its own peripheral species, in the same proportions.

A nested backbone and after?

UK data



Roughly nested

- The 506 rare insects (I3) mostly visit the most generalist plants (P1, P2).
- The generalist insects (I1: Bombus pascuorum, Apis mellifera, B. terrestris…) visit plant blocks P1 to P4.

Departures from nestedness

- P1 (Bellis, Taraxacum, Heracleum: open, accessible flowers) is visited more by I2 (small Bombus) than by I1 in every city.
- The bees I1 visit P2-P3 more than P1.
- Structure that is not recovered when studying nestedness

Take-home message

Modelling the collection, not one network at a time:

- a finer structure: 5×3 groups instead of 2–3 $\times$ 3
- the species of each site matched to common functional groups, defined by α : comparison of role from a network to another

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

- Introduction
- Joint modelling
- Statistical inference
- Clustering a collection of networks
- Application to the city networks
- Application to the land uses

Conclusion

The 9 land-use networks

UK data

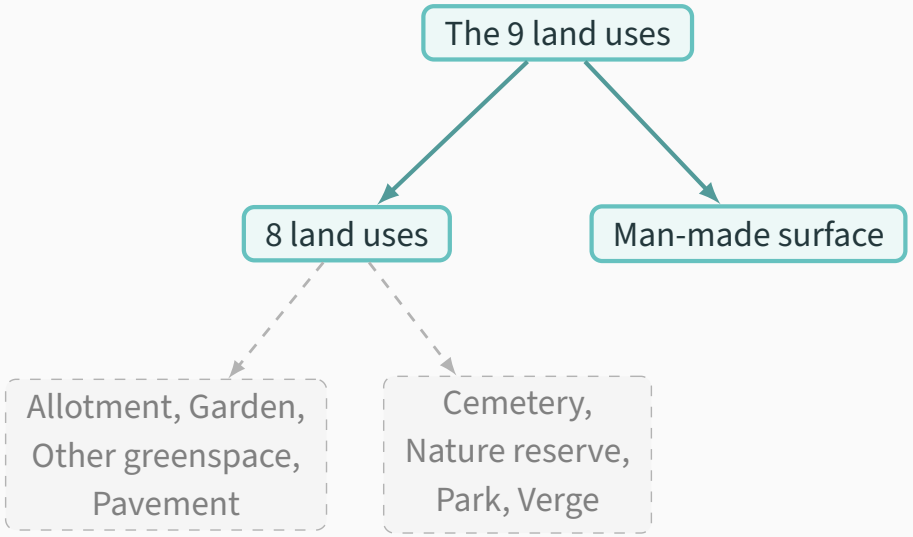
One network per land use, pooled over the 4 cities (binary)

Land use	Visits	Plants	Insects	Connectance
Allotment	1657	180	155	0.024
Garden	1060	237	145	0.019
Cemetery	540	50	120	0.043
Other greenspace	530	25	62	0.072
Nature reserve	467	52	113	0.037
Park	385	42	100	0.051
Verge	207	28	72	0.061
Pavement	47	20	22	0.075
Man-made surface	28	12	15	0.117

→ Do the land uses share the same structure?

Land-use: same structure?

UK data



score = −8739.0

−8659.7 − 67.4 = −8727.1
better: split kept

−5775.2 − 2888.1 = −8663.3
worse than −8659.7: rejected

Caveat

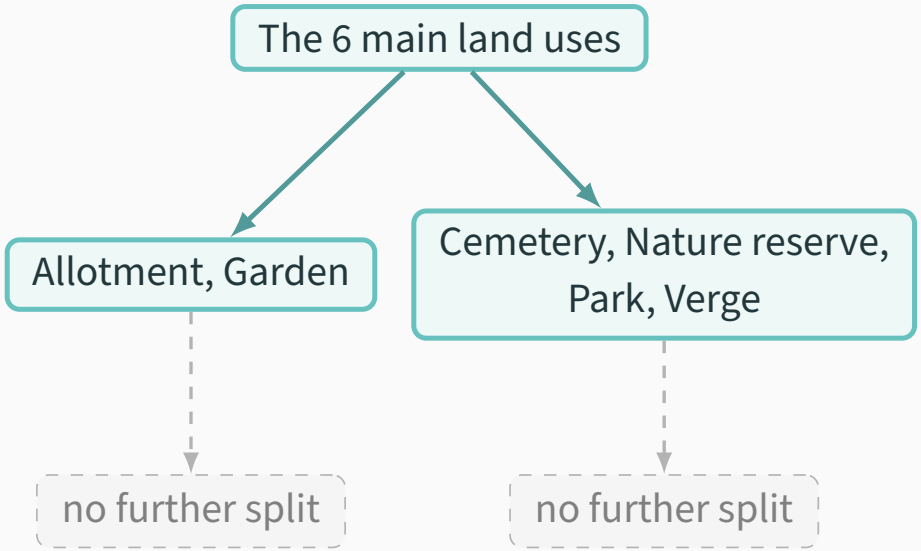
Very different densities (0.019 to 0.117) and tiny networks

