Comparaison de structures de réseaux. Applications à des réseaux écologiques

Audition candidature de thèse à l'EDMH

Louis Lacoste

23 mai 2024

Parcours

Formations

- 2023–2024, M2 Mathématiques pour les Sciences du Vivant, Université Paris-Saclay
  UC à choix 2nd semestre : modèles à variables latentes, statistiques spatiales et méthodes de stats en grande dimensions
- 2022–2023, Année de césure
- 2020–2022, 1ère et 2ème année en formation Ingénieur AgroParisTech
  Cours optionnels suivis : statistiques spatiales, mathématiques pour la santé, ingénierie par la simulation informatique ...
- 2018–2020, Classe Préparatoire BCPST

Expériences professionnelles

- 2024 Avril–Sept., Détection de structures et clustering de réseaux écologiques. Stage dans l'UMR MIA Paris-Saclay, supervisé par Pierre Barbillon.
- 2023 Janv.–Juillet, Détection de structures dans des collections de réseaux bipartites et écriture du package implémentant la méthode. Stage dans l'UMR MIA Paris-Saclay, supervisé par Pierre Barbillon.
- 2022 Mai–Déc., Stage assistant ingénieur en Qualité chez Eurofins Food France

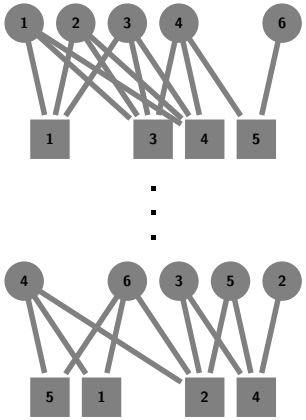
Axes de recherche

Contexte écologique

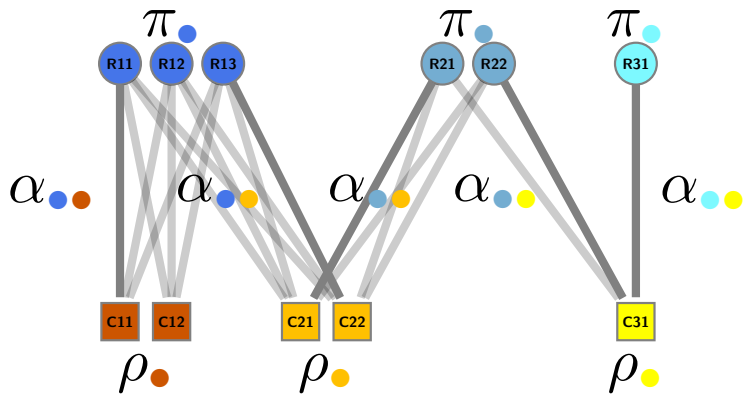
- Faire de la détection de structure sur un réseau (SBM, LBM) mais intérêt à le faire sur plusieurs
- De nombreux réseaux disponibles (« Web of Life : Ecological Networks Database », s. d.) et décrivant des interactions similaires. Par exemple des interactions proies-prédateurs, plantes-pollinisateurs ...
- Re-grouper les réseaux selon leur similarité (*clustering* de réseaux)
- Transférer de l'information grâce à la collection (par exemple reconstitution de données manquantes)
- Déterminer des structures d'interactions fines de manière agnostique

Contexte mathématiques

Collections bipartites



$\overset{iid}{\sim}$



LBM

Pour

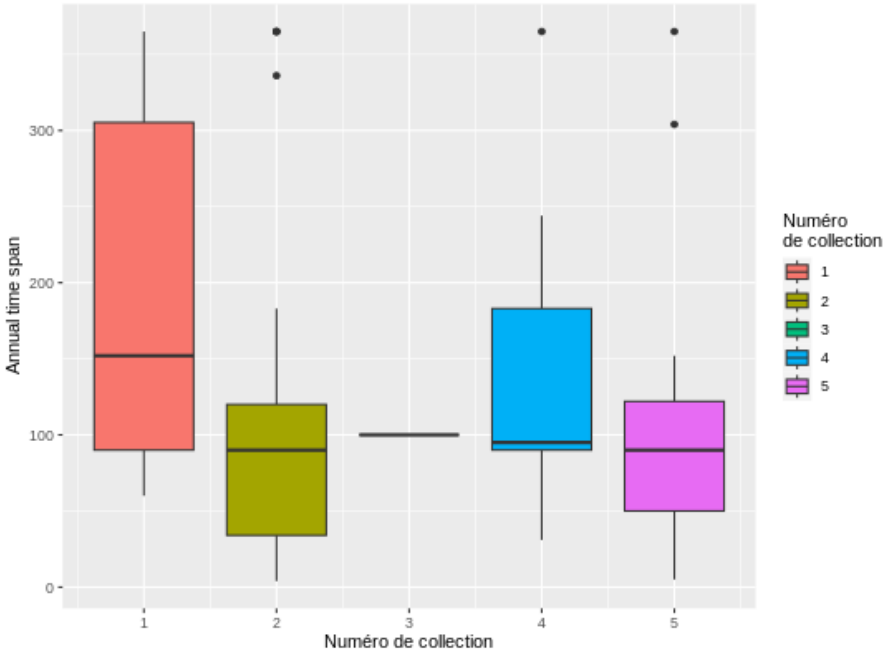
- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

