Statistical Learning of Collections of Networks with Applications in Ecology and Sociology

Saint-Clair Chabert-Liddell

Statistical learning of collections of networks with applications in ecology and sociology

Apprentissage statistique de collections de réseaux avec applications en écologie et en sociologie

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques, Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche UMR MIA Paris-Saclay (Université Paris-Saclay, AgroParisTech, INRAE), sous la direction de Sophie DONNET, chargée de recherche, la co-direction de Pierre BARBILLON, professeur

Thèse soutenue à Paris-Saclay, le 17 mars 2022, par

Saint-Clair CHABERT-LIDDELL

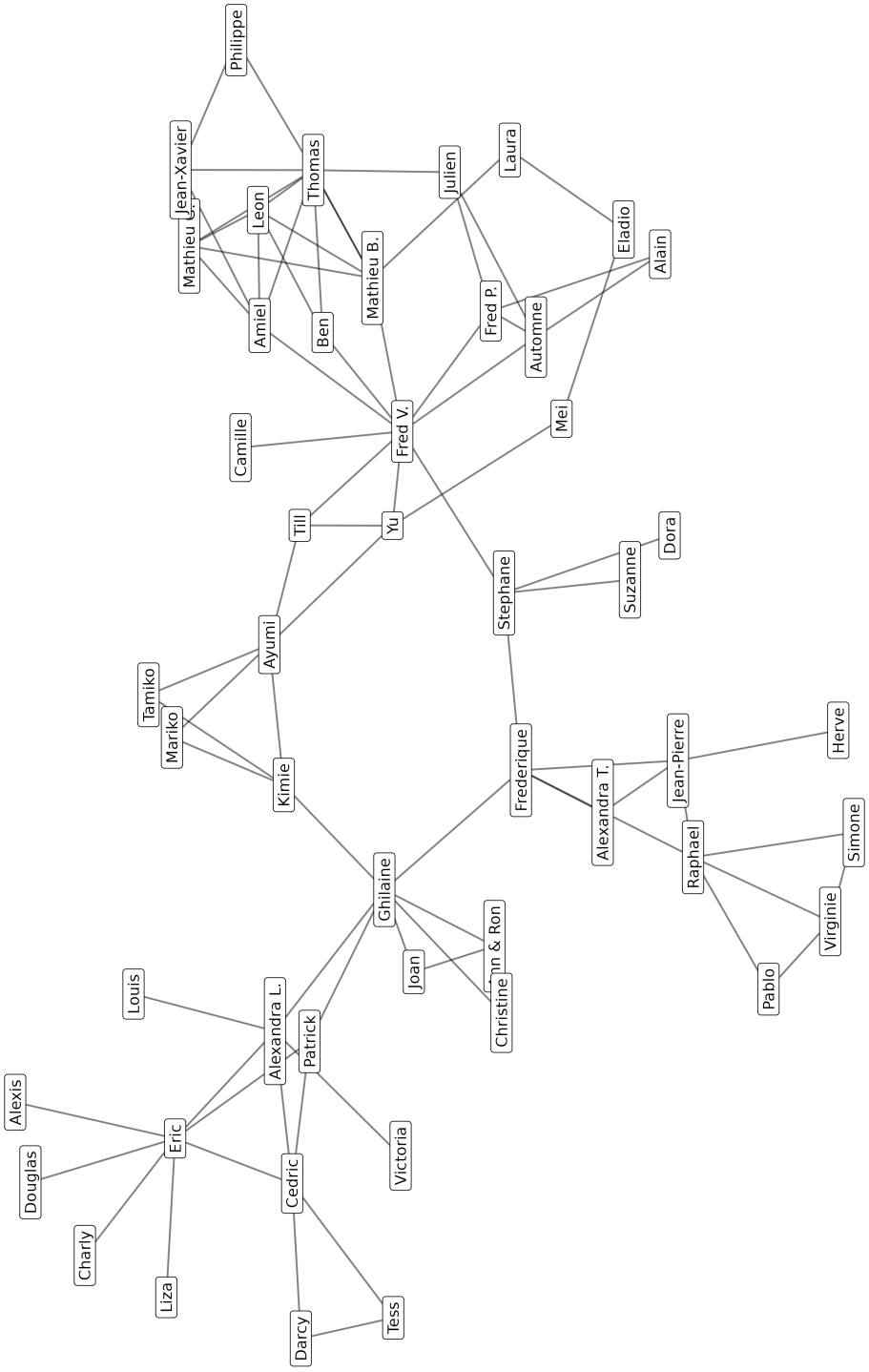
Composition du jury

Liliane Bel Professeure, AgroParisTech, Université Paris-Saclay	Présidente
Pierre Latouche Professeur, Université Paris Cité	Rapporteur & Examinateur
Nathalie Peyrard Directrice de Recherche, INRAE	Rapporteur & Examinatrice
Stéphane Dray Directeur de recherche, CNRS, Université Lyon 1	Examinateur
Tabea Rebafka Maîtresse de conférences, Sorbonne Université	Examinatrice
Aleš Žiberna Professeur, Université de Lubjana	Examinateur
Sophie Donnet Chargée de recherche, INRAE, Université Paris-Saclay	Directrice de thèse
Pierre Barbillon Professeur, AgroParisTech, Université Paris-Saclay	Directeur de thèse

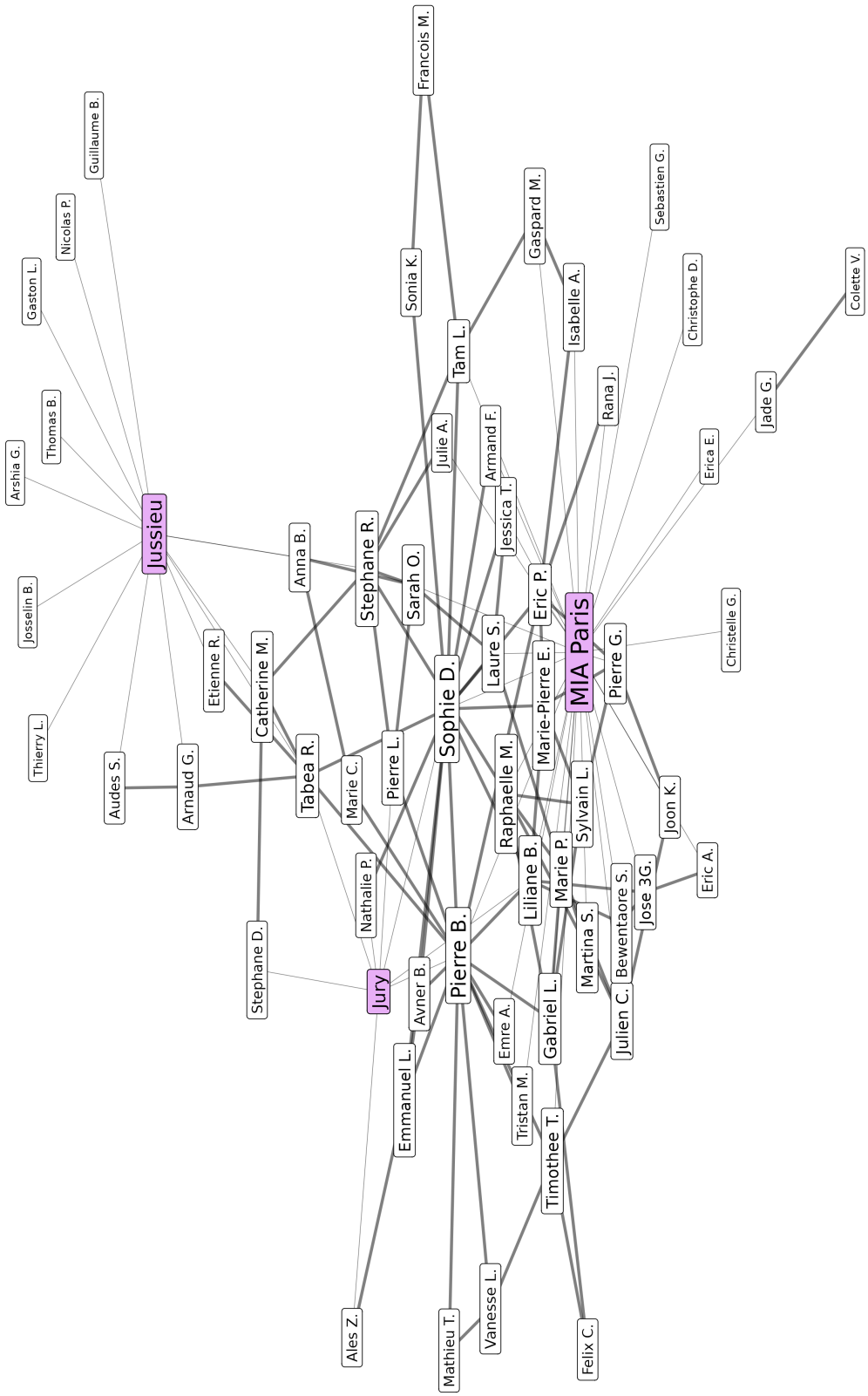
THESE DE DOCTORAT

NNT : 2022UPASM005

Remerciements



Ma famille et mes amis et leur famille



Le jury, mes collègues et camarades

Liste des productions scientifiques

Publications scientifiques

- **S-C. Chabert-Liddell**, P. Barbillon, S. Donnet et E. Lazega (2021). A Stochastic block model approach for the anlysis of multilevel networks. *Computational Statistics & Data Analysis*, 158:107179.
- **S-C. Chabert-Liddell**, P. Barbillon et S. Donnet (2022). Estimating the robustness of a bipartite ecological networks through a probabilistic modeling *Environmetrics*, 33(2), e2709.

