

# Comparaison de structures de réseaux. Applications à des réseaux écologiques

Audition de candidature de thèse à l'EDMH

Louis Lacoste

23 mai 2024

## 1 Parcours

## 2 Axes de recherche

- Axe 1 : Modèles à variables latentes pour une collection de réseaux bipartites
- Axe 2 : Embedding de nœuds par apprentissage profond pour comparaison des topologies de réseaux
- Axe 3 : Inférence jointe de réseaux

## 3 Organisation de la thèse

# Parcours

- 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

- 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 . . .
- Ces réseaux permettent un suivi de la biodiversité, de détecter et d'analyser la robustesse et les changements subies par ces écosystèmes et notamment les risques d'effondrement de la biodiversité.
- En écologie microbienne, les réseaux sont construits sur la base de co-occurrences et reconstruits par inférence des liens mais rarement par observation directe.

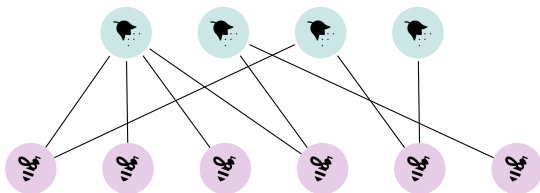


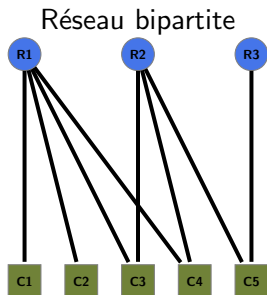
Figure 1 – Exemple d'un réseau plantes-pollinisateurs

- Faire de la détection de structure réseau par réseau de manière agnostique (SBM, LBM) est bien connu. Mais il y a de l'intérêt à le faire sur plusieurs :
  - ▶ Des espèces différentes dans plusieurs réseaux pourrait remplir des rôles similaires
  - ▶ Les petits réseaux pourraient bénéficier d'une estimation avec des réseaux plus grands et révéler une structure plus fine.
  - ▶ Certains réseaux étant moins bien échantillonnés que d'autre une prise en compte en collection de réseaux pourrait aider à transférer de l'information
- Re-grouper les réseaux selon leur similarité (*clustering* de réseaux)
- Transférer de l'information grâce à la collection (par exemple reconstitution de données manquantes)
- Proposer des comparaisons en extrayant plus d'informations que les métriques classiques



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

# Réseaux bipartites



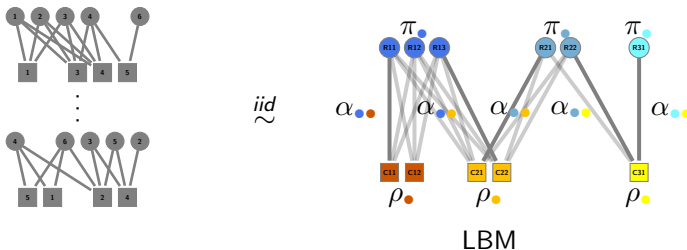
Matrice d'incidence

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

Permet de décrire des interactions impliquant deux agents dont les rôles sont de natures différentes.

Par exemple : hôtes-parasites, plantes-pollinisateurs, graines-disperseurs ...

# Collections bipartites



Pour

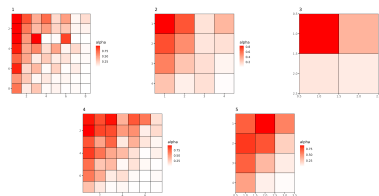
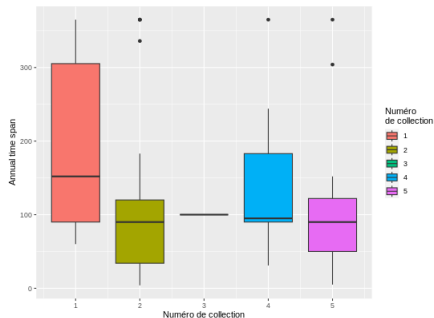
- $Q_1 = |\{\bullet, \bullet, \bullet\}|$  blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$  blocs fixés en colonne

