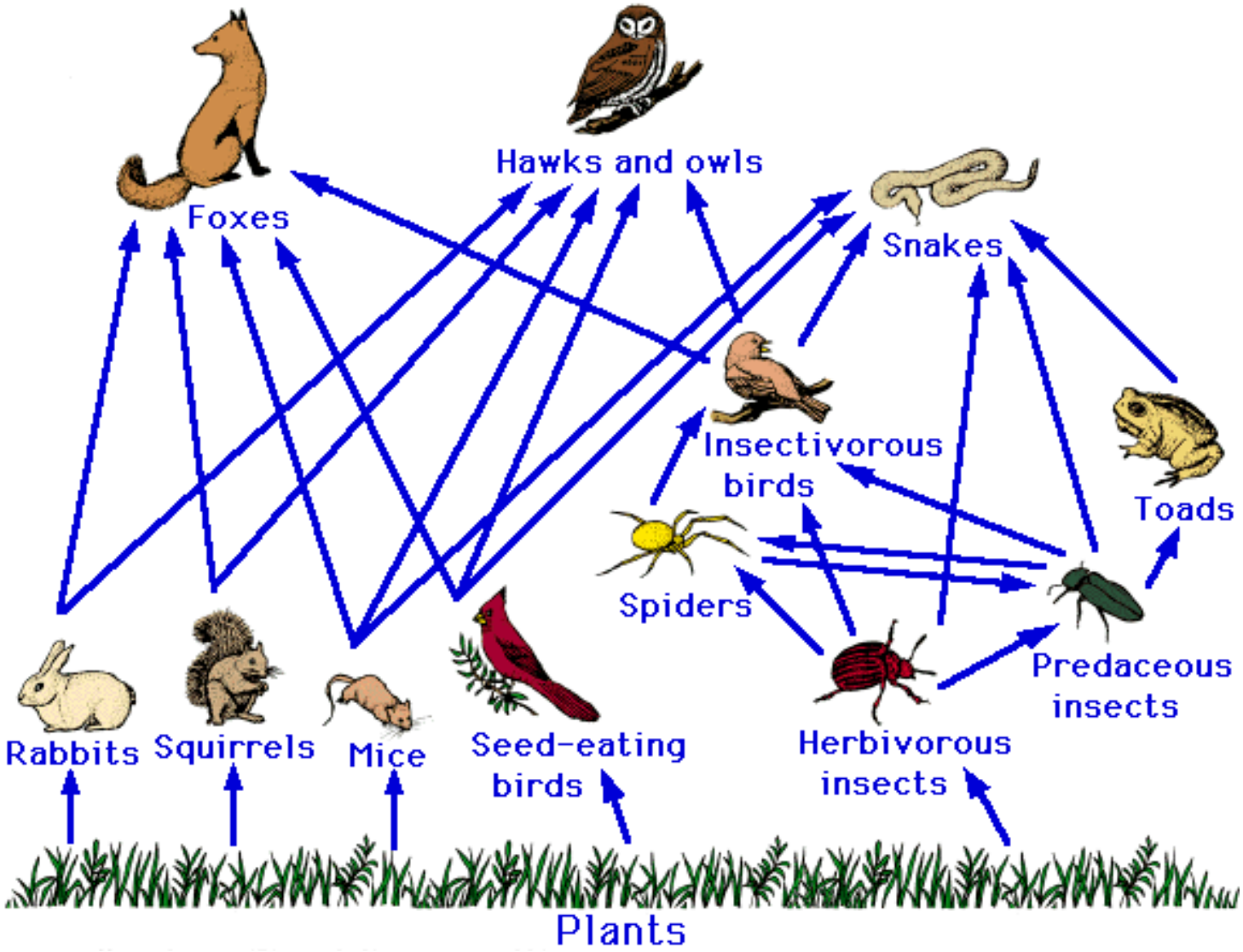
Graph clustering of ecological networks

Tabea Rebafka

LPSM, Sorbonne Université

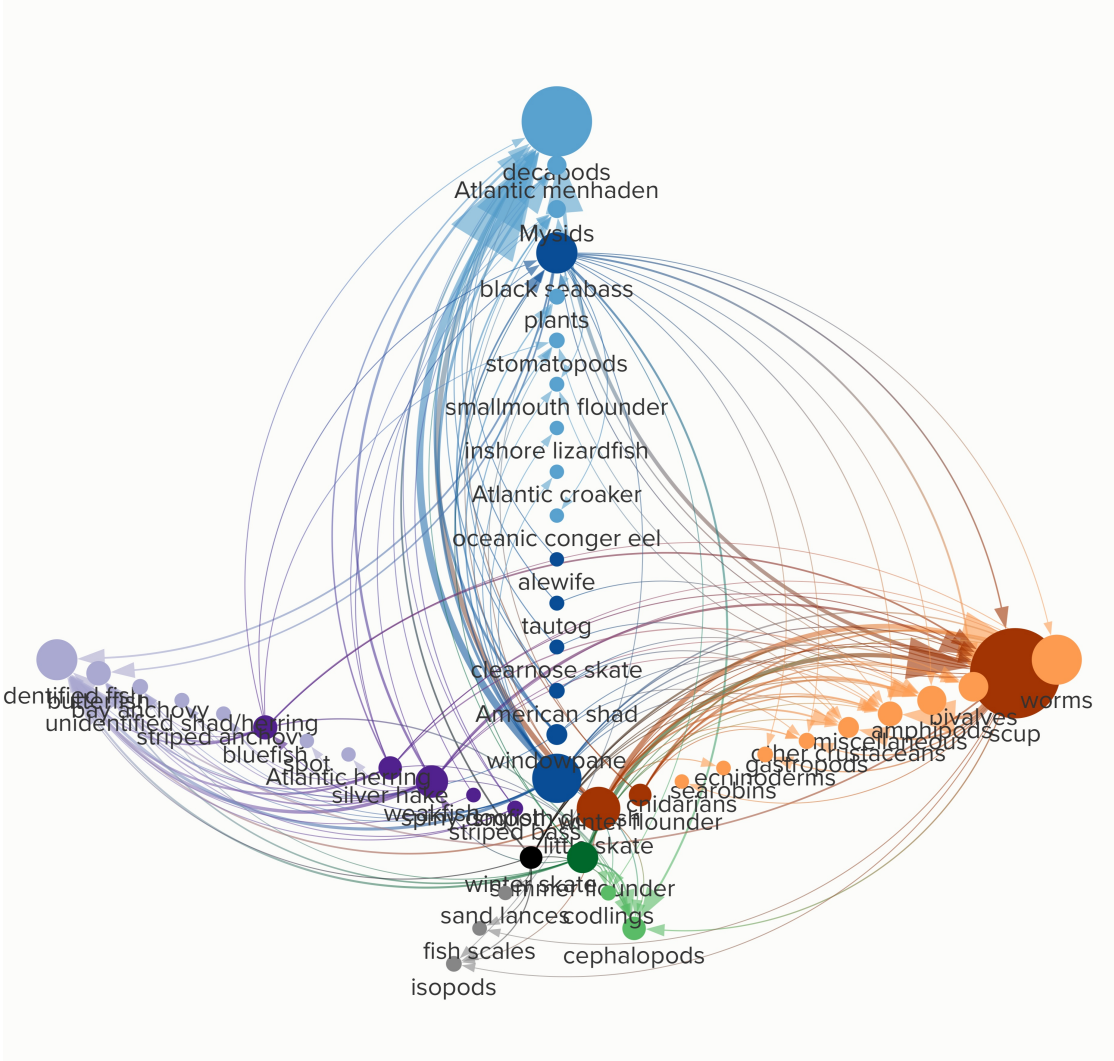
Rencontre MMB
13 octobre 2022

Foodwebs I



Source: www.pw.live/chapter-our-environment/food-web

Foodwebs II



Fall 2021 NEAMAP Food Webs – Mid-Atlantic Coastal Food Web

Source: Virginia Institute of Marine Science

Foodwebs III

Mangal database

- collection of foodwebs on a planetary scale
- 1.300 networks, 120.000 interactions across 7.000 taxa
- contains further information on
  - type (predation, mutualism, parasitism)
  - geographic location
  - climate conditions
- Poisot et al. (2016) + R package `rmangal`

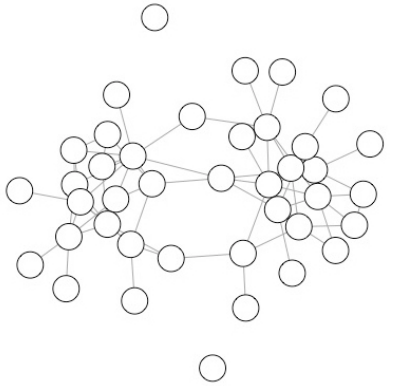
Foodwebs IV

Questions

- analyze and compare many networks
- identify networks with similar structure/organization
- detect outliers/changes...

$\hookrightarrow$ do it in an **automatic** way