Bibliothèques logicielles

- **Chabert-Liddell, S.-C.** (2021). *MLVSBM: A Stochastic Block Model for Multilevel Networks.* R package version 0.2.2. https://CRAN.R-project.org/package=MLVSBM
- **Chabert-Liddell, S.-C.** (2021). *robber: Using block model to estimate the robustness of ecological networks.* R package version 0.2.2. https://CRAN.R-project.org/package=robber

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Motivation

Un réseau est un moyen naturel de représenter un ensemble d'objets interconnectés. Ces objets sont les nœuds du réseau. Ceux-ci sont reliés entre eux par des arêtes signifiant leurs interactions. Selon le domaine d'application, les nœuds et les arêtes du réseau peuvent revêtir diverses significations. Dans cette thèse, je m'intéresse à développer des méthodes pour analyser des réseaux issus des sciences sociales et de l'écologie.

Dans un réseau social, les nœuds représentent des acteurs, en général des individus. L'observation des liens existants entre acteurs se fait par exemple suite aux réponses obtenues lors d'entrevues ou de questionnaires. Dans un réseau d'amitié, un acteur est un individu déclarant une relation d'amitié avec d'autres individus. Dans ce cas, la relation d'amitié est généralement réciproque et le réseau est dit non dirigé. Par contre, dans un réseau de relations de conseil, l'interaction signifie la demande de conseil qui est un lien dirigé d'un individu vers un autre.

Un réseau d'interactions écologiques représente les relations entre espèces dans un écosystème. Ces données peuvent-être le fruit d'observations de terrain ou de connaissances préexistantes dans la littérature. Il existe de nombreux types d'interactions dans un écosystème impliquant des espèces différentes. On peut par exemple s'intéresser aux relations de pollinisation, l'ensemble des espèces est divisé en deux groupes, celui des insectes pollinisateurs et celui des plantes. Ces espèces ne peuvent interagir qu'avec des espèces de l'autre groupe. La relation est mutualiste, c'est à dire qu'elle bénéficie aux deux espèces. Ce type de réseau est dit bipartite.

Un réseau est un système complexe dont l'analyse n'est pas aisée. Suivant les questions que l'on se pose nous allons vouloir analyser le réseau à différentes échelles. Localement, à l'échelle microscopique, on peut s'intéresser à détecter les nœuds les plus centraux d'un réseau ou bien détecter l'existence de certains motifs revêtant une signification particulière. On peut également regarder à l'échelle macroscopique les caractéristiques plus globales du réseau. Y a-t-il beaucoup de nœuds ? Quelle est la densité des interactions ? À quelle distance sont les nœuds les plus éloignés ? Enfin, on peut s'intéresser à résumer le réseau en trouvant des structures de tailles intermédiaires dîtes *structures méso-échelles* en regroupant des nœuds dont le profil d'interaction est similaire. Je m'intéresserai essentiellement à cette dernière approche dans ce manuscrit.

Un autre objet d'intérêt est l'évaluation de la résilience d'un système aux perturbations. Dans l'analyse de la *robustesse* d'un réseau écologique, on cherche à quantifier l'impact que la disparition d'espèces a sur l'écosystème. Par exemple, dans un réseau trophique, où les arêtes sont dirigées et représentent le flux d'énergie apporté par la consommation de ressources, la disparition d'une espèce impacte toute la chaîne trophique. Ce type d'analyse peut également être fait pour des réseaux bipartites tels que des réseaux mutualistes ou antagonistes (un réseau où l'interaction bénéficie à une espèce au détriment de l'autre, tel que le parasitisme ou l'herbivorie). L'analyse de ce phénomène passe par la compréhension du lien entre la structure du réseau et sa robustesse (Dunne et al., 2002).

Collection de réseaux et réseaux multicouches

Considérer des données provenant de plusieurs réseaux simultanément permet de répondre à de nouvelles questions. Ces réseaux peuvent concerner un même type d'interaction pris dans des contextes différents comme le temps, l'espace, le type des nœuds impliqués, etc... Ou bien concerner des réseaux représentant plusieurs types d'interactions mais dont une partie des nœuds est commun. Ceci ne doit pas être confondu avec un grand réseau où l'on aurait regroupé la totalité des nœuds et des interactions de manière indifférenciée. Pour l'analyse, des informations de différentes natures ne doivent pas être mises au même niveau et la signification des différentes interactions doit être préservée pour les différents réseaux d'une collection.

Étudier une collection de réseaux permet par exemple de se demander si des réseaux de relations de conseil impliquant des acteurs différents ont une structure particulière, spécifique à ce type de relations entre acteurs. Une autre application est de comprendre quelles sont les différences structurelles, si elles existent, entre des réseaux mutualistes et antagonistes en écologie (Michalska-Smith and Allesina, 2019).

Lorsque les réseaux d'une collection pourront s'agencer ensemble de manière à créer un système commun via des relations entre les nœuds des différents réseaux, je parlerai de réseaux multicouches. La popularité des réseaux multicouches a explosé récemment comme en témoigne la figure 1.1. En considérant qu'un système n'est pas composé d'un réseau isolé, mais d'une collection de réseaux en interactions. Ils représentent un moyen naturel pour étendre l'analyse des systèmes écologiques (Pilosof et al., 2017) et sociaux (Lazega and Snijders, 2015). Pour en comprendre l'intérêt, je donne ici, en terme très générique, quelques exemples caractéristiques de différents types de réseaux multicouches.

Les réseaux multiplexes permettent de représenter différents types d'interactions entre mêmes nœuds. Par exemple, des avocats peuvent échanger des conseils, développer des liens d'amitiés ou travailler ensemble (Lazega, 2001). Étudier conjointement ces différents types d'interactions, permet d'analyser plus finement les mécanismes sociaux en jeu entre les acteurs.

Dans un réseau dynamique, les couches représentent les interactions à un état

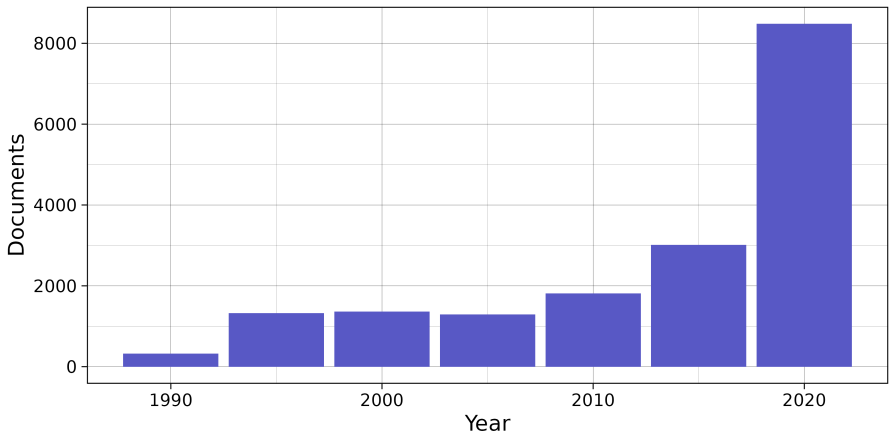


FIGURE 1.1 – Nombre de documents incluant les mots "multilayer networks" par période de 5 ans sur Google Scholar.

donné le long d'un gradient (temporel, spatial...). Les couches sont ordonnées et on s'intéresse alors à l'évolution de la structure du réseau à travers le temps ou l'espace.

Un réseau multipartite est composé de plusieurs types d'interactions, chacune d'elle impliquant un sous-ensemble de nœuds distincts. Dans un réseau en ethnobiologie, d'échange de semence, des individus possèdent des semences (ou graines) qu'ils sont susceptibles de s'échanger. Le réseau multicouche est alors constitué d'un réseau dirigé d'échange entre individus et d'un réseau bipartite individus-semences où les interactions représentent un lien de possession. On cherche alors à comprendre le lien entre possession de graines, interaction sociale et diversité des cultures (Labeyrie et al., 2016; Thomas and Caillon, 2016).