## Paramètres

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

# 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 2 – Connectivités de la partition

## Apport déjà réalisé

- Adaptation du modèle mathématique<sup>a</sup> proposé par [Chabert-Liddell et al., 2024](#) aux réseaux bipartites
- Développement algorithmique pour l'exploration de l'espace de paramètres.
- Écriture du code de la partie bipartite s'intégrant au package<sup>b</sup> écrit par Saint-Clair Chabert-Liddell.

a. Notamment des formules des étapes VE et M et du calcul de dissimilarité.

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

## À finir/à faire

- Finaliser l'analyse d'applications sur données réelles commencée sur [Doré et al., 2021](#) ; [Thébault et Fontaine, 2020](#) avec les interprétations des écologues.
- 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

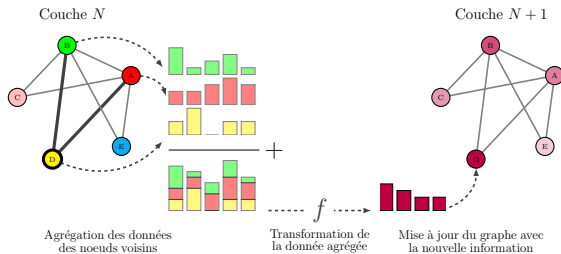


Figure 3 – Illustration du *message passing* au sein d'un graphe.<sup>1</sup>

# Graph Neural Networks II

En utilisant les *Graph Convolutional Networks* (GCN) il est possible de réaliser un *embedding* des graphes (Veličković et al., 2018 ; Hamilton et al., s. d. ; Xu et al., 2019) en tenant compte des invariances qui sont inhérentes à ces objets.

## 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.

Pour, par exemple, utiliser des auto-encodeur variationnels ou VAE (Kipf & Welling, 2017, 2016) par exemple et donc permettant d'obtenir par réseau une distribution. Cela permet alors par exemple de calculer une distance de Gromov-Wasserstein afin de comparer les réseaux et de pouvoir réaliser un *clustering* ou une 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

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

## Organisation de la thèse

## Planning prévisionnel de la thèse

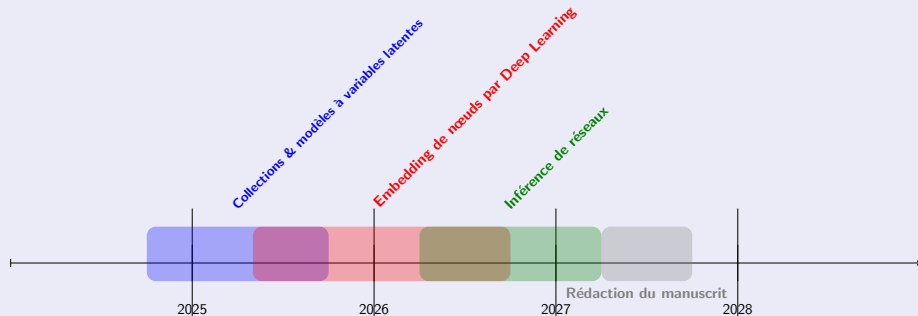


Figure 4 – Chronologie prévue

## Financement

L'INRAE, par le département MathNum accorde déjà 50% des financements 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>
- 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>
- 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>

# Bibliographie II

- 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/](http://www.numdam.org/item/JSFS_2015__156_3_120_0/)
- 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>



# Bibliographie III

- 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>

# Bibliographie IV

- 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>
- 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>
- 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>