$\hookrightarrow$ use an **objective criterion** for comparison

Graph clustering task



From a **mathematical point of view**.

- consider a **collection of networks** or graphs
- goal: graph **clustering**
- many other fields of application : social sciences, transport, biology and medicine (e.g. metabolic networks)

Clustering in machine learning literature

Classical clustering approaches

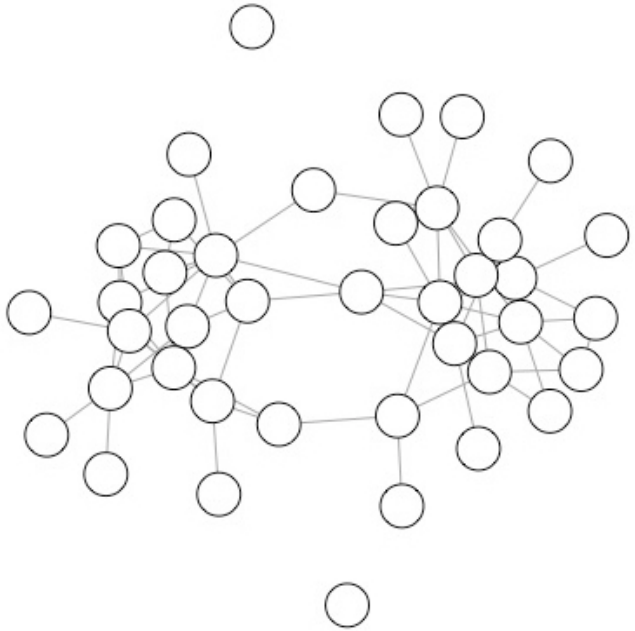
- clustering of **vectors**: kmeans, DBSCAN, GMM
- straightforward solution for networks:
  - compute a **graph embedding** (based on hand-designed statistics or deep learning)
  - apply classical ML clustering method

New network clustering approach

Our approach

- is a statistical one
- introduce a **statistical model** for each graph
- perform **model-based clustering** (like a classical mixture model)
- hierarchical **agglomerative cluster algorithm**
- interpretable output