Considérons un réseau de pollinisation et un réseau de parasitisme partageant un même ensemble de plantes, ces deux réseaux peuvent être réunis sous la forme d'un réseau tripartite à deux couches interconnectées par les plantes. Dans ce nouveau réseau l'extinction d'espèce de pollinisateurs a un impact indirect sur les parasites à travers son impact sur les plantes. Les conclusions qualitatives de la robustesse des parasites aux extinctions de cet écosystème diffèrent suivant que l'on considère conjointement plusieurs réseaux sous la forme d'un réseau tripartite ou sous la forme de deux réseaux bipartites (Pocock et al., 2012; Pilosof et al., 2017).

Enfin, dans un réseau multiniveau, chaque couche est composée de nœuds différents interagissant entre eux et le lien entre les nœuds des différentes couches est hiérarchique. Un cas typique est celui des réseaux multiniveaux prenant en compte divers niveaux d'une organisation. Des échanges de conseils ont lieu au niveau inter-individuel (par exemple entre chercheurs), tandis que les organisations auxquelles ces individus sont affiliés s'échangent des ressources (par exemple entre laboratoires). On peut alors s'intéresser à l'influence mutuelle des différents types de relations (Lazega et al., 2008).

Cadre des travaux

Avant de présenter un état de l'art, je tiens à préciser trois choses. Bien que justifiées par des applications en écologie ou en sciences sociales, une grande partie des méthodes décrites ci-dessous est applicable à des réseaux issus d'autres domaines (biologie, information...). Dans le cadre de cette thèse les réseaux existants dans les applications qui nous intéressent sont de tailles moyennes, de l'ordre de dizaines à quelques centaines de nœuds et sont relativement denses. Ces réseaux sont observés, c'est-à-dire que l'interaction est directement collectée et ne demande pas à être inférée à partir d'autres types de données (données de coocurrences...).

1.2. État de l'art

1.2.1. Formulation mathématique de réseaux et réseaux multicouches

Un réseau simple (unipartite) est représenté mathématiquement par un graphe $G = (\mathcal{V}, \mathcal{E})$, où $\mathcal{V}$ est l'ensemble des nœuds (sommets) et $\mathcal{E}$ est l'ensemble des arêtes (liens). Dans les réseaux d'interaction qui nous intéressent, un nœud sera par exemple un individu ou une espèce. Les nœuds peuvent être étiquetés par le nom des espèces ou des individus et l'ensemble des labels est alors en bijection avec l'ensemble des $n = |\mathcal{V}|$ premiers entiers, où n est le nombre de nœuds du graphe :

$$\{\text{labels}\} \simeq \{1, \dots, n\}.$$

Une arête est une paire $(i, i') \in \mathcal{V} \times \mathcal{V}$, à laquelle on peut attribuer une valeur $w_{ii'}$ qui représente la force ou la fréquence de l'interaction entre les nœuds i et i'. Deux cas nous concernent particulièrement, lorsque l'on s'intéresse simplement à l'existence ou non d'une arête comme dans les réseaux binaires, où alors $w_{ii'} = 1$ pour tout $(i, i') \in \mathcal{E}$ et lorsque le nombre d'interactions entre deux nœuds importe, où l'on a alors $w_{ii'} \in \mathbb{N}^*$.

Un réseau simple $G = (\mathcal{V}, \mathcal{E})$ peut être représenté par sa matrice d'adjacence A, de taille $n \times n$ où pour tout $(i, i') \in \mathcal{V}^2$:

$$A_{ii'} = \begin{cases} w_{ii'} & \text{si les nœuds } i \text{ et } i' \text{ sont liés avec la valeur } w_{ii'} \\ 0 & \text{sinon.} \end{cases}$$

On nommera par dyade un couple de nœuds et on notera par $\mathcal{D}$ la restriction de $\mathcal{V}^2$ aux dyades pouvant être en interactions. Ainsi $\mathcal{E} \subset \mathcal{D} \subset \mathcal{V}^2$. Dans les réseaux que nous étudierons, les nœuds n'interagissent pas entre eux et il n'y a donc pas de lien entre i et i, ainsi $A_{ii} = 0$ pour tous $i \in \mathcal{V}$ et $\mathcal{D} = \{(i, i') \in \mathcal{V}^2, i \neq i'\}$. Certains réseaux peuvent également être non dirigés (réseau d'amitié...), alors $A_{ii'} = A_{i'i}$ et $\mathcal{D} = \{(i, i') \in \mathcal{V}^2 : i < i'\}$.

Réseaux bipartites Un réseau bipartite est un triplet $G_B = (\mathcal{V}_r, \mathcal{V}_c, \mathcal{E})$. Dans ce type de réseau, les nœuds sont divisés en deux ensembles disjoints $\mathcal{V} = \mathcal{V}_r \cup \mathcal{V}_c$, $\mathcal{V}_r \simeq \{1, \dots, n_r\}$, $\mathcal{V}_c \simeq \{1, \dots, n_c\}$ et $n_r + n_c = n$, où l'interaction n'est possible qu'entre nœuds de différents ensembles, i.e. $\mathcal{D} = \{(i,j) : i \in \mathcal{V}_r, j \in \mathcal{V}_r\}$. Ce type de réseau peut-être représenté par sa matrice d'incidence B de taille $n_r \times n_c$ où pour tout $(i,j) \in \mathcal{V}_r \times \mathcal{V}_c$:

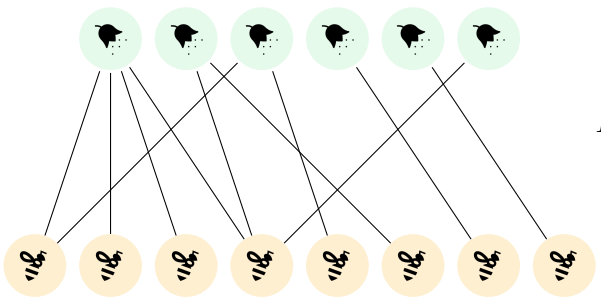
$$B_{ij} = \begin{cases} w_{ij} & \text{si les nœuds } i \text{ et } j \text{ sont liés avec la valeur } w_{ij}, \\ 0 & \text{sinon.} \end{cases}$$

Dans les réseaux d'interaction qui nous intéressent, soit l'interaction est réciproque (bien qu'elle puisse revêtir une signification différente) soit nous ne sommes concernés que par un seul sens de celle-ci. Par convention, nous posons que les interactions ont lieu de $\mathcal{V}_r$ vers $\mathcal{V}_c$ c'est-à-dire des lignes vers les colonnes de B.

Un réseau bipartite peut également être représenté par une matrice d'adjacence (symétrique par bloc) :

$$A = \begin{bmatrix} 0 & B \\ B^\intercal & 0 \end{bmatrix}.$$

Un réseau de pollinisation ainsi que sa matrice d'incidence sont représentés en figure 1.2.



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

FIGURE 1.2 – Représentation d'un réseau jouet de pollinisation et de sa matrice d'incidence.

Collection de réseaux simples Une collection de réseaux est un ensemble de M réseaux $\mathcal{C} = \{G^1, \dots G^M\}$, que nous représentons par un ensemble de M matrices $\mathbf{X} = \{X^1, \dots X^M\}$, chaque matrice X^m pouvant être une matrice d'adjacence ou d'incidence suivant le type du réseau considéré.

$$X^m = \begin{cases} A^m \in \mathbb{R}^{n_m \times n_m} & \text{si le réseau } m \text{ est un réseau unipartite} \\ B^m \in \mathbb{R}^{n_{m,r} \times n_{m,c}} & \text{si le réseau } m \text{ est un réseau bipartite.} \end{cases}$$

Réseaux multicouches

Il existe plusieurs formalismes pour définir des réseaux multicouches (Kivelä et al., 2014). J'adapterai dans ce qui suit celui de Bianconi (2018). Un réseau multicouche à M couches est un triplet $\mathcal{G} = (\mathcal{M}, \mathbf{G}, \mathbf{B})$, où $\mathcal{M}$ est l'ensemble des couches,

$$\mathcal{M} \simeq \{1, \dots, M\},$$

$\mathbf{G} = (G_1, \dots, G_M)$ est un M-uplet de réseaux unipartites représentant les interactions à l'intérieur d'une même couche :

$$G_m = (\mathcal{V}_m, \mathcal{E}_m) \quad \forall m \in \mathcal{M},$$

où $\mathcal{V}_m \simeq \{1, \dots, n_m\}$ et $\mathcal{E}_m \subset \mathcal{D}_m \subset \mathcal{V}_m^2$ et $\mathbf{B}$ est un $M(M-1)$-uplet dont les éléments sont des réseaux bipartites représentant les relations inter-couches :

$$G_{B,m,m'} = (\mathcal{V}_m, \mathcal{V}_{m'}, \mathcal{E}_{m,m'}) \quad \forall (m \neq m') \in \mathcal{M}^2.$$

Comme chaque réseau peut être représenté par sa matrice d'adjacence dans le cas d'un réseau unipartite et sa matrice d'incidence pour un réseau bipartite, dans la plupart des travaux, nous considérerons un réseau multicouche comme une collection de matrices :

$$\mathbf{X} = \left((A^m)_{m \in \mathcal{M}}, (B^{m,m'})_{(m \neq m') \in \mathcal{M}^2} \right),$$

qui peut à son tour être aplatie pour définir la matrice de supra-adjacence suivante :

$$\mathcal{A} = \begin{bmatrix} A^1 & B^{1,2} & \cdots & B^{1,M} \\ B^{1,2} & A^2 & \cdots & B^{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ B^{1,M} & B^{2,M} & \cdots & A^M \end{bmatrix}.$$

Parfois dénommés "réseaux liés", "réseaux multipartites généralisés" ou "réseaux de réseaux", il s'agit du type le plus général de réseaux multicouches qui concerne à la fois les interactions inter-couches et intra-couches. En pratique, la plupart des réseaux multicouches n'impliquent qu'un sous-ensemble des réseaux possibles $\mathbf{G}$ et $\mathbf{B}$.