Paramètres

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

Application, données plantes pollinisateurs

Voici des résultats du modèle *iid-colBiSBM* sur des données plantes-pollinisateurs (Doré et al., 2021 et Thébault et Fontaine, 2020)



N°de collection	1	2	3	4	5
Nombre de réseaux	38	45	1	20	19

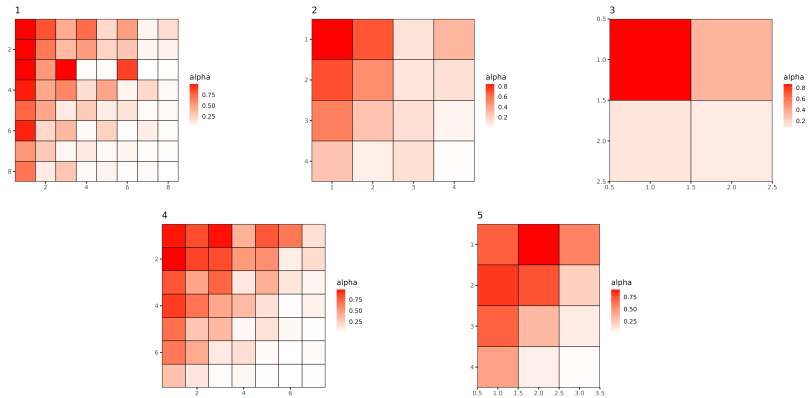


Figure – Connectivités de la partition

Organisation de la thèse

Planning prévisionnel de la thèse

TODO Ici une timeline

Financement

L'INRAE, par le département MathNum donne 50% des financements de la thèse.

Merci pour votre attention.

Bibliographie I

Web of Life : Ecological Networks Database. (s. d.). Récupérée 17 juin 2023, à partir de https://www.web-of-life.es/map.php

Doré, M., Fontaine, C., & 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. https://doi.org/10.1111/gcb.15474

Thébault, E., & Fontaine, C. (2020, décembre 1). *A Database of Plant-Pollinator Networks* (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4300427

Govaert, G., & Nadif, M. (2005). An EM Algorithm for the Block Mixture Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(4), 643-647. https://doi.org/10.1109/TPAMI.2005.69

Bibliographie II

Chabert-Liddell, S.-C., Barbillon, P., & Donnet, S. (2023, mars 27). *Learning Common Structures in a Collection of Networks. An Application to Food Webs*. arXiv : 2206.00560 [stat]. https://doi.org/10.48550/arXiv.2206.00560

Anakok, E., Barbillon, P., Fontaine, C., & Thebault, E. (2022, novembre 29). *Disentangling the structure of ecological bipartite networks from observation processes*. arXiv : 2211.16364 [stat]. Récupérée 14 juin 2023, à partir de http://arxiv.org/abs/2211.16364

Modèles à variables latentes pour collection de réseaux bipartites

Latent Block Model (LBM [2])

Proposé par Govaert et Nadif, 2005.

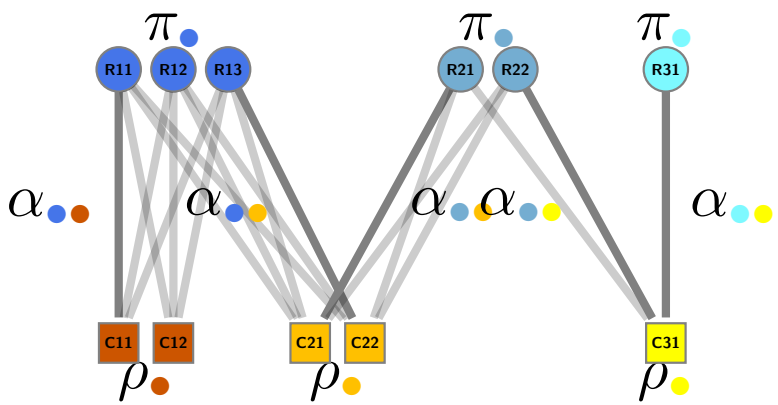


Figure – Exemple de LBM [1]

