

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

Parcours

Axes de
recherche

Axe 1

Axe 2

Axe 3

Organisation
de la thèse

Références

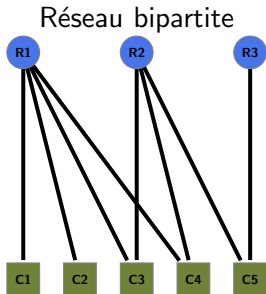
- 2024 Avril–Sept., Détection de structures et clustering de réseaux écologiques. Stage dans l'UMR MIA Paris-Saclay, supervisé par Pierre Barbillon.
- 2023 Janv.–Juillet, Détection de structures dans des collections de réseaux bipartites et écriture du package implémentant la méthode. Stage dans l'UMR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

Axes de recherche

- 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.

- 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

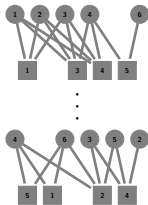


Matrice d'incidence

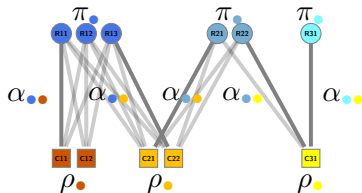
$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

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

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



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

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

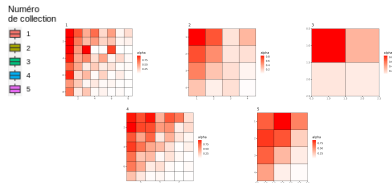
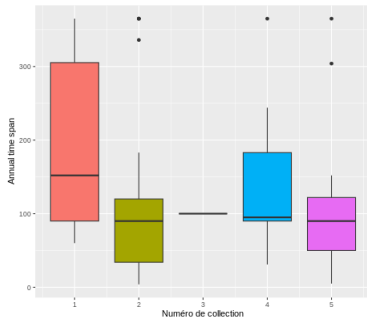



Figure – Connectivités de la partition

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

Apport déjà réalisé

- Adaptation du modèle mathématique^a proposé par [Chabert-Liddell et al., 2023](#) aux réseaux bipartites
- Développement algorithmique pour l'exploration de l'espace de paramètres.
- Écriture du code de la partie bipartite s'intégrant au package^b 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., 2023](#) ; [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

TODO Ici une illustration des GCN

En utilisant les *Graph Convolutional Networks* (GCN) il est possible de réaliser un *embedding* ou plongement 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.

Permettant derrière d'appliquer, par exemple, 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

Organisation de la thèse

Planning prévisionnel de la thèse

TODO Ici une timeline




- Axe 1 : Oct. 2024 – Oct. 2025
- Axe 2 : Oct. 2025 – Oct. 2026
- Axe 3 : Avril 2026 – Avril 2027
- Rédaction du manuscrit de thèse : Avril 2027 – Octobre 2027

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 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>




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>
-  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>



Bibliographie III

-  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 15 mai 2024, à partir de http://www.numdam.org/item/JSFS_2015__156_3_120_0/
-  Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018, février 4). *Graph Attention Networks*. arXiv : 1710.10903 [cs, stat]. Récupérée 14 mai 2024, à partir de <http://arxiv.org/abs/1710.10903>
-  Hamilton, W., Ying, Z., & Leskovec, J. (s. d.). Inductive Representation Learning on Large Graphs.

Bibliographie IV

-  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>

Bibliographie V

-  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>
-  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>

Annexes

Modèles à variables latentes pour collection de réseaux bipartites

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

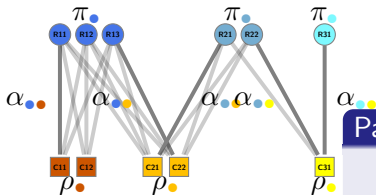
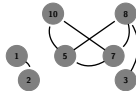
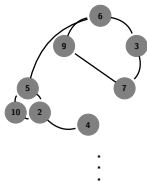


Figure – Exemple de LBM¹

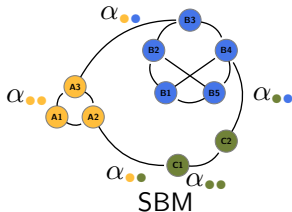
Paramètres

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



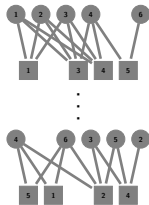
iid



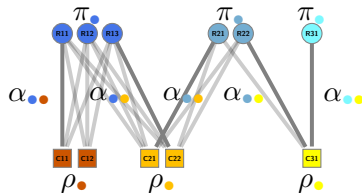
Pour $Q = |\{\bullet, \bullet, \bullet\}|$ blocs fixés :

Paramètres

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



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

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 paramètres²

$\pi\rho$ -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 $\forall q, m \in \llbracket 1, Q_1 \rrbracket \times \llbracket 1, M \rrbracket, \pi_q^m \in [0, 1]$ 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.