Notons que les réseaux unipartites et bipartites sont des réseaux multicouches selon cette définition. En utilisant différentes contraintes, nous pouvons alors définir dans ce cadre certains types de réseaux multicouches d'un intérêt particulier en écologie et en sociologie.

Réseaux multipartites Un réseau multipartite est une généralisation des réseaux bipartites. Les nœuds du réseau sont partitionnés en type et les nœuds n'interagissent pas avec leurs propres types. En d'autres termes, c'est un réseau multicouche sans interaction intra-couche. Souvent appelés réseaux multimodaux dans la littérature sur les réseaux sociaux, les réseaux bipartites – à 2-modes – en sont un cas particulier.

Ainsi un réseau K-partite est défini par

$$\mathcal{G} = (\{1, \dots, K\}, \emptyset, \mathbf{B}),$$

qui peut être représenté par la matrice supra-adjacence suivante :

$$\mathcal{A} = \begin{bmatrix} 0 & B^{1,2} & \cdots & B^{1,K} \\ B^{2,1^\intercal} & 0 & \cdots & B^{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ B^{K,1^\intercal} & B^{K,2^\intercal} & \cdots & 0 \end{bmatrix}.$$

Généralement, seulement une partie des interactions inter-couches est possible (ou est intéressante) et $|(k \neq k') : G_{B,k,k'} \neq \emptyset| \leq K(K-1)$. Un exemple typique issu des réseaux écologiques est celui des réseaux tripartites où seulement un ensemble d'espèces est en interaction avec les deux autres, comme les réseaux plantes-pollinisateurs-oiseaux (mutualiste-mutualiste) où les pollinisateurs et les oiseaux n'interagissent qu'avec les plantes ou les réseaux plantes-herbivores-parasites (antagoniste-antagoniste) où les parasites et les plantes ne sont pas en interactions.

Réseaux multiplexes Un réseau multiplexe est un réseau multicouche utilisé pour décrire plusieurs types d'interactions entre un même ensemble de nœuds. En sociologie, des individus peuvent être amis ou échanger des conseils tandis qu'en écologie des espèces peuvent être en compétition pour des ressources tout en coopérant pour d'autres taches.

Dans un réseau multiplexe, les nœuds sont des nœuds répliqués entre les couches et sont justes concernés par les interactions intra-couches :

$$\mathcal{G} = (\mathcal{M}, \mathbf{G}, \emptyset),$$

où $\mathcal{V}_m = \mathcal{V}_{m'}$ pour tout $(m, m') \in \mathcal{M}^2$.

La relation entre les nœuds des différentes couches peut être explicitée, donnant le réseau multicouche $\mathcal{G} = (\mathcal{M}, \mathbf{G}, \mathbf{B})$ où les arêtes des réseaux de $\mathbf{B}$ sont trivialement celles que relient les nœuds ayant le même label. Cela peut être représenté par la matrice de supra-adjacence suivante :

$$\mathcal{A} = \begin{bmatrix} B^1 & \mathbf{I}_n & \cdots & \mathbf{I}_n \\ \mathbf{I}_n & B^2 & \cdots & \mathbf{I}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{I}_n & \mathbf{I}_n & \cdots & B^M \end{bmatrix},$$

où $\mathbf{I}_n$ est la matrice identité de taille $n \times n$.

Notons que dans cette définition les interactions inter-couches sont redondantes et qu'il est possible de se ramener à un réseau simple dont les interactions sont M-dimensionnel de matrice d'adjacence :

$$A_{ij} = (a_{ij}(1), \dots, a_{ij}(M)) \quad \forall (i \neq j) \in \mathcal{V}^2.$$

Réseaux multislices (multicoupes) Un réseau multislice (multicoupe) est utilisé pour représenter des interactions impliquant un même ensemble de nœuds à plusieurs points ordonnés (par exemple le long d'une ligne temporelle ou spatiale). C'est un réseau multicouche où chaque couche représente le réseau d'interaction à un point donné et où les interactions entre les couches identifient les nœuds communs entre points adjacents. C'est donc un réseau multicouche $\mathcal{G} = (\mathcal{M}, \mathbf{G}, \mathbf{B})$ avec les contraintes suivantes :

- $\mathcal{V}_m = \mathcal{V}_{m'}, \qquad \forall (m, m') \in \mathcal{M}^2$
- $\mathcal{E}_{m,m'} = \{(i, i') \in \mathcal{V}_m \times \mathcal{V}_{m'} : i = i'\}, \qquad \forall (m, m') \in \mathcal{M}^2 : m' = m + 1$

et avec une matrice de supra-adjacence de la forme :

$$\mathcal{A} = \begin{bmatrix} A^1 & \mathbf{I}_n & 0 & \cdots & \cdots & 0 \\ 0 & A^2 & \mathbf{I}_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & A^{M-1} & \mathbf{I}_n \\ 0 & \cdots & \cdots & \cdots & 0 & A^M \end{bmatrix}.$$

On peut créer un réseau multislice à partir d'interactions continues dans le temps en segmentant la ligne temporelle en intervalle et en définissant les intersections d'une couche comme celles ayant eu lieu durant un intervalle de temps donné.

Réseaux multiniveaux Les réseaux multiniveaux sont des réseaux multicouches avec une relation hiérarchique entre les couches et où les interactions inter-couches sont possibles uniquement entre couches adjacentes dans la hiérarchie. Dans ce cadre, ils peuvent être vus comme des réseaux multislices sans l'hypothèse de nœuds communs entre les couches et peuvent donc être également utilisés pour représenter des réseaux temporels impliquant des nœuds différents à travers le temps. C'est un réseau multicouche $\mathcal{G} = (\mathcal{M}, \mathbf{G}, \mathbf{B})$ avec les contraintes suivantes :

$$\mathbf{B} = (G_{B,m,m'} : m' = m + 1 \quad \forall m \in \{1, \ldots, M - 1\}),$$

qui peut-être représenté par la matrice d'adjacence :

$$\mathcal{A} = \begin{bmatrix} A^1 & B^{1,2} & 0 & \cdots & \cdots & 0 \\ 0 & A^2 & B^{2,3} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & A^{M-1} & B^{M-1,M} \\ 0 & \cdots & \cdots & \cdots & 0 & A^M \end{bmatrix}.$$

1.2.2. Quelques modèles probabilistes pour graphes aléatoires

Modèle d'Erdős-Rènyi

Erdős and Rényi (1960) ont proposé le premier modèle classique pour les graphes aléatoires binaires non dirigés, le modèle est paramétré par le nombre de nœuds du graphe et son nombre d'arêtes. Un réseau $G = (\mathcal{V}, \mathcal{E})$ à n nœuds et m arêtes est tiré uniformément parmi tous les réseaux possibles ayant le même nombre de nœuds et d'arêtes :

$$\mathbb{P}(G) = \binom{\binom{n}{2}}{m}^{-1}.$$

À la même époque, Gilbert (1959) propose une autre formulation où l'existence de chaque arête est indépendante et identiquement distribuée (*i.i.d.*) :

$$\mathbb{P}(X_{ij} = 1) = p \quad p \in (0, 1), \quad (i, j) \in \mathcal{V}^2, i < j.$$

Le premier modèle est souvent noté par $G(n, m)$ où le nombre d'arêtes est fixé, tandis que le second l'est par $G(n, p)$ où le nombre d'arête est donné en espérance. Par un abus de langage, on nommera par modèle Erdős-Rènyi (ER) n'importe quel modèle de graphe aléatoire où l'existence d'une arête est équiprobable et indépendante des autres arêtes.