Pour

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

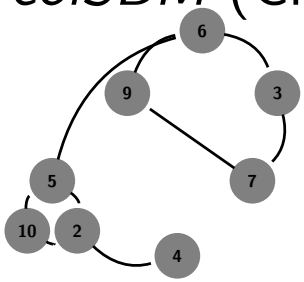
Paramètres

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

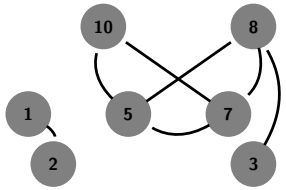
1. Que j'appellerai par la suite BiSBM

colSBM

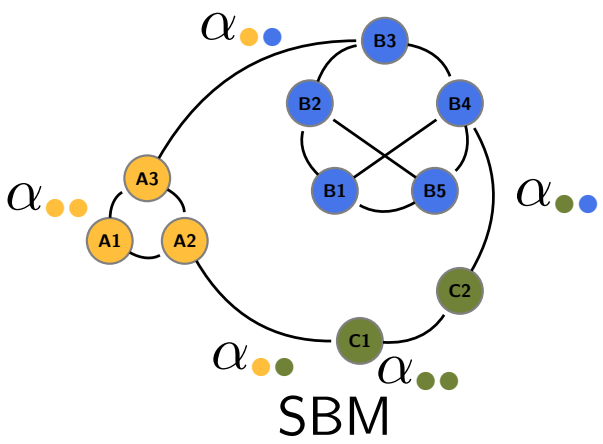
Le modèle *colSBM* (Chabert-Liddell et al., 2023).



$\vdots$



$\overset{iid}{\sim}$

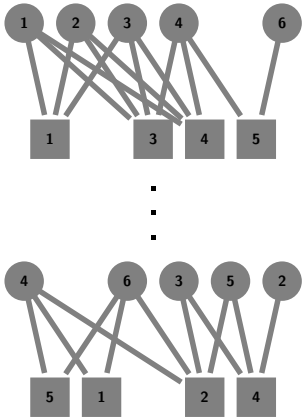


Pour $Q = |\{\bullet, \bullet, \bullet\}|$ blocs fixés :

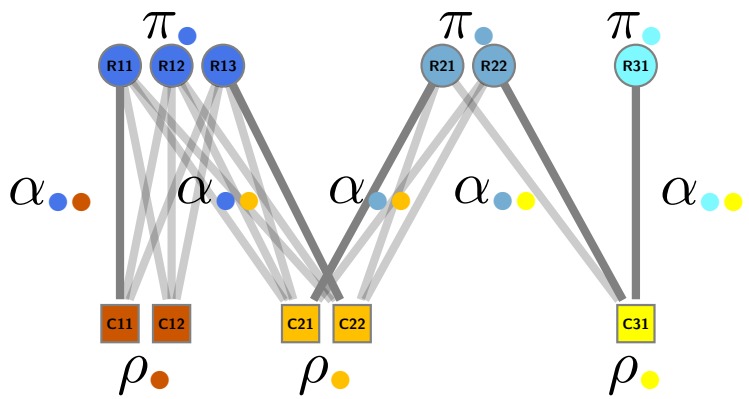
Paramètres

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, Z_j = \bullet)$

Collections bipartites



$\overset{iid}{\sim}$



BiSBM

Pour

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

Paramètres

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

Différents modèles

iid-colBiSBM

$\boldsymbol{\pi} = (\pi_1, \dots \pi_{Q_1})$ et $\boldsymbol{\rho} = (\rho_1, \dots \rho_{Q_2})$, tous les réseaux partagent les mêmes paramètres [2]

$\pi\rho$-colBiSBM

$\boldsymbol{\pi} = ((\pi_1{}^m, \dots \pi_{Q_1}{}^m))_{m=1,\dots M}$ et $\boldsymbol{\rho} = ((\rho_1{}^m, \dots \rho_{Q_2}{}^m))_{m=1,\dots M}$
avec $\forall q, m \in [\![1, Q_1]\!] \times [\![1, M]\!], \pi_q^m \in [0,1]$ et $\forall r, m \in [\![1, Q_2]\!] \times [\![1, M]\!], \rho_r^m \in [0,1]$

Et également deux autres modèles (π-colBiSBM et ρ-colBiSBM) où seulement une des deux dimensions est libre.