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

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

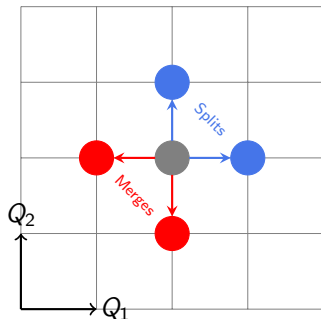
$$\begin{aligned} \ell(\mathbf{X}; \theta) \geq & \sum_{m=1}^M \left(\sum_{i=1}^{n_1^m} \sum_{j=1}^{n_2^m} \sum_{q \in \mathcal{Q}_{1,m}} \sum_{r \in \mathcal{Q}_{2,m}} \tau_{i,q}^{1,m} \tau_{j,r}^{2,m} \log f(X_{ij}^m; \alpha_{qr}) \right. \\ & + \sum_{i=1}^{n_1^m} \sum_{q \in \mathcal{Q}_{1,m}} \tau_{i,q}^{1,m}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

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

Le VEM se fait à Q_1, Q_2 fixés, il faut donc déterminer les “meilleures” coordonnées. Nous maximisons un BIC-L³.

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

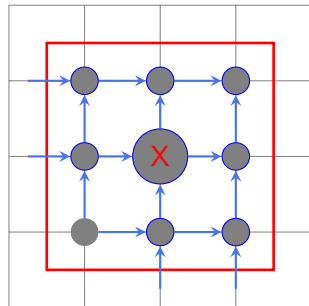
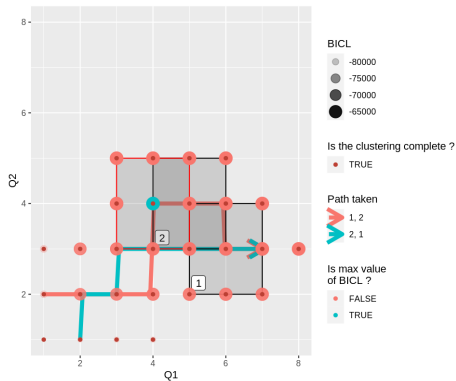


Figure – Exemple de parcours de fenêtre glissante

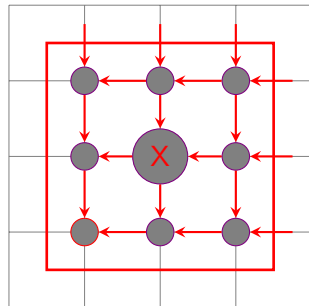
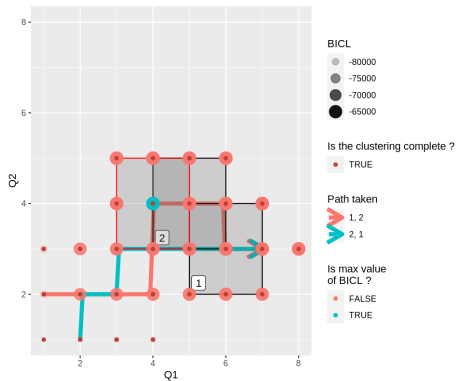
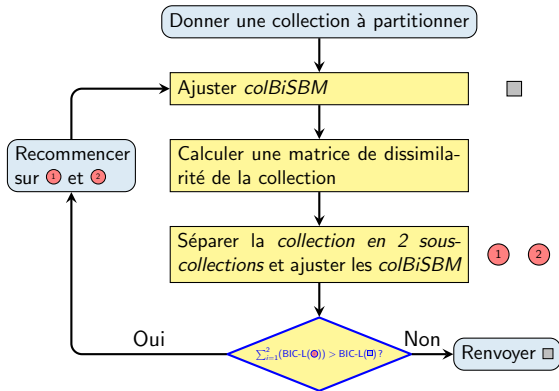


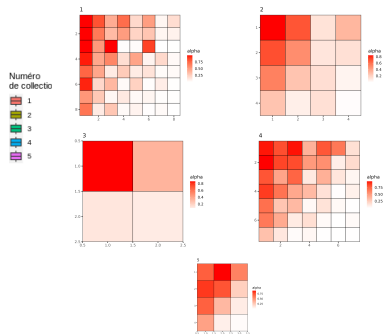
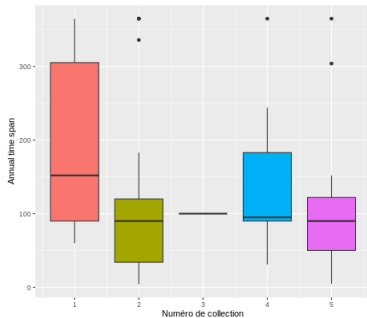
Figure – Exemple de parcours de fenêtre glissante

Objectif

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



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

- 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>