→ the partition is hard to find

Without the 3 small land uses

UK data

Other greenspace, Pavement, Man-made surface removed



score = −8194.1

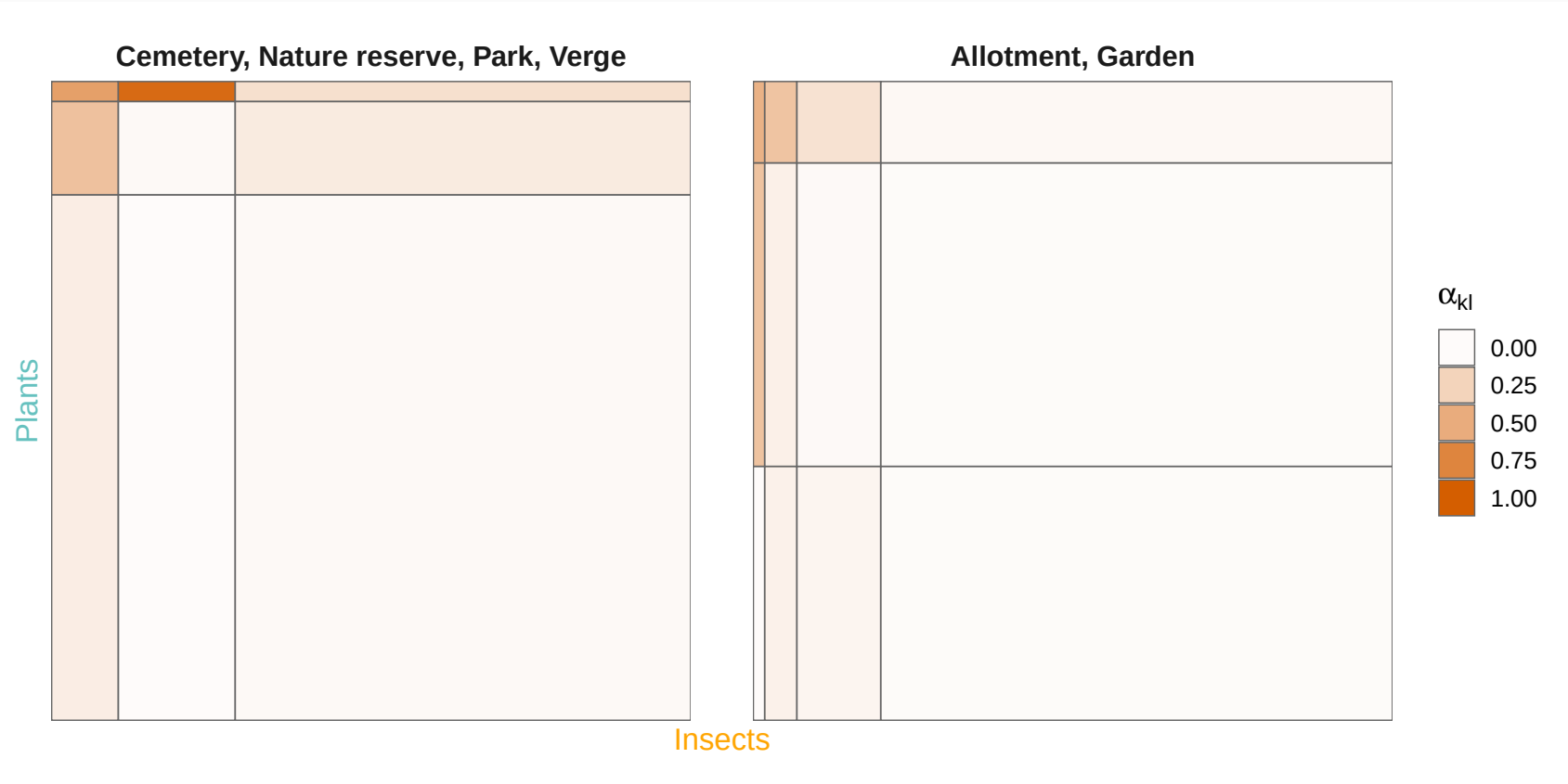
−5291.8 − 2867.3 = −8159.1

better: split kept

→ iid-colBiSBM: gardened vs greenspaces

Two different structures: graphon view

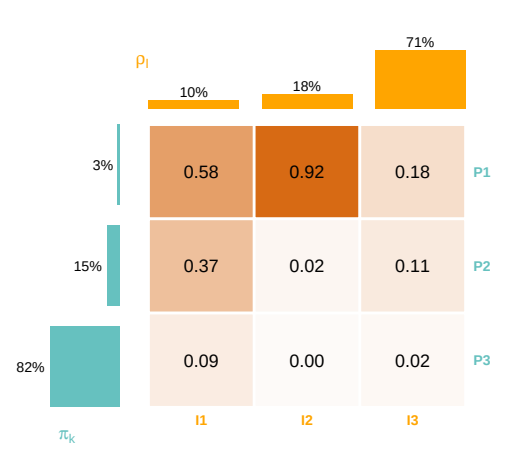
UK data



→ Block sizes = $(\hat{\pi}_k, \hat{\rho}_\ell)$, shade = $\hat{\alpha}_{k\ell}$

Greenspaces: a core-periphery

UK data



- Cemetery, Nature reserve, Park, Verge
- A core-periphery of native wild flower

- Plants are ranked by attractiveness.
  - 3 hubs: *Bellis*, *Taraxacum*, *Heracleum*
  - 13 native wild flowers, mostly of grassland and lawns: *Scorzoneroides*, *Senecio jacobaea*, *Ranunculus*, *Hypochaeris*...All native.
  - A periphery holding all the non-native plants
- Insects
  - A core of generalist insects (I1: *Episyrphus*, *Apis*, *common Bombus*): best-connected insects, with 13 plant partners on average.
  - A block of smaller insects (I2) visit only P1: 62% of *other Diptera*.

Gardens: bee flowers vs fly flowers

UK data



Insects

- I1 (the dominant generalist bees) and I4 (the rare insects): clear-cut.
- I2 and I3 overlap: many species are in I2 in one land use and in I3 in the other.

Plants

- P1: 21 native and 25 non-native: wild and weedy plants (*Taraxacum, Ranunculus repens...* + ornamentals and crops (*Lavandula, Geranium, thyme...*
- P2 and P3 do not share partners I1
  - P2: flowers visited mainly by bees (I1)