Modelling I

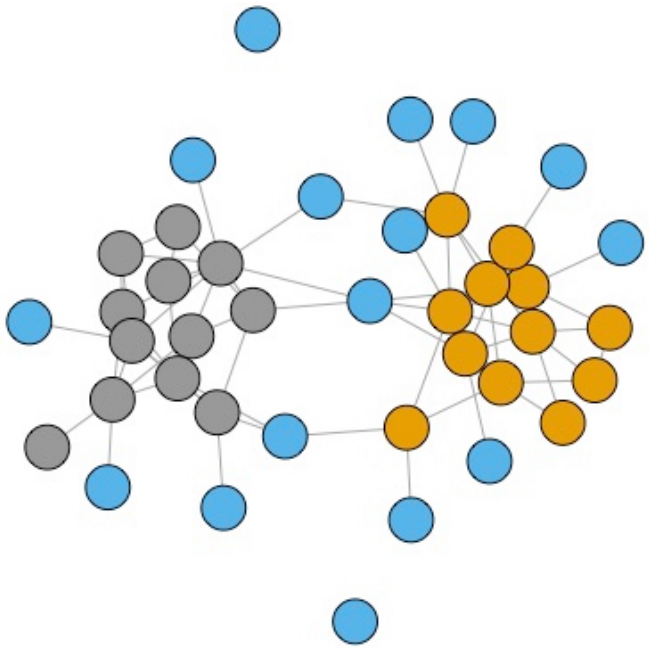


$\Longleftrightarrow$

```
> adjMatrix
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
 [1,]    0    0    0    0    0    0    0    0    0     0     1     0     1
 [2,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [3,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [4,]    0    0    0    0    0    0    0    0    1     0     0     0     0
 [5,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [6,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [7,]    0    0    0    0    0    0    0    0    0     0     0     0     0
 [8,]    0    0    0    0    0    0    0    0    0     1     0     0     0
 [9,]    0    0    0    1    0    0    0    0    0     0     0     0     0
[10,]    0    0    0    0    0    0    0    1    0     0     1     0     0
[11,]    1    0    0    0    0    0    0    0    0     1     0     0     0
[12,]    0    0    0    0    0    0    0    0    0     0     0     0     0
[13,]    1    0    0    0    0    0    0    0    0     0     0     0     0
```

Adjacency matrix

Modelling II



Stochastic block model (SBM)

- **Block memberships** For node i, $Z_i \in \{1, \ldots, K\}$ is drawn independently with probabilities

$$\mathbb{P}(Z_i = \bullet) = \pi_\bullet$$

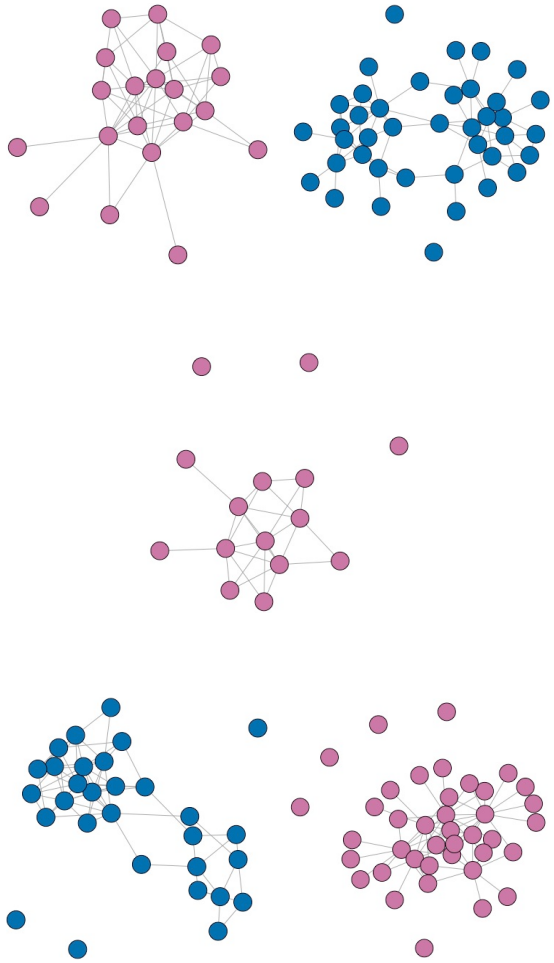
- **Edges** Conditionally on $Z_1, \ldots, Z_n$, $A_{i,j}$ are drawn independently

$$A_{i,j} | (Z_i = \bullet, Z_j = \bullet) \sim \text{Bernoulli}(\gamma_{\bullet\bullet})$$

- **Model parameters** of the SBM:

$$\theta^{\text{SBM}} = ((\pi_k)_{1 \leq k \leq K}, (\gamma_{k,l})_{1 \leq k,l \leq K})$$

Modelling III



Mixture model of SBMs

- C different SBM parameters $\theta_c^{\mathrm{SBM}}, c = 1, \ldots, C$
- **Cluster membership** For network m, $U_m \in \{1, \ldots, C\}$ is drawn independently with probabilities

$$\mathbb{P}(U_m = c) = p_c$$

- Conditionally on $U_1, \ldots, U_M$, the adjacency matrix $A^{(m)}$ is drawn from a SBM:

$$A^{(m)} | (U_m = c) \sim \mathrm{SBM}(\theta_c^{\mathrm{SBM}})$$

Estimation I

- Observed collection of networks: $\mathcal{A} = \{A^{(m)}, m = 1, \ldots, M\}$
- Latent variables: $\mathcal{U} = (U_1, \ldots, U_M)$, $\mathcal{Z} = \{Z^{(m)}, m = 1, \ldots, M\}$
- Model parameters of the mixture:
  $\theta = \left((p_1, \ldots, p_C), (\boldsymbol{\pi}^{(c)}, \boldsymbol{\gamma}^{(c)})_{c=1,\ldots,C}\right)$

Estimation II

Integrated classification likelihood (ICL)

- Bayesian approach: put a prior $p(\theta)$ on the parameters θ
- Likelihood criterion:

$$\mathrm{ICL}(\mathcal{U}, \mathcal{Z}; \mathcal{A}) = \log(p(\mathcal{U}, \mathcal{Z}, \mathcal{A})) = \log\left(\int p(\mathcal{U}, \mathcal{Z}, \mathcal{A}|\theta)p(\theta)\mathrm{d}\theta\right)$$

- Estimates of $\mathcal{U}$ and $\mathcal{Z}$:

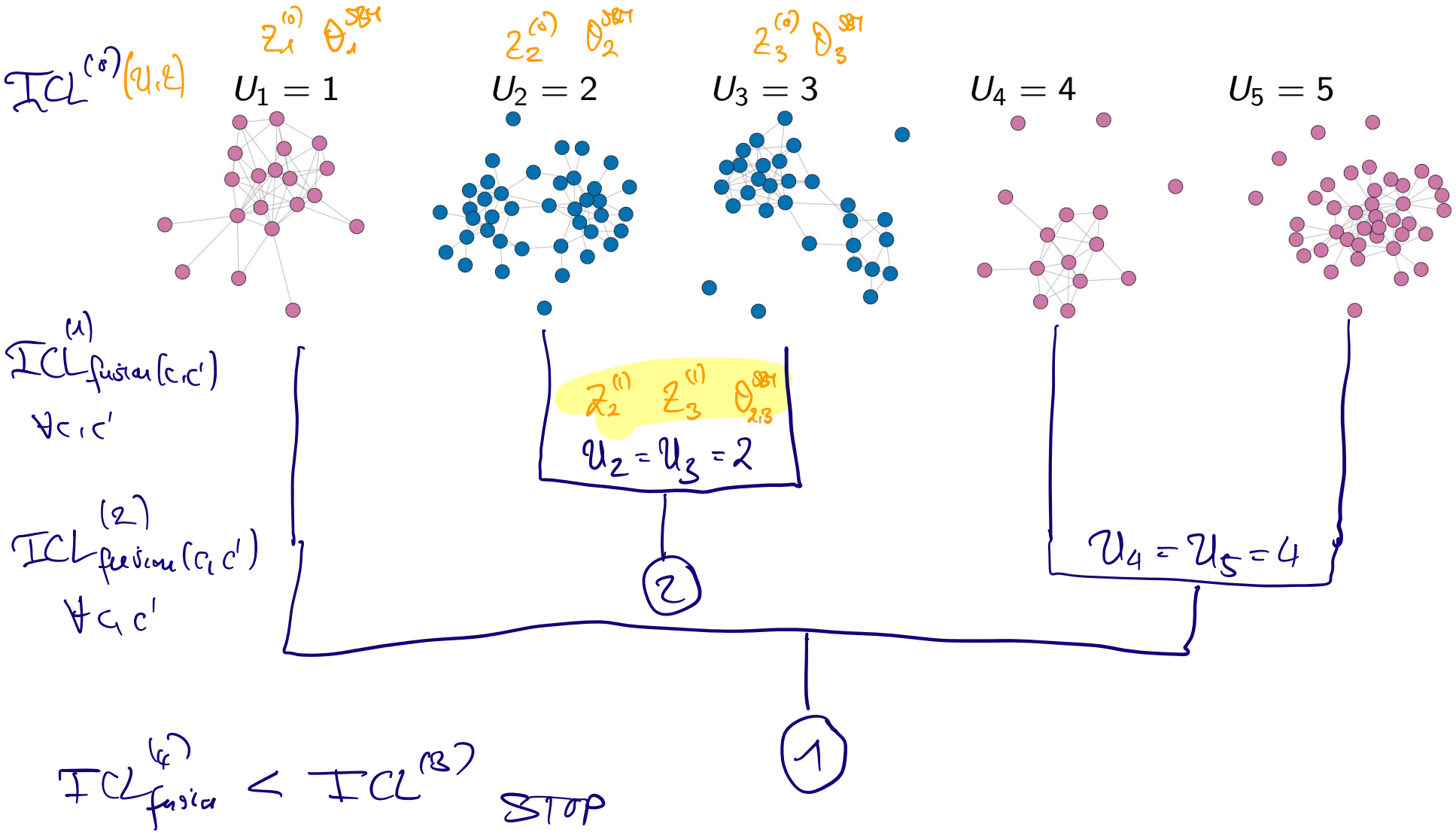
$$(\hat{\mathcal{U}}, \hat{\mathcal{Z}}) = \arg\max_{\mathcal{U},\mathcal{Z}} \mathrm{ICL}(\mathcal{U}, \mathcal{Z}; \mathcal{A})$$

(Côme and Latouche, 2015)

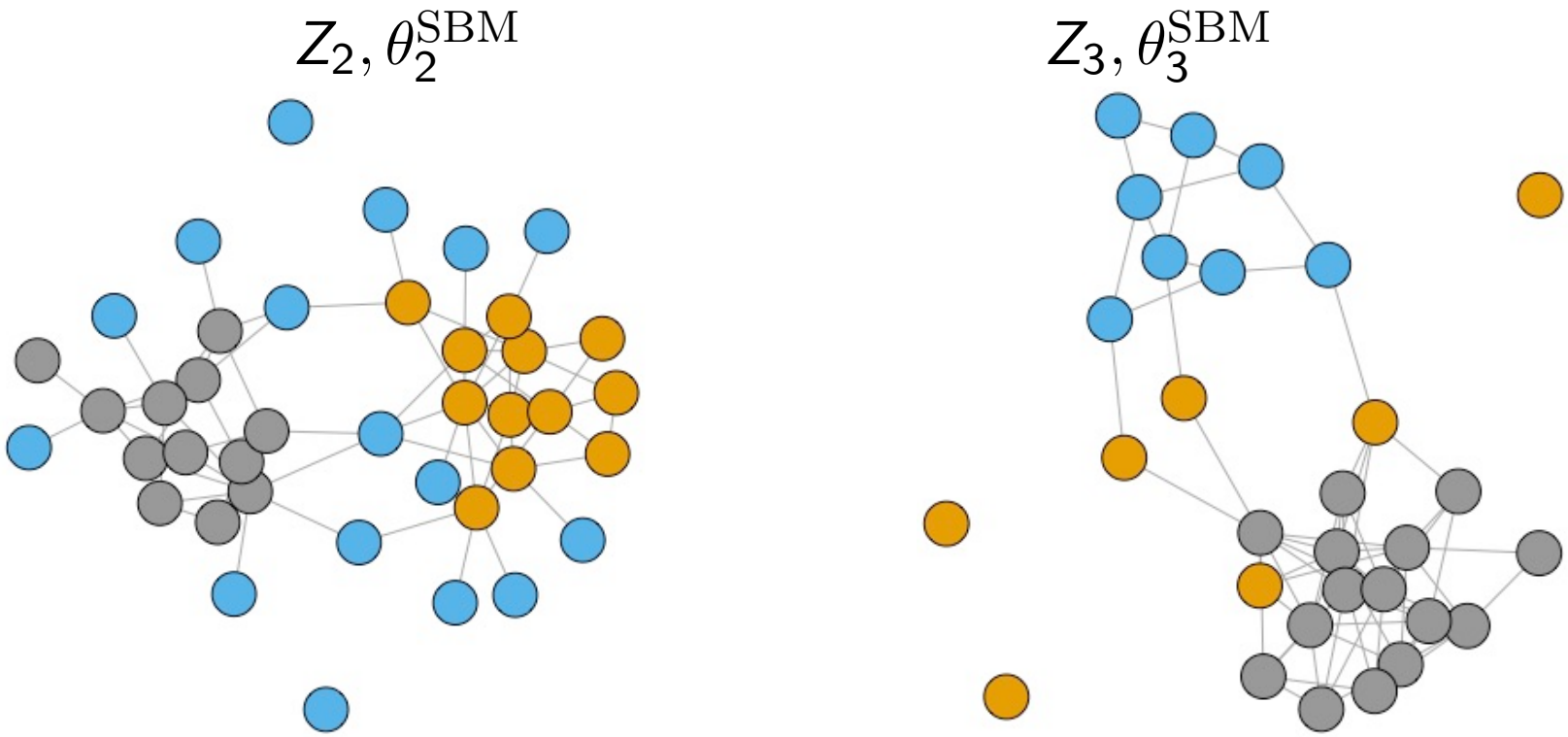
- With an appropriate prior $p(\theta)$, the ICL has closed-form expression
- Discrete optimization