2. Dans tous les modèles la structure de connectivité est supposée identique au sein de la collection.

Estimation des paramètres

Maximisation d'une borne inférieure de la log-vraisemblance des données observées.

$$\ell(\mathbf{X};\boldsymbol{\theta}) \geq \sum_{m=1}^{M} \Bigg(\sum_{i=1}^{n_1^m} \sum_{j=1}^{n_2^m} \sum_{q\in\mathcal{Q}_{1,m}} \sum_{r\in\mathcal{Q}_{2,m}} \tau_{i,q}^{1,m}\tau_{j,r}^{2,m} \log f(X_{ij}^m;\alpha_{qr}) + \sum_{i=1}^{n_1^m} \sum_{q\in\mathcal{Q}_{1,m}} \tau_{i,q}^{1,m} \log {\pi_q}^m + \sum_{j=1}^{n_2^m} \sum_{r\in\mathcal{Q}_{2,m}} \tau_{j,r}^{2,m} \log {\rho_r}^m - \sum_{i=1}^{n_1} \tau_{i,q}^{1,m} \log \tau_{i,q}^{1,m} - \sum_{j=1}^{n_2} \tau_{j,r}^{2,m} \log \tau_{j,r}^{2,m} \Bigg) =: J(\boldsymbol{\tau};\boldsymbol{\theta})$$

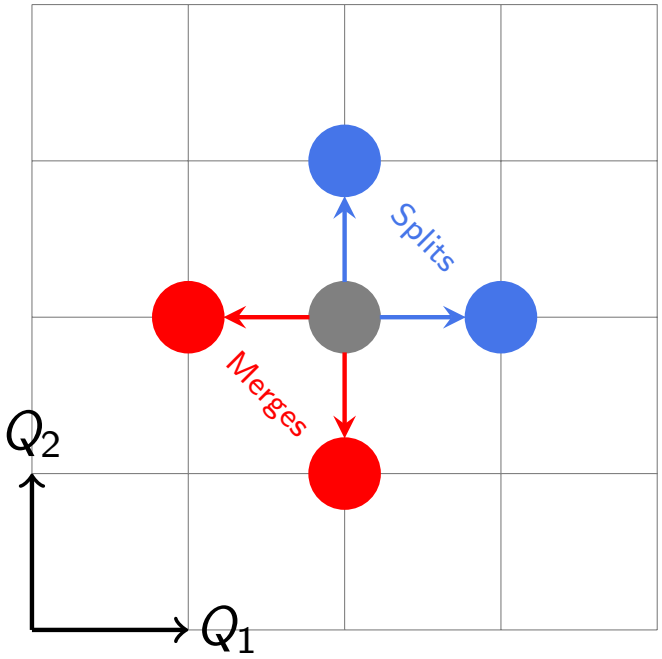
Approximation variationnelle

$\tau_{i,q}^{1,m} = P(Z_i = q|X_{ij}^m)$ et $\tau_{j,r}^{2,m} = P(W_j = r|X_{ij}^m)$ tels que
$P(Z_i = q, W_j = r|X_{ij}^m) = \tau_{i,q}^{1,m} \times \tau_{i,r}^{2,m}$

Sélection de modèle : choix de (Q_1, Q_2) - Approche gloutonne

Le VEM se fait à Q_1, Q_2 fixés, il faut donc déterminer les "meilleures" coordonnées. Nous maximisons un BIC-L [3].