Modèles exponentiels de graphes aléatoires (ERGM)

L'ERGM, aussi nommé p^* dans la littérature en réseaux sociaux (Wasserman and Pattison, 1996) s'inspire des modèles de régression statistique pour modéliser des graphes aléatoires. L'idée est d'écrire la loi d'un graphe aléatoire comme celle d'une loi de la famille exponentielle :

$$\mathbb{P}_\theta(X = x) = \frac{1}{C(\theta)} \exp(\theta^\intercal S(x)),$$

où $C(\theta) = \sum_x \exp(\theta^\intercal S(x))$ est une constante de normalisation et $S(x)$ est un vecteur de statistiques exhaustives du modèle. Par conséquent, tous les graphes ayant la même valeur observée de S ont la même probabilité selon le modèle prédéfini. Ce modèle est largement utilisé dans les sciences sociales, en utilisant $S(x)$ comme un comptage de différents motifs locaux (triangles, k-étoiles, k-triangles...) qui revêtent des significations sociologiques (Robins et al., 2009). Cette approche s'étend aux réseaux multicouches : temporels (Hanneke et al., 2010), multiniveaux (Wang et al., 2013)... Les ERGMs s'accompagnent d'importantes limites théoriques et pratiques : le problème de la maximisation de la vraisemblance est mal posé et les modèles sont souvent dégénérés car la distribution se concentre autour d'un mélange de graphes vides et complets (Chatterjee and Diaconis, 2013).

Modèle Expected Degree Distribution (EDD)

On peut définir un modèle de graphe aléatoire en s'intéressant à la distribution des degrés des nœuds. Dans le modèle EDD (Chung and Lu, 2002), on définit une suite de degrés attendus $\mathbf{D} = (D_1, \ldots, D_n)$, et les arêtes du graphe aléatoire $G \in G(\mathbf{D})$ sont tirées indépendamment avec une probabilité qui est proportionnelle aux degrés attendus :

$$p_{ij} = \frac{D_i D_j}{C} \in (0,1) \quad C \in \mathbb{R}_+^* \quad \forall (i,j) \in \mathcal{V}^2.$$

De même que le modèle d'Erdős-Rènyi a sa version $G(n,m)$ à nombre d'arêtes fixées, dans le modèle de configuration (Bollobás, 1980), le degré de chaque nœud est fixé. Le graphe est tiré uniformément parmi tous les graphes respectant cette suite de degré (si celle-ci est graphique). Le chapitre 12 du livre de Newman (2018) est consacré à ces deux modèles et à certaines de leurs propriétés.

1.2.3. Modèles à espaces latents pour graphes aléatoires

Les modèles à espaces latents supposent l'existence d'une variable aléatoire latente dont la valeur caractérise la distribution de l'observation. Dans le cas des graphes aléatoires, nous considérerons en général que les variables latentes vont caractériser le comportement des nœuds, apportant de l'hétérogénéité dans les connexions du réseau. Les variables latentes sont notées $\mathbf{Z} = (Z_i, i \in \mathcal{V})$ et permettent d'obtenir un clustering des nœuds du réseau. Les observations $\mathbf{X} = (X_{ij})_{(i,j)\in\mathcal{D}}$ sont indépendantes conditionnellement à ces variables latentes. L'espace des variables latentes peut être continu, par exemple à valeur dans un *espace social* dont les distances entre individus définissent leur probabilité d'interaction (Hoff et al., 2002; Handcock et al., 2007). Le modèle EDD présenté juste au-dessus, peut également être vu comme un modèle à espaces latents où les variables latentes des nœuds seraient les degrés attendus. Dans ce qui suit, je vais m'intéresser aux modèles à espaces latents discrets, en présentant en détail le modèle à blocs stochastiques et certaines de ses extensions. Pour plus de détails sur les modèles à espaces latents pour graphes aléatoires, j'oriente vers la revue relative à ce sujet de Matias and Robin (2014).

Modèle à blocs stochastiques (SBM) et extensions

Le modèle à blocs stochastiques (Holland et al., 1983) permet de modéliser de l'hétérogénéité dans les structures de connexion du réseau. Les nœuds appartiennent à des blocs latents qui façonnent leur profil de connectivité.

Soit $\mathcal{Q} = \{1, \ldots, Q\}$ un ensemble de blocs latents et Z_i la variable latente d'appartenance au bloc du nœud i tel que $Z_i = q$ si et seulement si le nœud i appartient au bloc q. Alors pour tout $i \in \mathcal{V}$, les Z_i sont des variables aléatoires *i.i.d.* tel que pour tout $q \in \mathcal{Q}$:

$$\mathbb{P}(Z_i = q) = \pi_q, \quad \sum_{q\in\mathcal{Q}} \pi_q = 1,$$

où $\pi = (\pi_1, \ldots, \pi_Q)$ est le paramètre de mélange. On considèrera également la matrice des appartenances aux blocs $Z \in \{0,1\}^{n \times Q}$ tel que

$$Z_{iq} = \begin{cases} 1 & \text{si } Z_i = q \\ 0 & \text{sinon.} \end{cases}$$

Alors, connaissant les blocs d'appartenances, les interactions sont indépendantes de loi :

$$X_{ij} | Z_i, Z_j \sim f(\cdot; \alpha_{Z_i Z_j}),$$

où f est la loi d'émission dépendant du type de valeur de l'arête et $\alpha = (\alpha_{qr})_{(q,r) \in \mathcal{Q}^2}$ est la matrice des paramètres de connexion.

En supposant une équivalence stochastique entre les nœuds, la structure modélisée par un SBM est plus universelle que les structures de communautés purement assortatives qui supposent que les probabilités de connexion à l'intérieur d'un bloc sont plus élevées que les probabilités de connexion entre blocs, i.e. $\alpha_{qq} > \alpha_{qr}$ pour tout $(q \neq r) \in \mathcal{Q}^2$. L'identifiabilité du modèle a été prouvée par Celisse et al. (2012) pour les réseaux binaires.

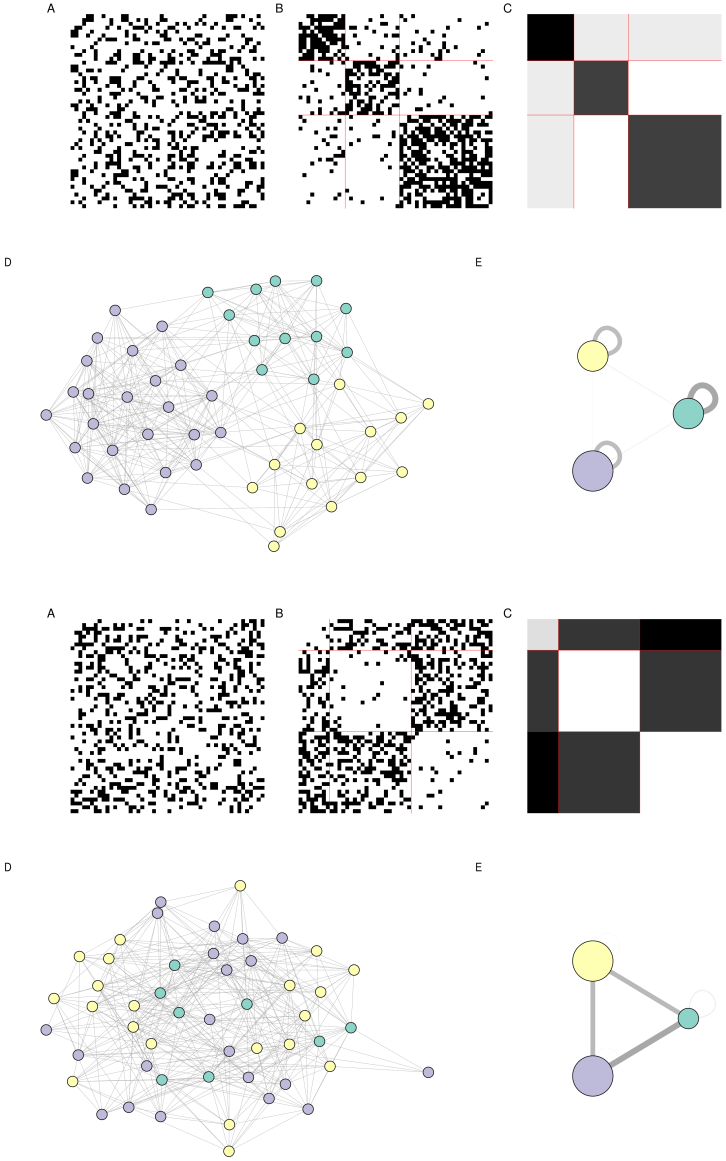
Le SBM est un outil performant pour la visualisation de graphes, car il résume l'information en quelques paramètres. En effet, la structure méso-échelle d'un réseau peut être décrite par un graphe dont les nœuds représentent les blocs, valués par la cardinalité du bloc, et les arêtes la probabilité ou le nombre attendu d'arêtes entre les blocs. Ce fait est astucieusement utilisé par Peixoto (2014b) pour proposer un SBM imbriqué à différents niveaux de granularité. Les paramètres peuvent également être résumés sous forme matricielle, tel un graphon constant par bloc, modèle que je présente à la fin de la section 1.2.3.

La figure 1.3 illustre les différents types de visualisation possible sur trois topologies classiques de réseaux : communautés assortatives où $\alpha_{qq} > \alpha_{qr}$ pour tout $(q \neq r) \in \mathcal{Q}^2$, communautés disassortatives où $\alpha_{qq} < \alpha_{qr}$ pour tout $(q \neq r) \in \mathcal{Q}^2$ et cœur-périphérie où un bloc du réseau (le cœur) est fortement connecté au reste du réseau alors que la périphérie est faiblement connectée à elle même.

