

Comparaison de structures de réseaux. Applications à des réseaux écologiques

Audition de candidature de thèse à l'EDMH

Louis Lacoste

Supervisé par Sophie Donnet et Pierre Barbillon, co-encadré par Julie Aubert

UMR MIA Paris-Saclay

23 mai 2024

- 1 Parcours
- 2 Sujet de thèse
- 3 Organisation de la thèse

Parcours

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

- 2022 Mai–Déc., Stage assistant ingénieur en Qualité chez Eurofins Food France
- 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.
- 2024 Avril–Sept., Détection de structures et clustering de réseaux écologiques. Stage dans l'UMR MIA Paris-Saclay, supervisé par Pierre Barbillon et Sophie Donnet.

Sujet de thèse

Contexte écologique

- Nombreux réseaux disponibles (« [Web of Life : Ecological Networks Database](#) », s. d.) pour interactions similaires. Par exemple, interactions proies-prédateurs, plantes-pollinisateurs . . .
- Suivi biodiversité, analyse de robustesse et risque d'effondrement

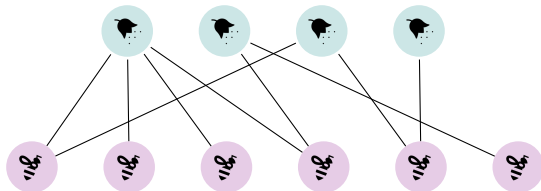


Figure 1 – Exemple d'un réseau plantes-pollinisateurs

- En écologie microbienne réseaux permettent le suivi de la qualité des sols.

Détection de structure¹ pour un réseau bien connu :

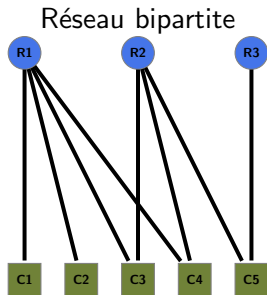
- Modèles de *clustering* à variables latentes
- *Embedding* par apprentissage profond

Mais des motivations pour considérer des collections :

- Espèces différentes, rôles analogues
- Transfert d'informations grands vers petits réseaux.
- Regrouper les réseaux selon leur similarité (*clustering* de réseaux)

Axe 1 : Modèles à variables latentes pour une collection de réseaux bipartites

Réseaux bipartites



Matrice d'incidence

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

Décrit des interactions (pas uniquement binaires) entre deux groupes d'agents :

- hôtes-parasites
- plantes-pollinisateurs
- graines-disperseurs
-

Latent Block Model (LBM)

Proposé par Govaert et Nadif, 2005.

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

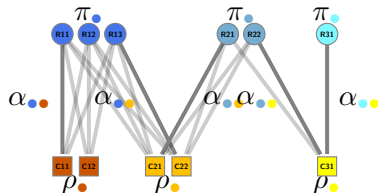
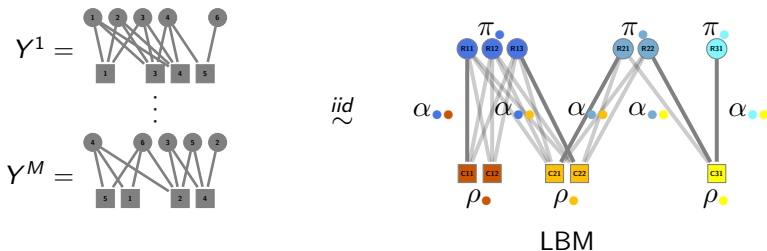


Figure 2 – Exemple de LBM²

2. Que j'appelle par la suite BiSBM

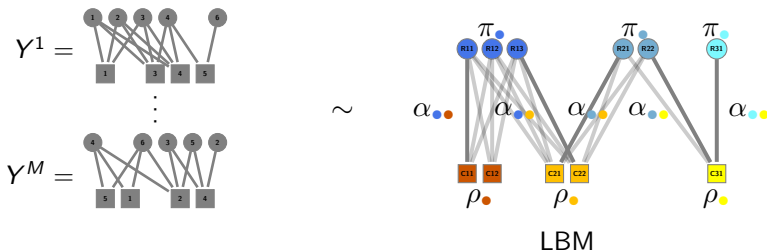
Collections bipartites



Modèle *iid*-colBiSBM

$$\forall m \in \llbracket 1, M \rrbracket, Y_i \sim LBM_{n_1^m, n_2^m}(\pi, \rho, \alpha)$$