Détermination d'un premier mode par approche *gloutonne*

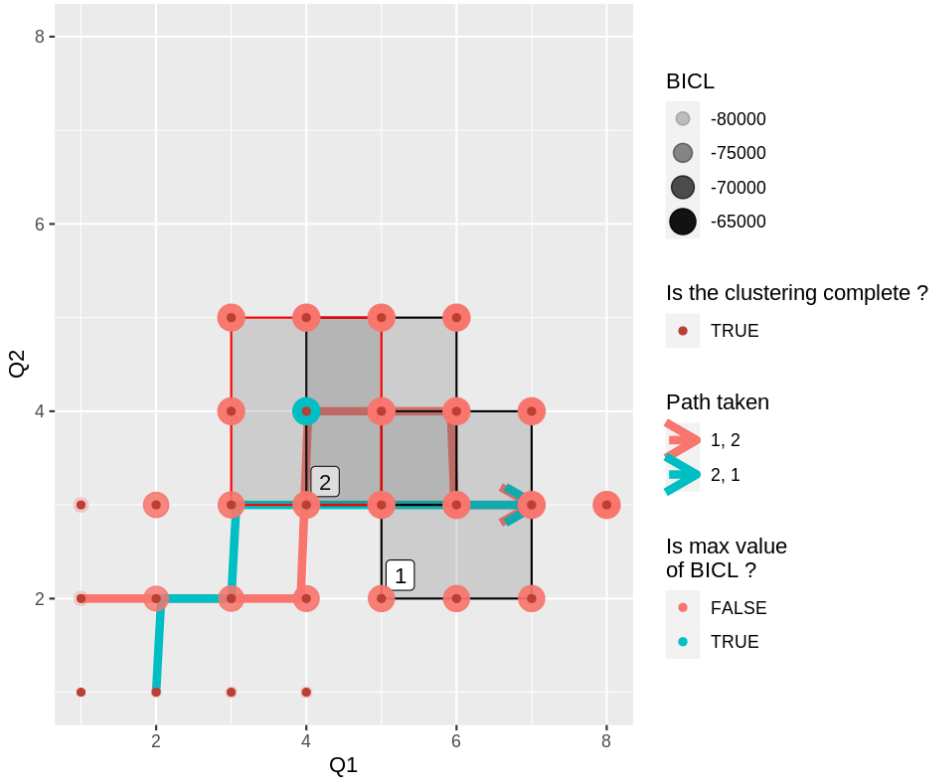


Exploration gloutonne

- Initialisation sur $(1, 2)$ et $(2, 1)$
- Exploration des 4 voisins et déplacement sur le meilleur des 4
- Arrêt après 2 étapes successives sans augmentation du BIC-L

3. *Bayesian Information Criterion - Like*, en adaptant les formules de Chabert-Liddell et al., 2023

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante



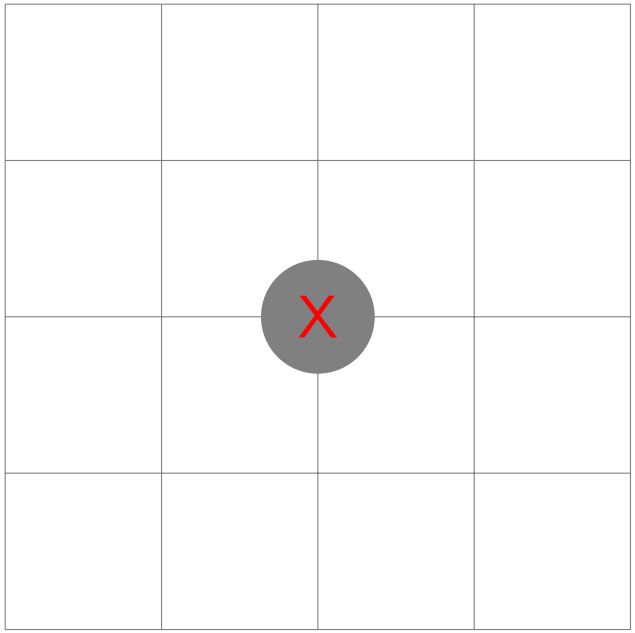
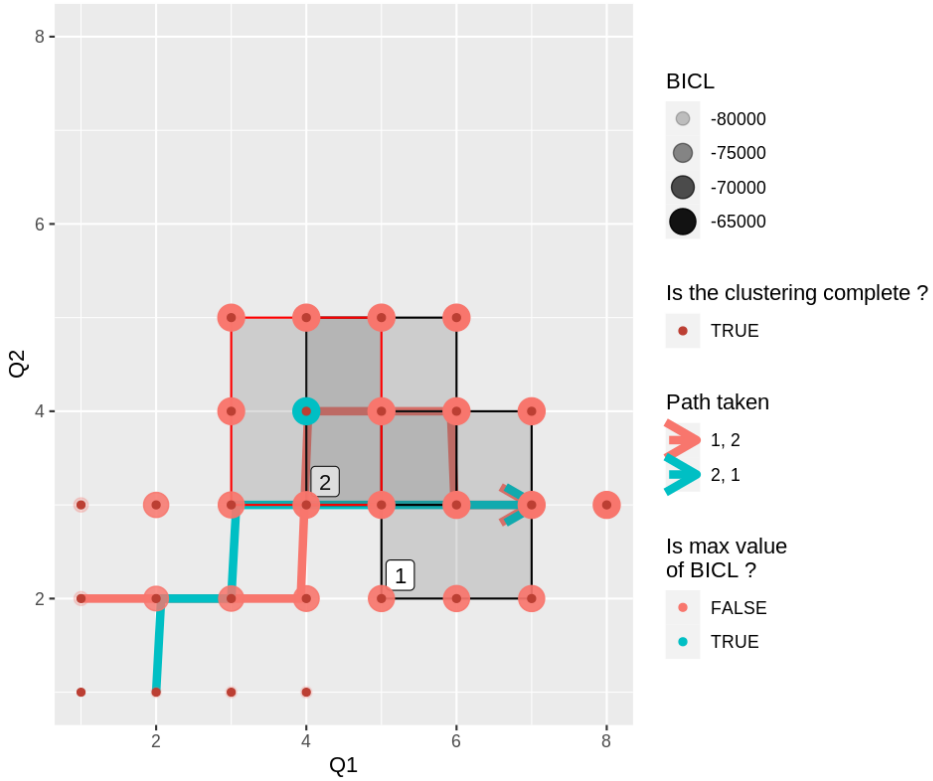