Estimation III

Agglomerative algorithm



Fusion of two clusters I

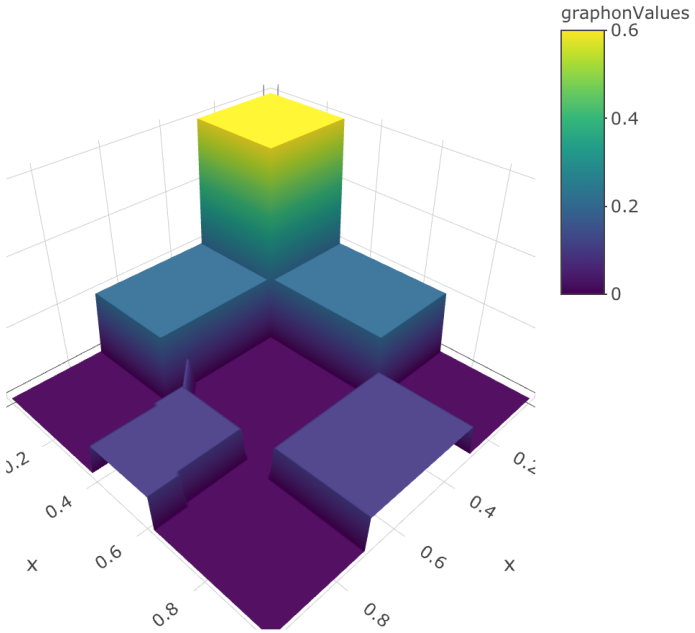


Label-switching problem in the SBM

Fusion of two clusters II

$$\boldsymbol{\pi} = (0.3, 0.4, 0.3)$$

$$\gamma = \begin{pmatrix} 0.6 & 0.2 & 0 \\ 0.2 & 0 & 0.1 \\ 0 & 0.1 & 0 \end{pmatrix}$$



Graphon of a SBM

- $g(u, v) = \gamma_{k,l}$ for all $u \in I_k$, $v \in I_l$
- intervals $I_k = \left(\sum_{s=1}^{k-1} \pi_s, \sum_{s=1}^{k} \pi_s\right]$
- piecewise constant function
- Graphon g depends on the block labelling!

(Lovász, 2012)

Fusion of two clusters III

- Consider L^2-distance of two graphons:

$$\|g_{\theta_1} - g_{\theta_2}\|_2 = \left(\int_{[0,1]^2} (g_{\theta_1}(u,v) - g_{\theta_2}(u,v))^2 \mathrm{d}(u,v) \right)^{\frac{1}{2}}$$

Matching blocks of two SBMs

Search the best permutations of block labels by

$$(\hat{\sigma}_1, \hat{\sigma}_2) \in \arg \min_{\sigma_1 \in \mathfrak{S}_{K_1}, \sigma_2 \in \mathfrak{S}_{K_2}} \|g_{\sigma_1(\theta_1)} - g_{\sigma_2(\theta_2)}\|_2,$$

where $\mathfrak{S}_K$ denotes the set of all permutations of $\{1, \ldots, K\}$ and $\sigma(\theta) = \left((\pi_{\sigma(1)}, \ldots, \pi_{\sigma(K)}), (\gamma_{\sigma(k),\sigma(l)})_{k,l} \right)$.

- does not depend on the clusterings, the data, the number of networks or the number nodes
- works also for $K_1 \neq K_2$