Collections bipartites



Modèle $\pi\rho$ -colBiSBM

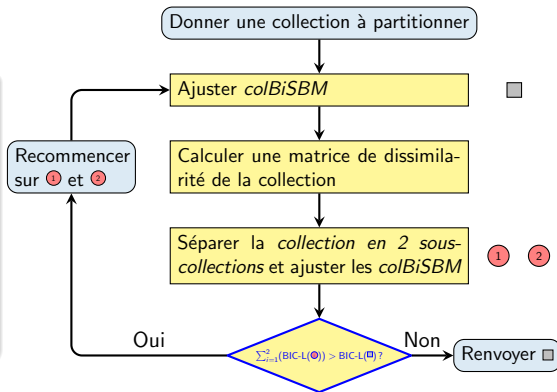
$$\forall m \in \llbracket 1, M \rrbracket, Y_i \sim \text{LBM}_{n_1^m, n_2^m}(\pi^m, \rho^m, \alpha)$$

- Écriture du modèle colBiSBM
- Dérivation des formules d'inférence et d'un critère de sélection de modèle de vraisemblance pénalisée
- Implémentation des formules et du critère et développement algorithmique pour l'exploration de l'espace de paramètres.
- Partitionnement d'une large collection de réseaux.
- Écriture du code s'intégrant au package³ écrit par Saint-Clair Chabert-Liddell.

Clustering de réseaux

Objectif

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



Même approche que [Chabert-Liddell et al., 2024](#)

Application du *clustering*, données plantes pollinisateurs

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

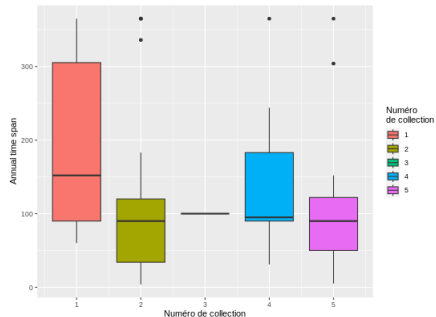
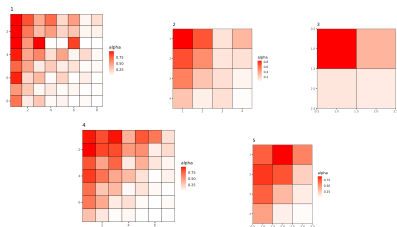


Figure 3 – Connectivités de la partition

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

- Finaliser l'analyse sur données réelles commencée sur [Doré et al., 2021](#) ; [Thébault et Fontaine, 2020](#) avec les interprétations des écologues en vue d'une publication.
- Preuve d'identifiabilité du modèle ([Chabert-Liddell et al., 2024](#) ; [Celisse et al., 2012](#) ; [Keribin et al., 2015](#) ; [Brault & Mariadassou, 2015](#))

Axe 2 : Embedding de nœuds par apprentissage profond pour comparaison des topologies de réseaux

Graph Neural Networks I

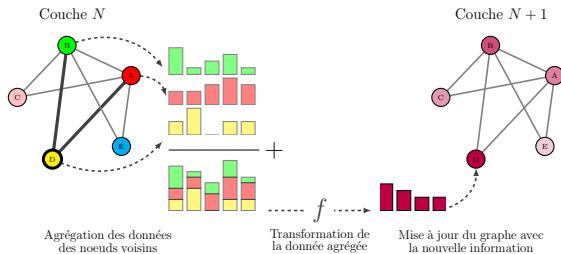


Figure 4 – Illustration du *message passing* au sein d'un graphe.⁴

Graph Neural Networks II

Avec les *Graph Convolutional Networks* (GCN) *embedding* de graphes (Veličković et al., 2018 ; Hamilton et al., s. d. ; Xu et al., 2019) tenant compte des invariances.