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante



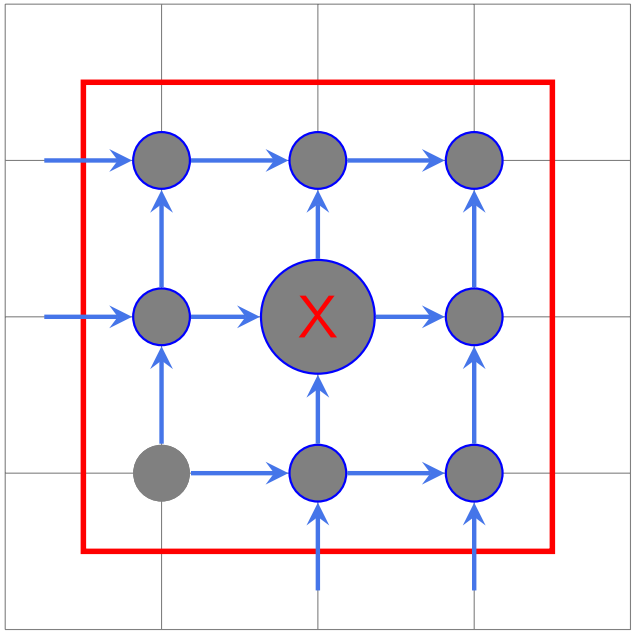
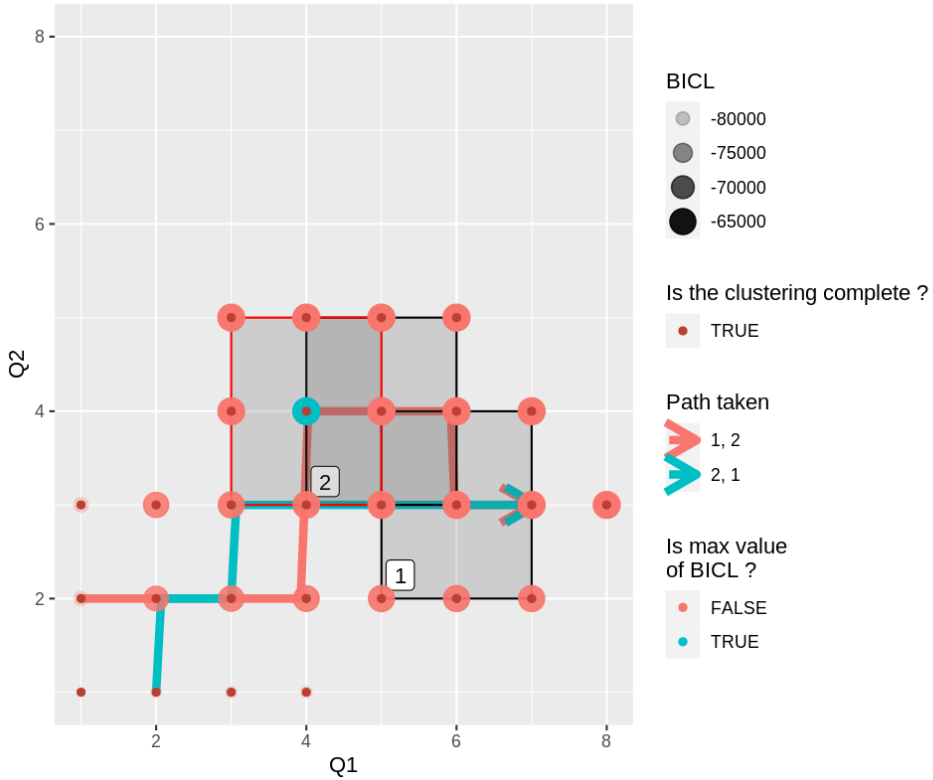