Mangal Foodwebs I

Data

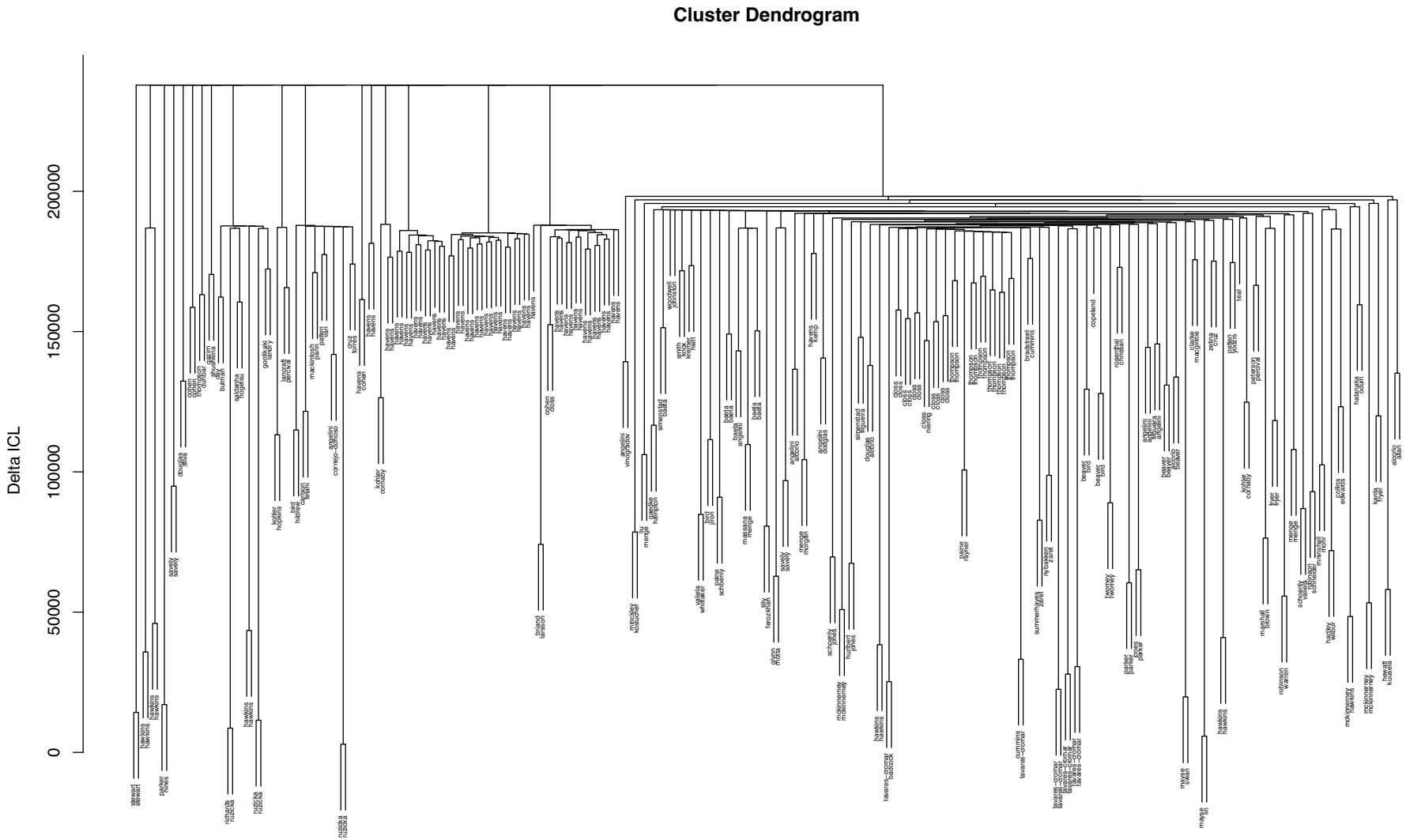
- 270 predator foodwebs
- mean number of species: 33 (min=5, max=317)
- mean number of edges: 120 (min=5, max=1.086)
- mean density: 0,12 (min=0,01, max=0,41)

Clustering results

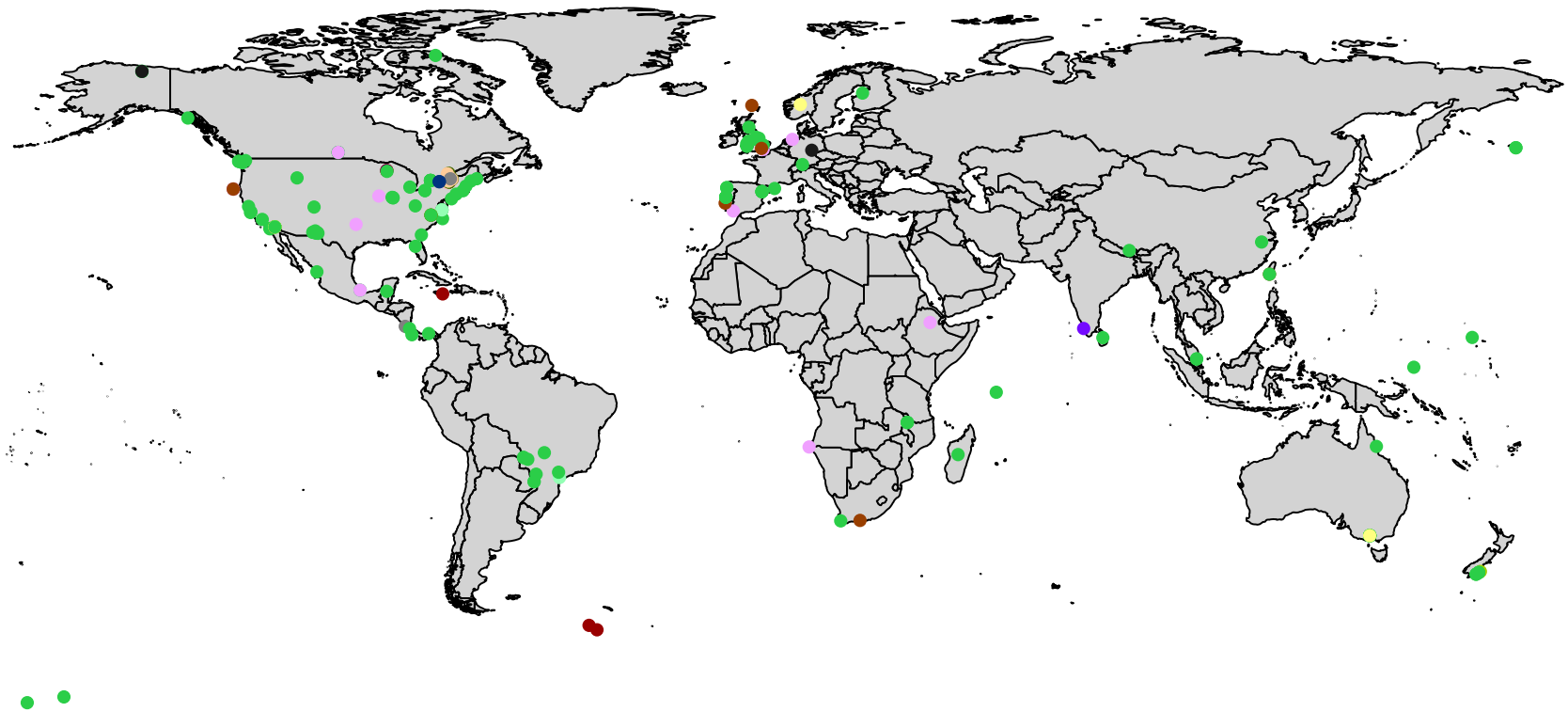
- partition into 18 clusters
- cluster sizes:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
166	19	18	11	4	2	2	14	4	12	2	2	2	2	2	2	4	2