  - P3: Mostly open flowers (*Leucanthemum*, *Aster*), never by I1. Visited mainly by flies

Take-home message: land uses

- Clustering may be fragile with very different densities and tiny networks $\rightarrow$ Check by hand
- Main land uses: two structures, gardened vs greenspaces
- To compare networks of different densities $\rightarrow$ model the density [Chabert-Liddell et al., 2024] but not yet for bipartite networks.

Motivation

Introduction to networks

Introduction to stochastic block models

Modelling collections of networks

Conclusion

Take-home message: the SBM

A flexible model for the structure of a network:

- no structure chosen in advance: modules, nestedness, core–periphery… are particular cases
- the data choose the number of groups (ICL)
- groups = species with the same role, read with their traits

Take-home message: comparing networks

Compare structures with a model, not with descriptive statistics:

- same structure or not? $\rightarrow$ a model choice (BICL)
- which networks go together? $\rightarrow$ clustering
- species of different sites matched by their role, even without shared species

Open question: what about abundances?

Abundant species meet more often → more visits, more links, even in a binary network

Remove the abundances

Interactions beyond chance encounters

→ abundances as covariates on the pairs

→ blocks = preferences

Keep the abundances

Structure as a result of heterogeneous abundances

→ abundances as covariates on the species

→ which blocks are explained by abundance?