Règle de propagation d'une couche de GCN

$$H^{(l+1)} = \sigma(\tilde{D}^{\frac{1}{2}} \tilde{A} \tilde{D}^{\frac{1}{2}} H^{(l)} W^{(l)}), \quad (1)$$

tirée de Kipf et Welling, 2017.

- Utiliser des *Variational Auto-Encoder* (VAE) (Kipf & Welling, 2017, 2016) et résume le réseau par une distribution. Calculer distance de Gromov-Wasserstein pour comparaison et classification.

Un des avantages principaux est le *passage à l'échelle* de ces méthodes permettant de traiter des réseaux de plus grande taille.

Axe 3 : Inférence jointe de réseaux

En écologie, réseaux inférés à partir de table de co-occurences. TODO
Insérer une table de co-occurrence Incertitude connue mais négliger dans la
suite de l'analyse.

Limites des techniques actuelles [Matchado et al., 2021](#). Rôle important
pour les réseaux reconstruits notamment en microbiologie.

Organisation de la thèse

Merci pour votre attention.

Bibliographie I

- Web of Life : Ecological Networks Database.* (s. d.). Récupérée juin 17, 2023, à partir de <https://www.web-of-life.es/map.php>
- 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>
- Chabert-Liddell, S.-C., Barbillon, P., & 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.
<https://doi.org/10.1214/23-AOAS1831>
- 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>

Bibliographie II

- Thébault, E., & Fontaine, C. (2020, décembre 1). *A Database of Plant-Pollinator Networks* (Version 1). *Zenodo*.
<https://doi.org/10.5281/zenodo.4300427>
- Celisse, A., Daudin, J.-J., & Pierre, L. (2012). Consistency of Maximum-Likelihood and Variational Estimators in the Stochastic Block Model. *Electronic Journal of Statistics*, 6, 1847-1899.
<https://doi.org/10.1214/12-EJS729>
- Keribin, C., Brault, V., Celeux, G., & Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Stat Comput*, 25(6), 1201-1216.
<https://doi.org/10.1007/s11222-014-9472-2>
- Brault, V., & Mariadassou, M. (2015). Co-clustering through Latent Bloc Model : a Review. *Journal de la société française de statistique*, 156(3), 120-139. Récupérée mai 15, 2024, à partir de http://www.numdam.org/item/JSFS_2015__156_3_120_0/

Bibliographie III

- Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltchko, A. B. (2021). A Gentle Introduction to Graph Neural Networks. *Distill*, 6(9), e33. <https://doi.org/10.23915/distill.00033>
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018, février 4). *Graph Attention Networks*. arXiv : 1710.10903 [cs, stat]. Récupérée mai 14, 2024, à partir de <http://arxiv.org/abs/1710.10903>
- Hamilton, W., Ying, Z., & Leskovec, J. (s. d.). Inductive Representation Learning on Large Graphs.
- Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019, février 22). *How Powerful Are Graph Neural Networks?* arXiv : 1810.00826 [cs, stat]. <https://doi.org/10.48550/arXiv.1810.00826>

- Kipf, T. N., & Welling, M. (2017, février 22). *Semi-Supervised Classification with Graph Convolutional Networks*. arXiv : 1609.02907 [cs, stat].
<https://doi.org/10.48550/arXiv.1609.02907>
- Kipf, T. N., & Welling, M. (2016, novembre 21). *Variational Graph Auto-Encoders*. arXiv : 1611.07308 [cs, stat].
<https://doi.org/10.48550/arXiv.1611.07308>
- Matchado, M. S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D., & List, M. (2021). Network Analysis Methods for Studying Microbial Communities : A Mini Review. *Computational and Structural Biotechnology Journal*, 19, 2687-2698. <https://doi.org/10.1016/j.csbj.2021.05.001>

Annexes

Modèles à variables latentes pour collection de réseaux bipartites

Latent Block Model (LBM²)

Proposé par Govaert et Nadif, 2005.