Mangal Foodwebs II

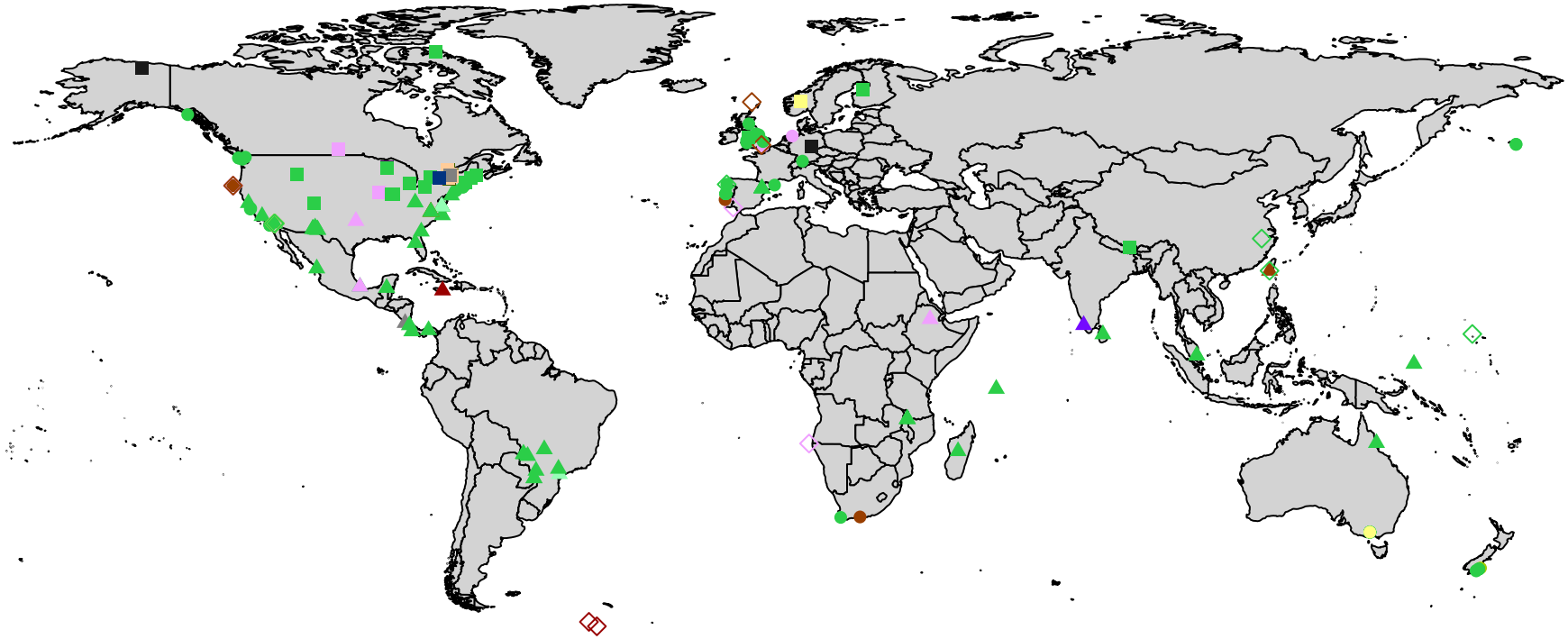


Mangal Foodwebs III



Sampling bias: the planet is not observed uniformly (Poisot et al., 2020)