De nombreuses variantes ont été développées à partir du SBM. Je présente ici celles que j'estime être les plus importantes.

Modèle à blocs latents (LBM) Le LBM (ou SBM bipartite) (Govaert and Nadif, 2003) sert à modéliser des réseaux bipartites. Les blocs latents sont divisés en deux ensembles disjoints $\mathcal{Q}_r$ et $\mathcal{Q}_c$, un pour chaque type de nœuds, de variables latentes respectives $\mathbf{Z}$ et $\mathbf{W}$ indépendantes. Les $(Z_i)_{i \in \mathcal{V}_r}$ sont *i.i.d.*, de même que les $(W_j)_{j \in \mathcal{V}_c}$:

$$\mathbb{P}(Z_i = q) = \pi_q, \quad \sum_{q \in \mathcal{Q}_r} \pi_q = 1,$$
$$\mathbb{P}(W_j = r) = \rho_r, \quad \sum_{r \in \mathcal{Q}_c} \rho_r = 1.$$



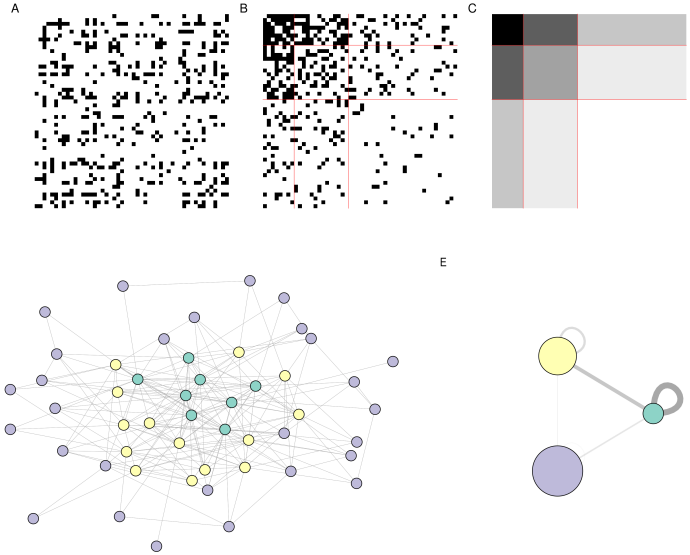


FIGURE 1.3 – Haut-gauche – Structure communautaire assortative : $\alpha^a = \begin{pmatrix} .8 & .1 & .1 \\ \cdot & .6 & .05 \\ \cdot & \cdot & .6 \end{pmatrix}$ and $\pi^a = (.3, .3, .4)$, Haut-droite – Structure cœur-périphérie avec un semi-cœur : $\alpha^{cp} = \begin{pmatrix} .8 & .5 & .2 \\ \cdot & .3 & .1 \\ \cdot & \cdot & .05 \end{pmatrix}$ and $\pi^{cp} = (.2, .3, .5)$, Bas-gauche – Structure communautaire disassortative : $\alpha^d = \begin{pmatrix} .1 & .4 & .5 \\ \cdot & .05 & .4 \\ \cdot & \cdot & .05 \end{pmatrix}$ and $\pi^d = (.2, .4, .4)$
A : Matrice d'un réseau de 50 nœuds généré aléatoirement d'après la structure donnée, **B** : La même matrice réordonnée par appartenance aux blocs, **C** : Vue de type graphon des paramètres du modèle, **D** : Réseau généré, la couleur des nœuds correspond aux blocs d'appartenances, **E** : Vue méso-échelle des paramètres du modèle, la taille des nœuds représente la cardinalité du bloc (paramètre π) et la largeur et l'intensité des arêtes désignent la probabilité de connexion (paramètre α).

Ensuite, sachant les blocs latents, les arêtes sont indépendantes de loi :

$$X_{ij}|Z_i, W_j \sim f(\cdot; \alpha_{Z_i W_j}).$$

La version Bernoulli du LBM est identifiable (Keribin et al., 2015).

Modèle avec covariables ou annotations Bien que le SBM soit habituellement utilisé pour les réseaux binaires, il peut être facilement étendu aux arêtes valuées. Mariadassou et al. (2010) montrent cela pour certaines lois d'émissions standards sur les arêtes, mais ils considèrent également un modèle avec des covariables sur les arêtes ajoutant un cadre de régression au SBM. En considérant des réseaux où les arêtes valuées sont modélisées par des variables aléatoires indépendantes de loi de Poisson, ils traitent deux cas, l'un où l'effet des covariables est homogène sur les arêtes :

$$X_{ij}|Z_i, Z_j \sim Pois(\alpha_{Z_i Z_j} e^{\beta^\intercal Y_{ij}}),$$

et l'autre où il dépend des blocs (effet inhomogène) :

$$X_{ij}|Z_i, Z_j \sim Pois(\alpha_{Z_i Z_j} e^{\beta_{ql}^\intercal Y_{ij}}),$$

où Y_{ij} est le vecteur des covariables β est le paramètre de régression du modèle linéaire généralisé.

Cette approche peut être adaptée aux réseaux binaires en changeant la fonction de lien pour celle d'une régression logistique, on a alors pour le modèle à effet inhomogène :

$$\mathbb{P}(X_{ij} = 1|Z_i, Z_j) = \frac{1}{1 + e^{-\alpha_{Z_i Z_j} - \beta^\intercal Y_{ij}}}. \tag{1.1}$$

Lorsque l'ensemble des covariables sur les nœuds est fini, celles-ci peuvent être transférées au niveau des arêtes. La structure de blocs retrouvée par ce modèle est la structure résiduelle, celle qui n'a pas été expliquée une fois que l'effet des covariables a été pris en compte.

Newman and Clauset (2016) adoptent une approche différente lorsqu'ils traitent des covariables sur les nœuds (métadonnées). Les covariables servent alors à définir des probabilités a priori pour les blocs latents. Lorsque le nombre de covariables est fini, discret et non ordonné, en notant Y_i la valeur de la covariable du nœud i, on a :

$$\mathbb{P}(Z_i = q) = \gamma_{qY_i}, \quad \sum_{q \in \mathcal{Q}} \gamma_{qY_i} = 1.$$

Cette approche est également étendue à des covariables continues et ordonnées.

Dans une troisième approche, Peel et al. (2017) essaient de relier la structure de covariables discrètes et finies sur les nœuds avec la structure par bloc du SBM en fixant un ensemble de nœuds dont les blocs sont entièrement définis par leurs covariables.

Degree Corrected Stochastic Block Model (DCSBM) Dans le SBM, les degrés des nœuds d'un bloc sont distribués suivant une loi de Poisson, alors que de nombreux réseaux réels observés ont une suite des degrés qui suit une loi de puissance ou une loi exponentielle. Pour corriger ce problème, Karrer and Newman (2011) ajoutent un paramètre de degré à chaque nœud. De manière à faciliter les calculs, la loi d'émission est une loi de Poisson, ce qui est raisonnable même pour un réseau binaire tant que le réseau est suffisamment creux. Soit η_i, $i \in \mathcal{V}$ un paramètre spécifique à chaque nœud avec pour des raisons d'identifiabilité $\sum_{i \in q} \eta_i = 1$, pour tout $q \in \mathcal{Q}$, alors conditionnellement aux blocs, les arêtes sont indépendantes et de loi :

$$X_{ij} | Z_i, Z_j \sim Pois(\eta_i \alpha_{Z_i Z_j} \eta_j). \tag{1.2}$$

Le DCSBM est particulièrement utile pour modéliser des structures de communautés assortatives avec une hétérogénéité des degrés à l'intérieur d'une communauté. Ceci a un coût puisque l'on perd l'équivalence stochastique entre les nœuds et en interprétabilité des paramètres du modèle.

Modèles avec partition souple Le SBM a aussi été étendu pour prendre en compte des partitions souples. Dans le mixed membership stochastic block model (Airoldi et al., 2008, MMSBM), on donne à chaque nœud $i \in \mathcal{V}$ un vecteur de probabilité d'appartenance π_i de longueur Q (issu d'une loi a priori de Dirichlet dans un cadre bayésien), alors l'appartenance des blocs est spécifique à chaque dyade. Pour un réseau dirigé, soit $Z_{i \to j}$ le bloc utilisé par le nœud i pour envoyer une arête vers le nœud j et $Z_{j \leftarrow i}$ le bloc utilisé par le nœud j pour recevoir une arête du nœud i, deux variables latentes indépendantes. Alors la loi conditionnelle des interactions est donnée par :

$$X_{ij} | Z_{i \to j}, Z_{j \leftarrow i} \sim Bern(\alpha_{Z_{i \to j}, Z_{j \leftarrow i}}).$$

Dans le overlapping stochastic block model (Latouche et al., 2011, OSBM), chaque nœud $i \in \mathcal{V}$ a un vecteur aléatoire latent d'appartenance $Z_i \in \{0,1\}^Q$ dont les entrées sont indépendantes et suivent une loi de Bernoulli : $Z_i \overset{ind}{\sim} \prod_{q \in \mathcal{Q}} Bern(\pi_q)$. Ainsi, un nœud peut aussi bien appartenir à plusieurs blocs qu'à aucun, ce qui est utile pour modéliser des "outliers". Alors la loi conditionnelle des arêtes sachant les blocs est donnée par :

$$\mathbb{P}(X_{ij} = 1 | Z_i, Z_j) \sim Bern(Z_i^\intercal Z_j + Z_i^\intercal U + Z_j^\intercal V + W^*),$$

où W est une matrice $Q \times Q$ modélisant les interactions entre blocs, U et V sont des vecteurs de dimension Q modélisant un effet envoyeur et receveur, tandis que W^* est un scalaire représentant un biais pour modéliser la parcimonie.