Pour

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

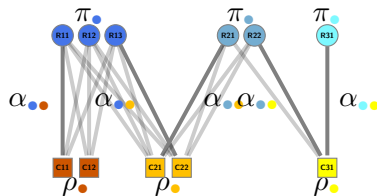
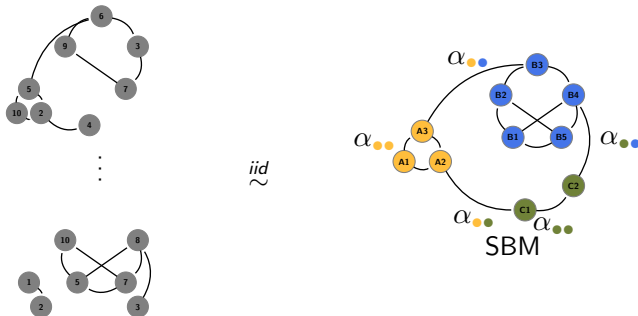


Figure 6 – Exemple de LBM⁵

Paramètres

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

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

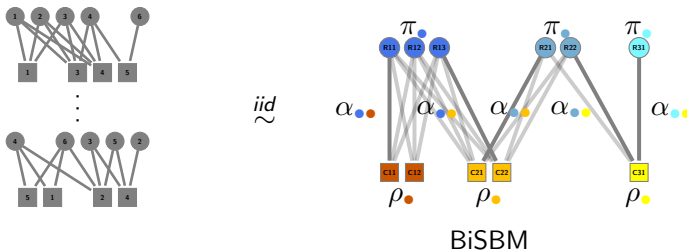


Pour $Q = |\{\bullet, \bullet, \bullet\}|$ blocs fixés :

Paramètres

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

Collections bipartites



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

iid-colBiSBM

$\pi = (\pi_1, \dots, \pi_{Q_1})$ et $\rho = (\rho_1, \dots, \rho_{Q_2})$, tous les réseaux partagent les mêmes paramètres⁶

$\pi\rho$ -colBiSBM

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

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

6. 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.

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

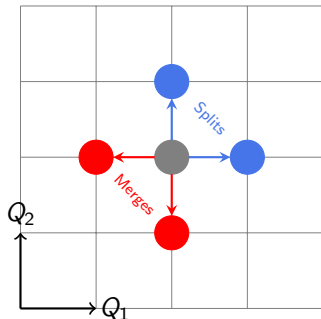
Approximation variationnelle

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

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-L⁷.