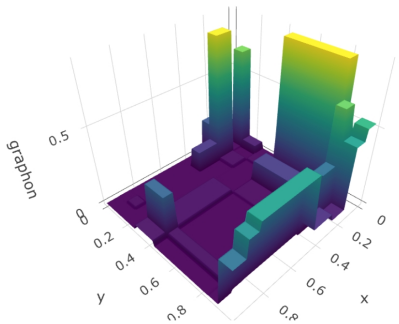
Mangal Foodwebs IV



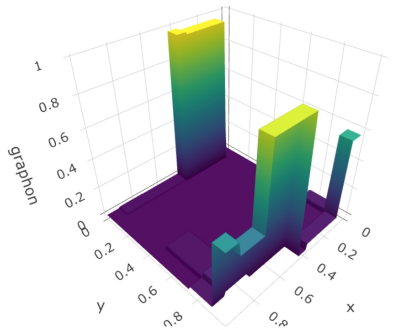
Impact of climat conditions

Mangal Foodwebs V

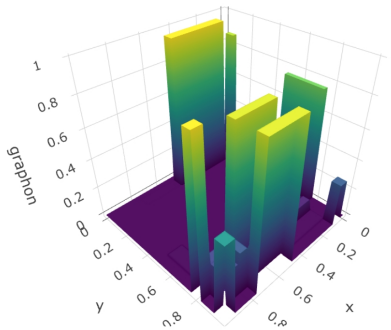
Cluster 1



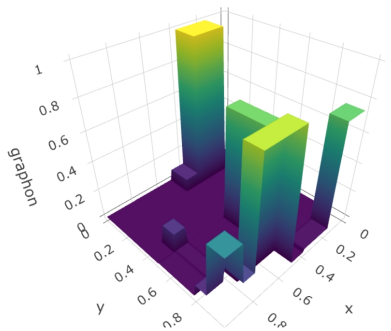
Cluster 2



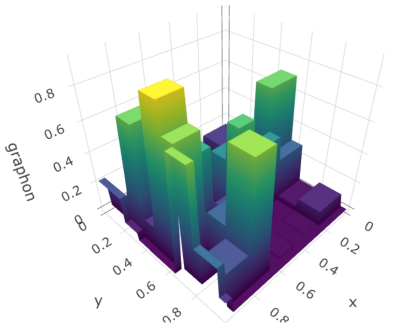
Cluster 3



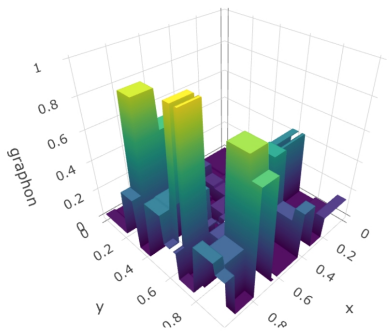
Cluster 4



Cluster 8

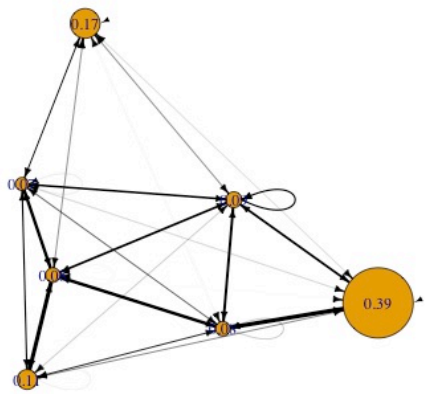


Cluster 10

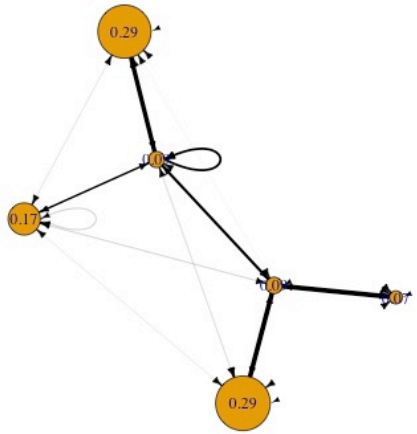


Mangal Foodwebs VI

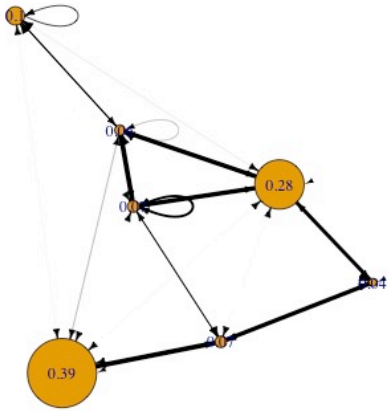
Cluster 1



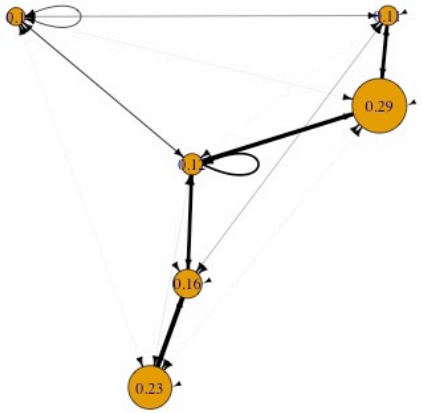
Cluster 2



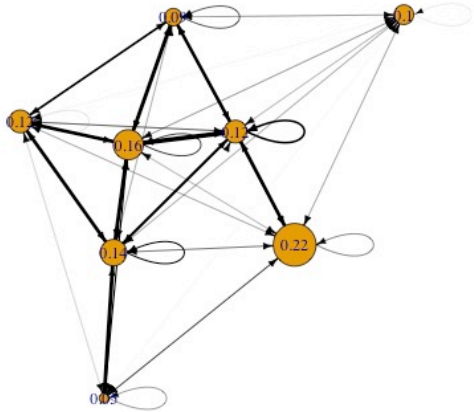
Cluster 3



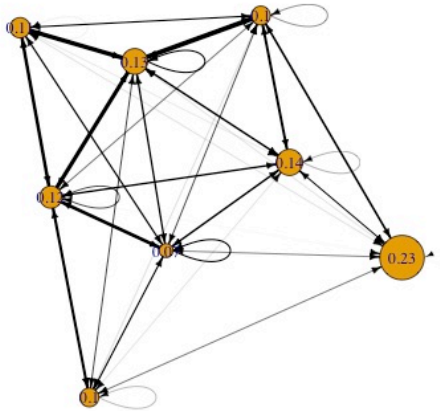
Cluster 4



Cluster 8



Cluster 10



Conclusion & perspectives I

Hierarchical algorithm

- better scalability than EM-type algorithms
- automatic search of the best number of clusters
- cluster hierarchy (dendrogram)
- interpretable output (SBM parameters)

Conclusion & perspectives II

Opens many questions on foodwebs

- interpretation of the network organizations
- analysis of species
- analysis of individual networks
- outlier detection (climate change)
- prediction of the cluster of a new network
- variants of SBM (varying intensity, varying group proportions etc.)
- include covariables

Bibliography

Côme, E. and Latouche, P. (2015). Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, 15(6):564–589.

Lovász, L. (2012). *Large Networks and Graph Limits.*, volume 60 of *Colloquium Publications*. American Mathematical Society.

Poisot, T., Baiser, B., Dunne, J. A., Kéfi, S., Massol, F. c., Mouquet, N., Romanuk, T. N., Stouffer, D. B., Wood, S. A., and Gravel, D. (2016). mangal – making ecological network analysis simple. *Ecography*, 39(4):384–390.

Poisot, T., Bergeron, G., Cazelles, K., Dallas, T., Gravel, D., Macdonald, A., Mercier, B., Violet, C., and Vissault, S. (2020). Environmental biases in the study of ecological networks at the planetary scale. *bioRxiv*.