Modèles dynamiques Certains travaux s'intéressent à des réseaux dynamiques, dont les interactions entre nœuds peuvent se répéter au cours du temps. La loi des arêtes est alors un processus stochastique. Le SBM peut être utilisé pour modéliser ce genre de données. Dans ce cas, les interactions sont indépendantes et suivent un processus de comptage (Poisson inhomogène) dont le paramètre d'intensité dépend des blocs des nœuds concernés (DuBois et al., 2013; Matias et al., 2018).

Formulation microcanonique De même qu'il existe une formulation $G(n, m)$ en plus d'une formulation $G(n, p)$ pour le modèle Erdős-Rényi ou le modèle EDD avec le modèle de configuration, Peixoto (2016) propose une formulation microcanonique du SBM Poisson et sa version avec degré corrigé. Pour cette dernière, il fixe des contraintes dures sur le degré de chaque nœud en plus du nombre d'arêtes entre chaque bloc et du nombre de nœud dans chaque bloc.

Graphon et limite de graphes aléatoires

Le modèle W-graphe représente un modèle de graphe aléatoire échangeable avec le système génératif suivant :

$$\begin{aligned} U_i &\sim Unif([0,1]) \quad \forall \in \mathcal{V} \\ X_{ij} = 1 | U_i, U_j &\sim Bern(W(U_i, U_j)) \quad \forall (i,j) \in \mathcal{V}^2. \end{aligned}$$

Il est caractérisé par le graphon, une fonction symétrique mesurable $W : [0,1]^2 \to [0,1]$. Les nœuds y possèdent des positions latentes continues qui déterminent leurs probabilités de connexion. Ce modèle définit la limite de la matrice d'adjacence d'un réseau dense (Lovász and Szegedy, 2006). Le modèle n'est pas identifiable, mais comme montré par Bickel and Chen (2009), ce problème peut être résolu en considérant $g : u \mapsto \int_0^1 W(u,v)\mathrm{d}v$ comme une fonction monotone croissante.

On peut relier le SBM à un graphon constant par morceau en considérant

$$W(u,v) = \alpha_{C(u)C(v)}, \quad C : u \mapsto 1 + \sum_{q \in \mathcal{Q}} \mathbb{1}_{\pi_q \leq u}.$$

1.2.4. Extensions du SBM aux réseaux multicouches et à des collections de réseaux

Travailler à partir du SBM permet d'obtenir un cadre très flexible. L'objectif de cette section est de fournir une brève revue sélective de la littérature qui a été développée autour des extensions du SBM et de ses variantes aux réseaux multicouches et aux collections de réseaux. En faisant différentes hypothèses sur la relation entre les blocs latents des différentes couches ou réseaux et sur la façon dont la probabilité des connexions peut varier, une grande variété de modèles a été développée pour les réseaux que nous avons présentés dans la Section 1.2.1.

Je tiens à préciser que dans de nombreux cas, la différence entre un réseau multicouche et un réseau simple ne réside que dans la formulation, comme par exemple un réseau multiplexe et un réseau simple avec des interactions multi-dimensionnelles. Certains réseaux dynamiques présentés ici sont en fait des réseaux simples dont la loi des arêtes suit un processus stochastique.

Réseaux multiplexes et collection de réseaux

Les extensions du SBM à des collections de réseaux se concentrent sur des réseaux ayant une correspondance entre les nœuds. Ils ont beaucoup en commun avec les réseaux multiplexes dans leur choix de modélisation, c'est pourquoi je les traite ensemble dans cette section.

On retrouve la volonté de prendre en compte de multiples relations entre les nœuds dès le premier article sur le stochastic block model de Holland et al. (1983). Les relations entre i et j, $(i \neq j) \in \mathcal{V}^2$ sont représentées par un vecteur $\{0,1\}^M$ où M est le cardinal du type de relation. En sociologie, cette représentation est utile pour modéliser la réciprocité dans les relations, en considérant un réseau dirigé comme un réseau non-dirigé à deux couches où la loi des arêtes partant des mêmes nœuds est jointe, i.e. on considère alors le couple $(X_{ij}, X_{ji}) = (X_{ij}^1, X_{ij}^2) \in \{0,1\}^2$ plutôt que chaque arête individuellement.

Dans ce même esprit, Barbillon et al. (2017) proposent un SBM pour réseaux multiplexes où les arêtes des différentes couches entre deux mêmes nœuds sont liées et dépendent des blocs d'individus :

$$\mathbb{P}(X_{ij} = k | Z_i, Z_j) = \alpha_{Z_i Z_j}(k), \qquad \sum_{k \in \{0,1\}^M} \alpha_{Z_i Z_j}(k) = 1.$$

L'hypothèse que les arêtes sont liées peut-être relâchée. On considère alors que chaque arête de chaque couche est générée indépendamment pour chaque $m \in \mathcal{M}$ (Han et al., 2015; Paul and Chen, 2020) :

$$X_{ij}^m | Z_i, Z_j \sim f(\cdot; \alpha_{Z_i Z_j}^m).$$

Une approche similaire pour le MMSBM est développée par Airoldi et al. (2008) et De Bacco et al. (2017), les premiers introduisant en plus un paramètre de parcimonie spécifique à chaque couche. Tandis que Pavlović et al. (2020) intègrent au modèle, dans le cadre d'une collection de réseaux, des covariables sur les réseaux en s'inspirant du travail de Mariadassou et al. (2010) présenté en section 1.2.3. C'est également en lien avec les travaux de Paul and Chen (2016) qui considèrent un effet spécifique à chaque couche en supposant que :

$$\mathbb{P}(X_{ij}^m = 1 | Z_i, Z_j) = \text{logit}^{-1}(\alpha_{Z_i Z_j} + \beta_m).$$

Certains travaux (ex : Stanley et al., 2016; Tarres-Deulofeu et al., 2019) s'intéressent à trouver un clustering des couches en plus de celui des nœuds. Stanley et al. (2016) considèrent une partition des M couches en S strates. Dans chaque strate $s \in \{1, \ldots, S\}$, les nœuds sont répartis dans $|\mathcal{Q}_s|$ blocs indépendamment avec probabilités π^s. Connaissant les strates et les appartenances aux blocs il y a indépendance des lois de chaque arête :

$$X_{ij}^m | m \in s, Z_i^s, Z_j^s \overset{ind}{\sim} Bern(\alpha_{Z_i^s Z_j^s}^s).$$

Finalement, Vallès-Català et al. (2016) se demandent si un réseau simple n'est pas l'agrégation des différentes couches d'un réseau multiplexe. Pour un réseau à deux couches $\mathbf{X} = (X^1, X^2)$, ils considèrent deux modes d'agrégation pour le réseau observé X^O. L'agrégation `AND` où $X^O_{ij} = X^1_{ij}X^2_{ij}$ et l'agrégation `OR` où $X^O_{ij} = 1 - (1 - X^1_{ij})(1 - X^2_{ij})$ pour tout $(i \neq j) \in \mathcal{V}^2$ et adaptent le SBM à ces deux cas. Peixoto (2015) se sert du réseau agrégé comme covariable pour générer des couches dépendantes d'un réseau multiplexe.

Pour les collections de réseaux, Le et al. (2018) supposent l'existence d'un véritable réseau X, dont les réseaux de la collection $\{X^m\}_{m\in\mathcal{M}}$ seraient des versions bruitées. Le réseau X suit un SBM et le bruit $P \in [0, .5]^{Q\times Q}$ et $Q \in [0, .5]^{Q\times Q}$ respecte la structure par bloc de X de manière que : $\mathbb{P}(X^m_{ij} = 1|X_{ij}, Z_i, Z_j) = X_{ij}P_{Z_iZ_j} + (1 - X_{ij})(1 - Q_{Z_iZ_j})$.