→ Two different ecological questions: choose before modelling

References i

- Almeida-Neto, M., Guimarães, P., Guimarães Jr, P. R., Loyola, R. D., and Ulrich, W. (2008). **A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement.** *Oikos*, 117(8):1227–1239.
- Baldock, K. C. R., Goddard, M. A., Hicks, D. M., Kunin, W. E., Mitschunas, N., Morse, H., Osgathorpe, L. M., Potts, S. G., Robertson, K. M., Scott, A. V., Staniczenko, P. P. A., Stone, G. N., Vaughan, I. P., and Memmott, J. (2019). **A systems approach reveals urban pollinator hotspots and conservation opportunities.** *Nature Ecology & Evolution*, 3(3):363–373.
- Bar-Hen, A., Barbillon, P., and Donnet, S. (2022). **Block models for generalized multipartite networks: Applications in ecology and ethnobiology.** *Statistical Modelling*, 22(4):273–296.
- Bascompte, J., Jordano, P., Melián, C. J., and Olesen, J. M. (2003). **The nested assembly of plant–animal mutualistic networks.** *Proceedings of the National Academy of Sciences*, 100(16):9383–9387.
- Biernacki, C., Celeux, G., and Govaert, G. (2000). **Assessing a mixture model for clustering with the integrated completed likelihood.** *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.
- Blüthgen, N., Menzel, F., and Blüthgen, N. (2006). **Measuring specialization in species interaction networks.** *BMC Ecology*, 6:9.
- Chabert-Liddell, S.-C., Barbillon, P., and Donnet, S. (2024). **Learning common structures in a collection of networks. An application to food webs.** *The Annals of Applied Statistics*, 18(2):1213–1235.
- Daudin, J.-J., Picard, F., and Robin, S. (2008). **A mixture model for random graphs.** *Statistics and computing*, 18(2):173–183.

References ii

- Dáttilo, W., Lara-Rodríguez, N., Jordano, P., Guimarães, Paulo R., J., Thompson, J. N., Marquis, R. J., Medeiros, L. P., Ortiz-Pulido, R., Marcos-García, M. A., and Rico-Gray, V. (2016). **Unravelling Darwin's entangled bank: architecture and robustness of mutualistic networks with multiple interaction types.** *Proceedings of the Royal Society B: Biological Sciences*, 283(1843):20161564.
- Erdős, P. and Rényi, A. (1959). **On random graphs I.** *Publicationes Mathematicae Debrecen*, 6:290–297.
- Freeman, L. C. (1977). **A set of measures of centrality based on betweenness.** *Sociometry*, 40(1):35–41.
- Govaert, G. and Nadif, M. (2008). **Block clustering with Bernoulli mixture models: Comparison of different approaches.** *Computational Statistics and Data Analysis*, 52(6):3233–3245.
- Kéfi, S., Miele, V., Wieters, E., Navarrete, S., and Berlow, E. (2016). **How Structured Is the Entangled Bank? The Surprisingly Simple Organization of Multiplex Ecological Networks Leads to Increased Persistence and Resilience.** *PLoS Biology*, 14(8):e1002527.
- Klein, A.-M., Vaissière, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., and Tscharntke, T. (2007). **Importance of pollinators in changing landscapes for world crops.** *Proceedings of the Royal Society B: Biological Sciences*, 274(1608):303–313.
- Lacoste, L., Barbillon, P., and Donnet, S. (2025). **Common structure discovery in collections of bipartite networks: Application to pollination systems.** arXiv:2512.01716.
- Lee, C. and Wilkinson, D. J. (2019). **A review of stochastic block models and extensions for graph clustering.** *Applied Network Science*, 4:122.
- Leger, J.-B. (2016). **Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates.** *arXiv preprint arXiv:1602.07587*.
- Mariadassou, M., Robin, S., and Vacher, C. (2010). **Uncovering latent structure in valued graphs: a variational approach.** *The Annals of Applied Statistics*, 4(2):715–742.

References iii

- Matias, C. and Robin, S. (2014). **Modeling heterogeneity in random graphs through latent space models: a selective review.** *ESAIM: Proc.*, 47:55–74.
- Memmott, J., Waser, N. M., and Price, M. V. (2004). **Tolerance of pollination networks to species extinctions.** *Proceedings of the Royal Society B: Biological Sciences*, 271(1557):2605–2611.
- Nowicki, K. and Snijders, T. A. B. (2001). **Estimation and prediction for stochastic blockstructures.** *Journal of the American Statistical Association*, 96(455):1077–1087.
- Ollerton, J., Winfree, R., and Tarrant, S. (2011). **How many flowering plants are pollinated by animals?** *Oikos*, 120(3):321–326.
- Pilosof, S., Porter, M. A., Pascual, M., and Kéfi, S. (2017). **The multilayer nature of ecological networks.** *Nature Ecology & Evolution*, 1(4).
- Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., and Kunin, W. E. (2010). **Global pollinator declines: trends, impacts and drivers.** *Trends in Ecology & Evolution*, 25(6):345–353.
- Rebafka, T. (2024). **Model-based clustering of multiple networks with a hierarchical algorithm.** *Statistics and Computing*, 34(1).
- Schwarz, G. (1978). **Estimating the dimension of a model.** *The Annals of Statistics*, 6(2):461–464.