# Annexes

# Modèles à variables latentes pour collection de réseaux bipartites

# Latent Block Model (LBM<sup>2</sup>)

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

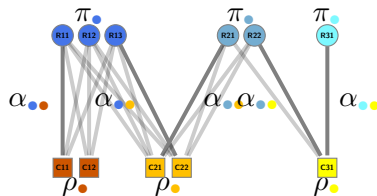


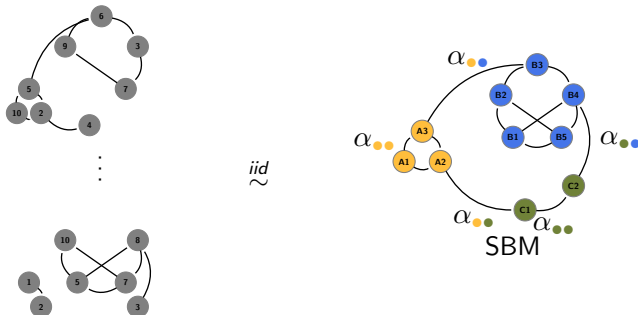
Figure 5 – Exemple de LBM<sup>2</sup>

## Paramètres

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

## 2. Que j'appellerai par la suite BiSBM

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

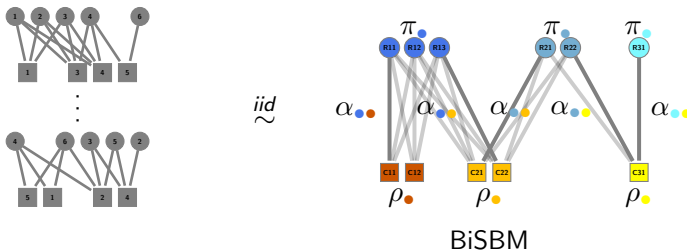


Pour  $Q = |\{\bullet, \bullet, \bullet\}|$  blocs fixés :

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



# Différents modèles

## *iid-colBiSBM*

$\pi = (\pi_1, \dots, \pi_{Q_1})$  et  $\rho = (\rho_1, \dots, \rho_{Q_2})$ , tous les réseaux partagent les mêmes paramètres<sup>3</sup>

## $\pi\rho$ -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 ( $\pi$ -colBiSBM et  $\rho$ -colBiSBM) où seulement une des deux dimensions est libre.

---

3. 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.

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

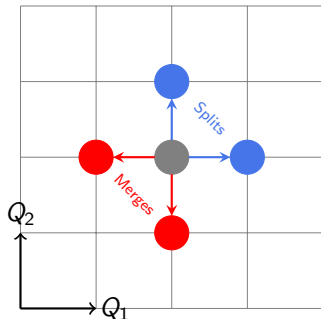
## Approximation variationnelle

$\tau_{i,q}^{1,m} = P(Z_i = q | X_{ij}^m)$  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-L<sup>4</sup>.