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante



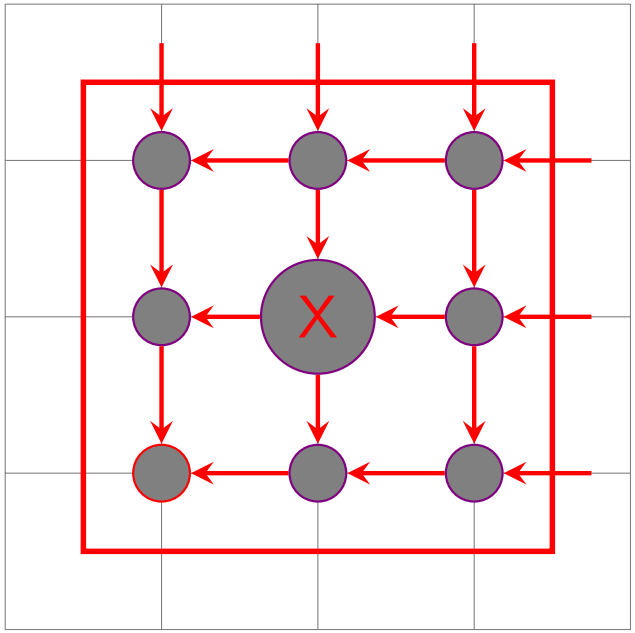
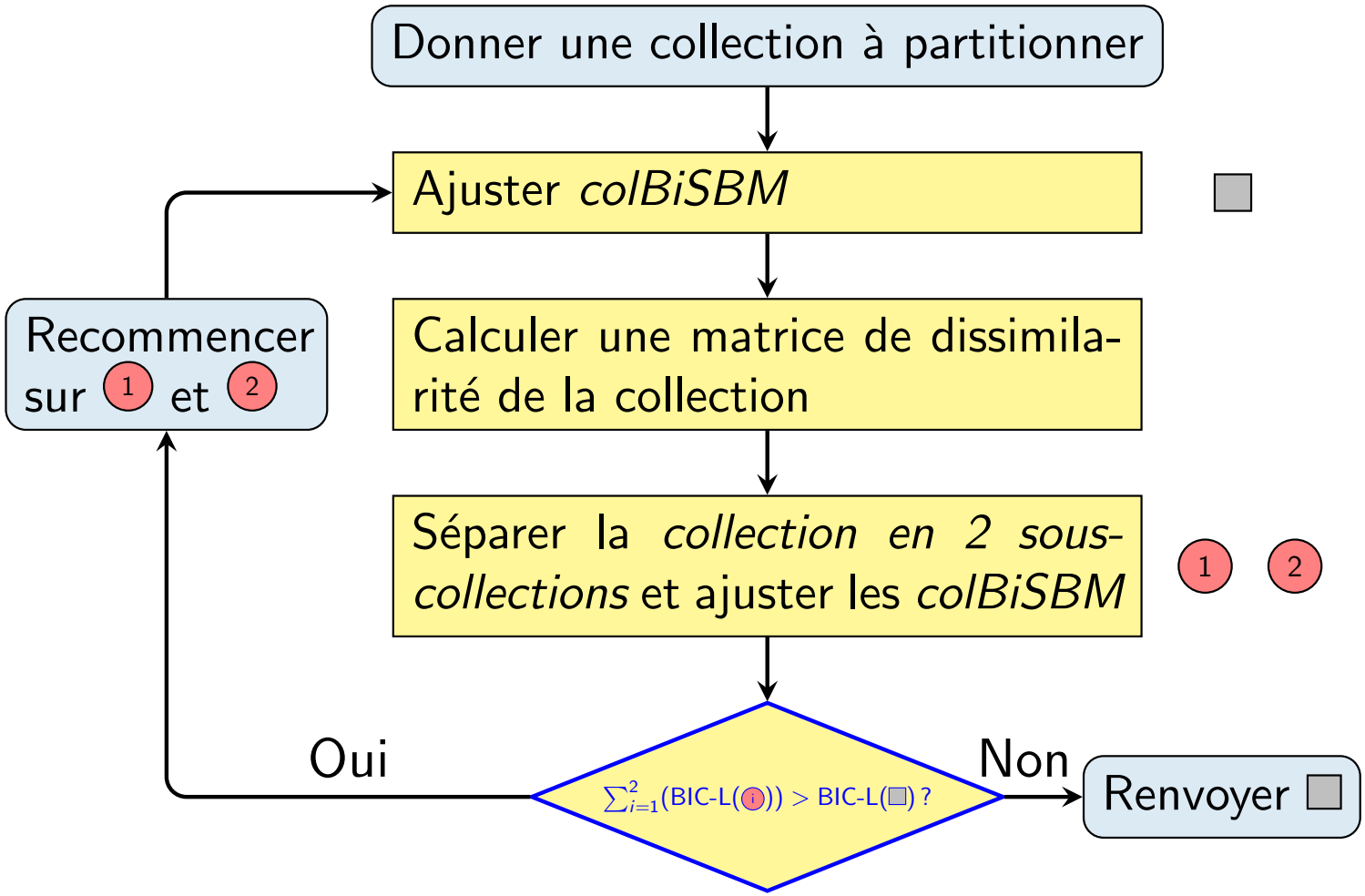