Appendix: proof of the central identity

Bayes: $\log \ell(Y;\boldsymbol{\theta}) = \log \ell_c(Y, \mathbf{Z}, \mathbf{W};\boldsymbol{\theta}) - \log p(\mathbf{Z}, \mathbf{W} \mid Y;\boldsymbol{\theta})$

Integrate against $\mathcal{R}$ (the left-hand side does not depend on $(\mathbf{Z}, \mathbf{W})$):

$$\log \ell(Y;\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{R}}[\log \ell_c(Y, \mathbf{Z}, \mathbf{W};\boldsymbol{\theta})] - \mathbb{E}_{\mathcal{R}}[\log p(\mathbf{Z}, \mathbf{W} \mid Y;\boldsymbol{\theta})]$$

Add and subtract $\mathbb{E}_{\mathcal{R}}[\log \mathcal{R}(\mathbf{Z}, \mathbf{W})]$:

$$\log \ell(Y;\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\mathcal{R}}[\log \ell_c(Y, \mathbf{Z}, \mathbf{W};\boldsymbol{\theta})] + \mathcal{H}(\mathcal{R})}_{\mathcal{I}(\boldsymbol{\theta}, \mathcal{R})} + \underbrace{\mathbb{E}_{\mathcal{R}}\left[\log \frac{\mathcal{R}(\mathbf{Z}, \mathbf{W})}{p(\mathbf{Z}, \mathbf{W} \mid Y;\boldsymbol{\theta})}\right]}_{\mathrm{KL}[\mathcal{R} \,\|\, p(\cdot \mid Y;\boldsymbol{\theta})] \geq 0}$$

◂ Back

Appendix: VE-step for the bipartite SBM

Under $\mathcal{R}_{\tau,\boldsymbol{\eta}}$, $\mathbb{E}[\mathbf{1}_{Z_i=k}\mathbf{1}_{W_j=\ell}] = \tau_{ik}\eta_{j\ell}$, so with $\log f_{k\ell}(Y_{ij}) = Y_{ij}\log\alpha_{k\ell} + (1-Y_{ij})\log(1-\alpha_{k\ell})$:

$$\mathcal{I}(\boldsymbol{\theta}, \mathcal{R}_{\tau,\boldsymbol{\eta}}) = \sum_{i,j}\sum_{k,\ell}\tau_{ik}\eta_{j\ell}\log f_{k\ell}(Y_{ij}) + \sum_{i,k}\tau_{ik}\log\pi_k + \sum_{j,\ell}\eta_{j\ell}\log\rho_\ell$$
$$- \sum_{i,k}\tau_{ik}\log\tau_{ik} - \sum_{j,\ell}\eta_{j\ell}\log\eta_{j\ell}$$

Maximise under $\sum_k \tau_{ik} = 1$, $\sum_\ell \eta_{j\ell} = 1$ (Lagrange multipliers) $\rightarrow$ fixed point:

$$\tau_{ik} \propto \pi_k \prod_j \prod_\ell f_{k\ell}(Y_{ij})^{\eta_{j\ell}}, \qquad \eta_{j\ell} \propto \rho_\ell \prod_i \prod_k f_{k\ell}(Y_{ij})^{\tau_{ik}}$$

◂ Back

Appendix: M-step for the bipartite SBM

Maximise $\mathbb{E}_{\mathcal{R}_{\tau,\eta}}[\log \ell_c(Y, \mathbf{Z}, \mathbf{W}; \theta)]$ under $\sum_k \pi_k = 1$, $\sum_\ell \rho_\ell = 1$: explicit

$$\hat{\pi}_k = \frac{1}{n} \sum_i \tau_{ik}, \qquad \hat{\rho}_\ell = \frac{1}{m} \sum_j \eta_{j\ell}, \qquad \hat{\alpha}_{k\ell} = \frac{\sum_{i,j} \tau_{ik} \eta_{j\ell} Y_{ij}}{\sum_{i,j} \tau_{ik} \eta_{j\ell}}$$

- $\hat{\alpha}_{k\ell}$: proportion of realised interactions between groups k and ℓ, weighted by the memberships
- Poisson model: same formula, $\hat{\lambda}_{k\ell}$ = mean number of visits
- With covariates: no closed form $\rightarrow$ numerical optimisation (Newton–Raphson)

◂ Back