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

7. *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

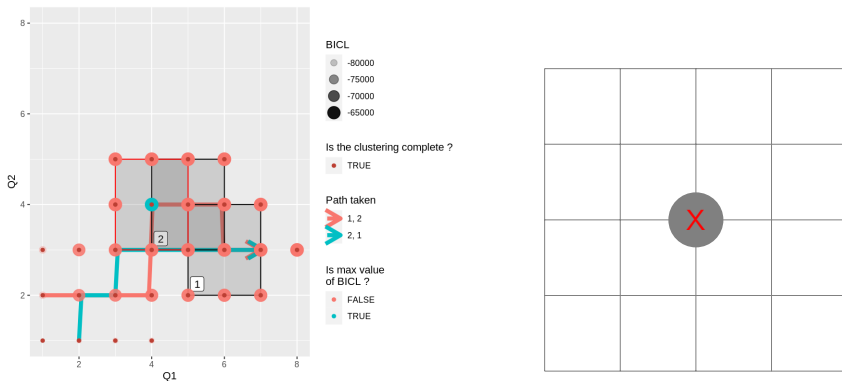


Figure 7 – Exemple de parcours de fenêtre glissante

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

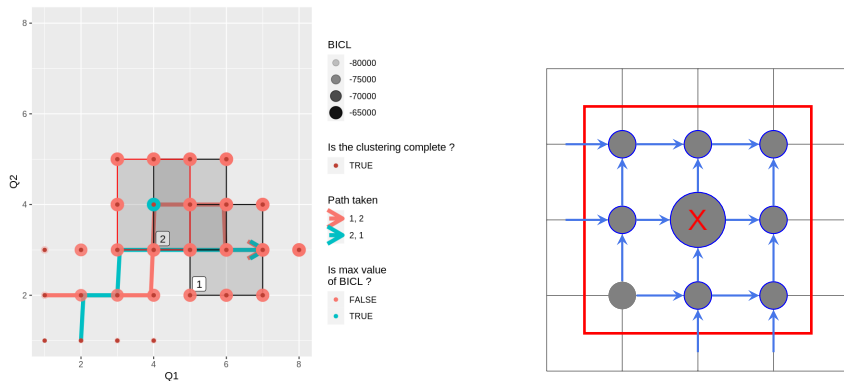


Figure 7 – Exemple de parcours de fenêtre glissante

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

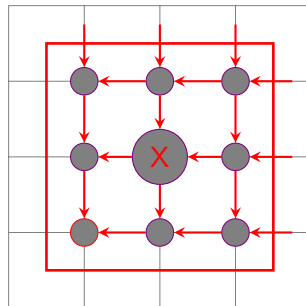
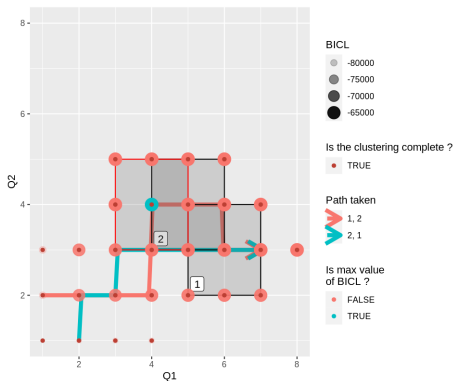
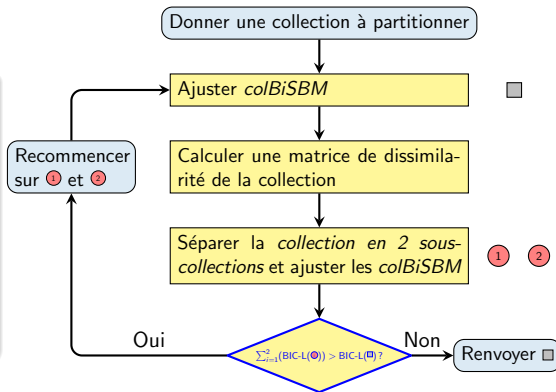


Figure 7 – Exemple de parcours de fenêtre glissante

Clustering de réseaux

Objectif

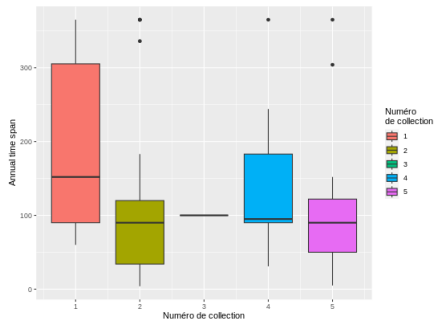
Déterminer une partition qui maximise la somme du BICL de ses sous-collections.



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

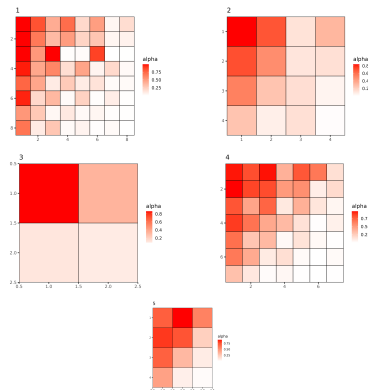



Figure 8 – Connectivités de la partition

- 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*⁸)

Le package est disponible sur GitHub :

 <https://github.com/Chabert-Liddell/colSBM>

Bibliographie des annexes I

- 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>
- 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>
- 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>

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 juin 14, 2023, à partir de <http://arxiv.org/abs/2211.16364>