Figure – Exemple de parcours de fenêtre glissante

Clustering de réseaux

Objectif

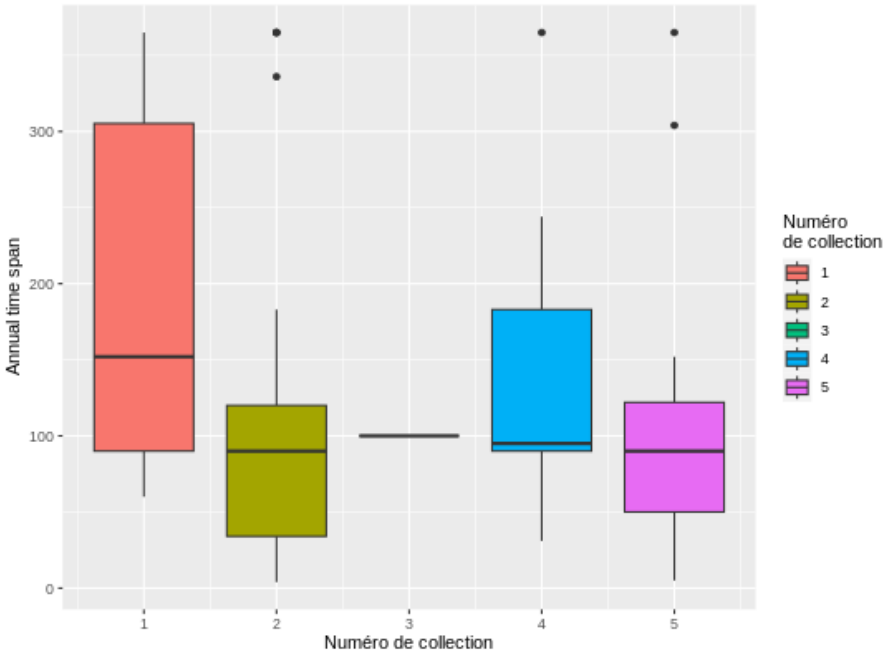
Déterminer une partition qui maximise la somme du BICL de ses sous-collections.



Même approche que Chabert-Liddell et al., 2023

Application, données plantes pollinisateurs

Voici des résultats du modèle *iid-colBiSBM* sur des données plantes-pollinisateurs (Doré et al., 2021 et Thébault et Fontaine, 2020)



N°de collection	1	2	3	4	5
Nombre de réseaux	38	45	1	20	19

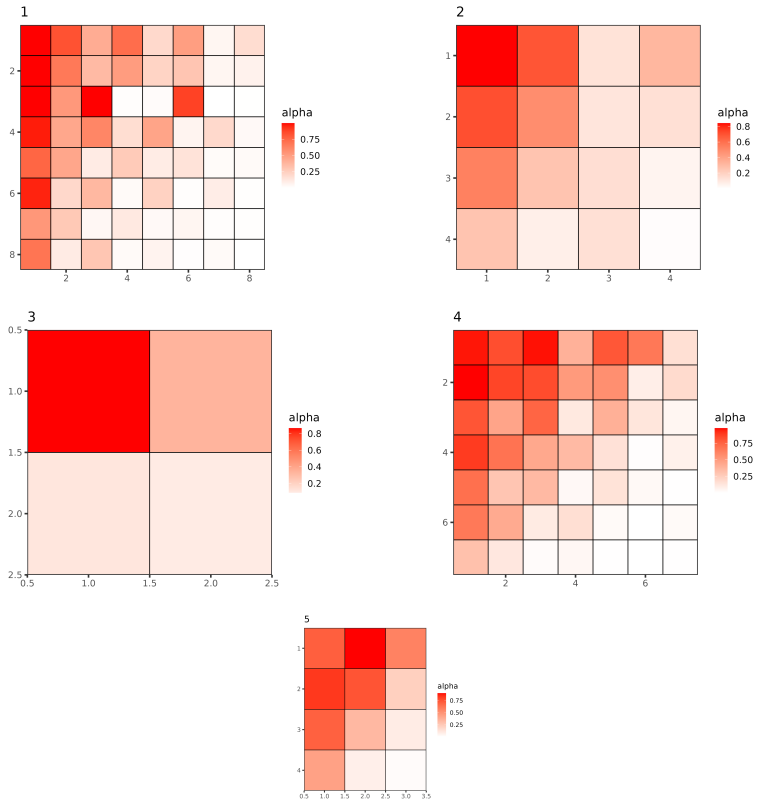


Figure – Connectivités de la partition

Perspectives sur *colSBM*

- 4 modèles dont 3 qui ont une flexibilité sur au moins une des dimensions (adaptabilité aux données)
- Partitionner un ensemble de réseaux selon leurs structures
- Comparer les *clusterings* de réseaux obtenus entre données brutes et données corrigées (par exemple par la méthode *CoOPLBM*[4])

Le package est disponible sur GitHub :
https://github.com/Chabert-Liddell/colSBM

4. Anakok et al., 2022