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

4. *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

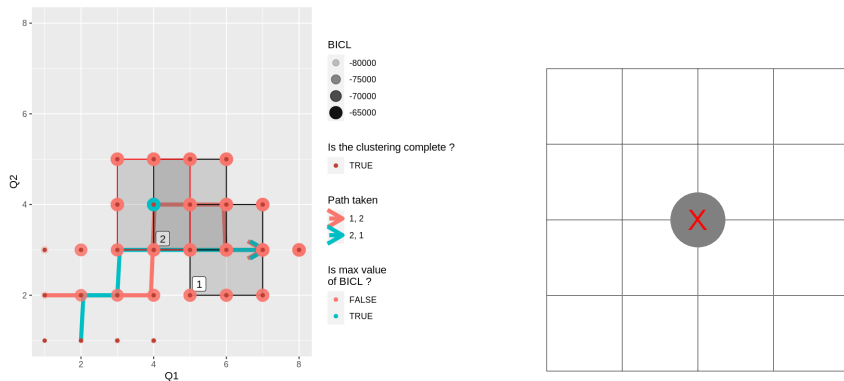


Figure 6 – Exemple de parcours de fenêtre glissante

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

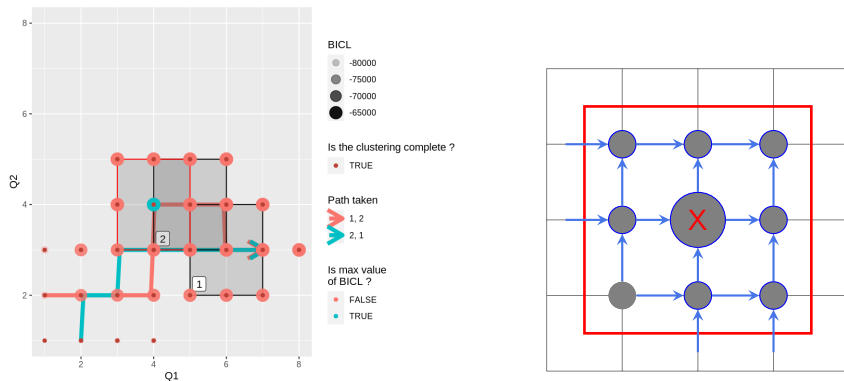


Figure 6 – Exemple de parcours de fenêtre glissante

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

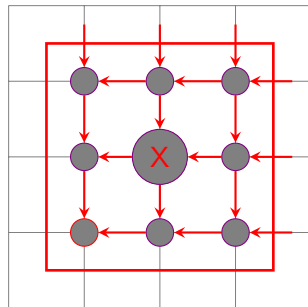
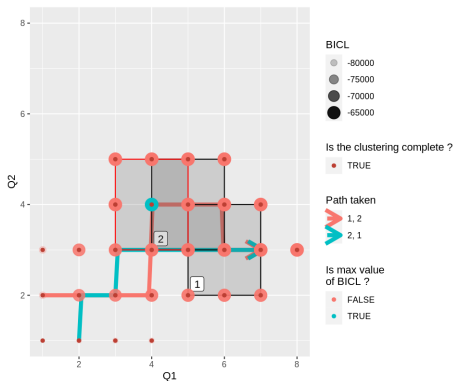
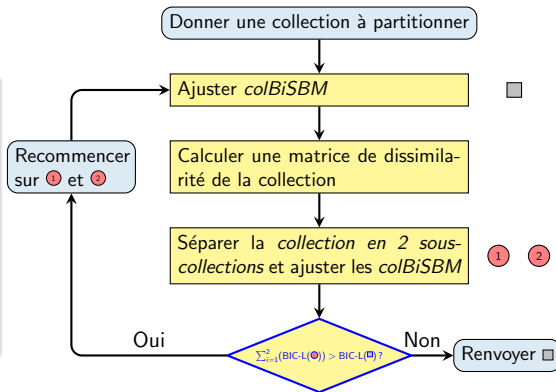


Figure 6 – Exemple de parcours de fenêtre glissante

# Clustering de réseaux

## Objectif

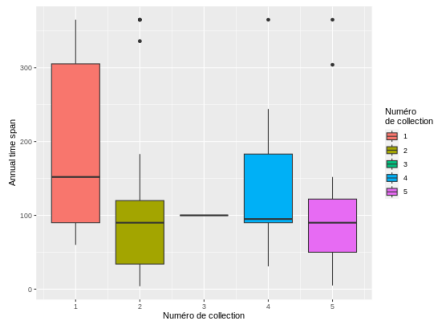
Déterminer une partition qui maximise la somme du BICL de ses sous-collections.



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

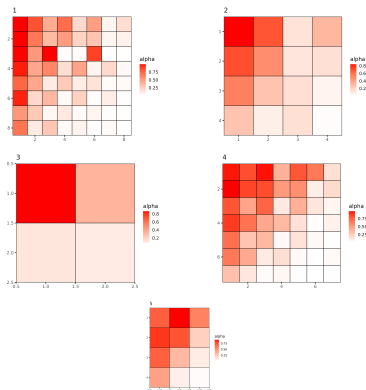



Figure 7 – Connectivités de la partition



- 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*<sup>5</sup>)

Le package est disponible sur GitHub :

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