D'autres auteurs s'intéressent à la variabilité des appartenances aux blocs entre les réseaux. Paul and Chen (2018) autorisent les nœuds à changer de blocs entre les réseaux. Pour cela, ils définissent $\bar{Z} \in [0, 1]^{n\times Q}$ comme étant la moyenne des appartenances de chacun des n nœuds, dans un des Q blocs à travers les M réseaux, puis définissent une matrice de transition T de taille $Q \times Q$, tel que

$$Z^m_i \sim Mult(1, \bar{Z}_iT), i \in \mathcal{V}, m \in \mathcal{M}.$$

La loi de $X^m|Z^m$ suit alors celle d'un SBM de paramètre de connexion α^m, avec pour des raisons d'identifiabilité la contrainte que le vecteur des connexions intra-blocs $(\alpha^m_{11}, \ldots, \alpha^m_{QQ})$ soit le même pour tout $m \in \mathcal{M}$.

Sweet et al. (2014) considèrent un mixed membership SBM (MMSBM) qui, par différentes configurations de hiérarchie de lois a priori sur les paramètres de modèles, permet de modéliser des structures de dépendances complexes entre les réseaux.

Réseaux temporels et spatiaux

Sur les modèles SBM pour réseaux dynamiques, je signale les récentes revues de Kim et al. (2018) et Lee and Wilkinson (2019). Des modèles dynamiques ont déjà été décrits dans les extensions du SBM section 1.2.3 (DuBois et al., 2013; Matias et al., 2018). Ici, nous sommes concernés par l'observation de réseaux en plusieurs points du temps ou de l'espace. L'évolution des blocs est modélisée par une chaine de Markov cachée. Des travaux considèrent également différents régimes à travers le temps, en classifiant des intervalles de temps en plus des nœuds (Corneli et al., 2016). Yang et al. (2011) proposent un SBM dont les blocs varieraient dans le temps, en dépendant du bloc au temps précédent, i.e. pour un réseau à T temps :

$$\mathbb{P}(Z^{t+1}_i = q|Z^t_i) = \pi_{qq'} \quad \forall t \in \{1, \ldots, T-1\}, \forall i \in \mathcal{N},$$

où π est une matrice de transition. Ils supposent que les probabilités de connexion restent constantes à travers le temps. Matias and Miele (2017) et Xu and Hero (2013) relâchent les contraintes sur les probabilités de connexion en les autorisant à varier dans le temps (à l'exception pour Matias and Miele (2017) des connexions

intra-blocs où pour des raisons d'identifiabilité des blocs à travers le temps $\alpha^t_{qq} = \alpha^{t'}_{qq}$ pour tout $q \in Q$, $(t, t') \in \{1, \dots, T\}^2$). Cette idée a été récemment étendue pour prendre en compte la réciprocité en modélisant un couple d'arête à valeur dans $\{(0,0), (0,1), (1,0), (1,1)\}$ par Bartolucci et al. (2018).

Enfin, bien que cela ne porte pas à proprement parler sur un réseau multicouche, Miele et al. (2014) proposent un SBM contraint spatialement en faisant, par une approche de vraisemblance pénalisée, un compromis entre un réseau d'infrastructure connu et les blocs du réseau à analyser.

Autres types de réseaux multicouches

Bar-Hen et al. (2020) proposent une formulation très générale appelée generalized multipartite SBM pour modéliser des réseaux dont la décomposition des nœuds en K groupes fonctionnels est déjà connue. On s'intéresse alors à retrouver des blocs à l'intérieur de ces groupes fonctionnels. Cela revient à étudier une collection de SBM et LBM dont certains partagent les mêmes nœuds et blocs. Ainsi, pour tout $k \in \{1, \dots, K\}$,

$$\mathbb{P}(Z^k_i = q) = \pi^k_q \quad q \in \{1, \dots, Q_k\}, i \in \mathcal{V}_k,$$

et pour tout pour $(k, k') \in \{1, \dots, K\}^2$,

$$X^{kk'}_{ij} | Z^k_i, Z^{k'}_j \sim f^{k,k'}(\cdot; \alpha^{kk'}_{Z^k_i Z^{k'}_j}), \quad (i,j) \in \mathcal{D}_{k,k'}.$$

1.2.5. Techniques d'inférence et algorithmes

Dans le cadre du SBM, l'objectif de la plupart des algorithmes d'inférence est de retrouver les blocs d'appartenance et d'estimer les paramètres du modèle. Pour un grand nombre de SBMs, une fois les blocs retrouvés, l'estimation des paramètres est triviale. Il existe une grande diversité dans les méthodes d'inférence. Des méthodes Monte Carlo par Chaîne de Markov (MCMC) ont été développées avec un échantillonneur de Gibbs (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001). Ces méthodes sont très coûteuses et ne permettent d'inférer que des petits réseaux. Ainsi, des travaux ont tenté d'en améliorer les performances, en proposant pour un nombre quelconque de blocs d'autres échantillonneurs (McDaid et al., 2013) ou en développant des heuristiques ayant des performances équivalentes mais computationnellement bien moins coûteuses qu'un MCMC exact (Peixoto, 2014a; Kuhn et al., 2020). Une autre famille de méthode très utilisée est celle des approches variationnelles (Jordan et al., 1999; Blei et al., 2017) que je développe plus en détail ci-dessous. Je mentionne également des méthodes d'inférence basées sur la méthode des moments (Bickel et al., 2011) ou sur la distribution des degrés empiriques (Channarond et al., 2012). De nombreuses revues de la littérature traitent des méthodes d'inférence pour le SBM (Lee and Wilkinson, 2019; Matias and Robin, 2014). Certaines se focalisent sur les résultats théoriques des différentes approches disponibles et leurs limites fondamentales (Abbe, 2016; Zhao, 2018), tandis que

d'autres se concentrent sur la comparaison empirique de l'efficacité des différents algorithmes (Funke and Becker, 2019; Ghasemian et al., 2020a).

Méthodes variationnelles

Dans une approche par maximum de vraisemblance, l'objectif est de maximiser la vraisemblance des données observées. Pour faire cela, une idée naturelle serait d'intégrer sur les variables latentes la vraisemblance complète :

$$\mathcal{L}_{\boldsymbol{\theta}}(\mathbf{X}) \quad = \quad \sum_{z \in \mathcal{Z}} \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z} = z),$$

où $\mathcal{Z}$ est l'ensemble de toutes les combinaisons d'affiliation de blocs possibles et $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\alpha}\}$ est l'ensemble des paramètres du modèle. Mais, pour un SBM sur un réseau unipartite, $|\mathcal{Z}| = Q^n$, ce qui rend le calcul inaccessible numériquement.

La méthode la plus utilisée pour inférer des modèle à variables latentes est l'algorithme EM (Dempster et al., 1977, Expectation-Maximization) qui réalise une approximation locale de l'estimateur du maximum de vraisemblance. L'EM nécessite le calcul de la loi des variables latentes sachant les variables observées. Mais, dans le cas du SBM, ce calcul n'est plus accessible numériquement dès lors que n et Q ne sont pas petits. En utilisant des outils issus des modèles graphiques (Lauritzen, 1996), on peut montrer que la loi de $\mathbf{Z}|\mathbf{X}$ ne peut pas être factorisée comme illustrée par la moralisation du graphe dirigé acyclique (DAG) de la figure 1.4. Ainsi, l'algorithme EM ne peut pas être appliqué à notre cas, mais l'on peut se rabattre sur une approximation variationnelle (Jordan et al., 1999).

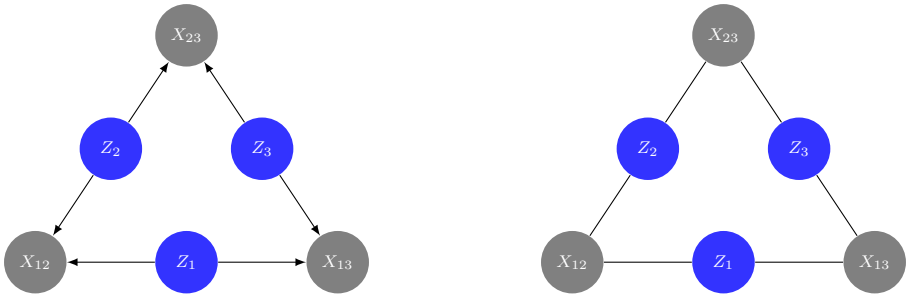


FIGURE 1.4 – Gauche : DAG d'un SBM à 3 nœuds. Droite : graphe moral associé au DAG.

Dans l'EM variationnel (VEM), nous voulons maximiser la borne variationnelle qui est une borne inférieure de la log-vraisemblance des données observées :

$$\begin{aligned} \ell_{\boldsymbol{\theta}}(\mathbf{X}) := \log \mathcal{L}_{\boldsymbol{\theta}}(\mathbf{X}) \geq \mathcal{J}(\mathcal{R}, \boldsymbol{\theta}) \quad &:= \quad \ell_{\boldsymbol{\theta}}(\mathbf{X}) - \mathrm{KL}(\mathcal{R} || \mathbb{P}_{\boldsymbol{\theta}}(\cdot|\mathbf{X})) && (1.3) \\ &= \quad \mathbb{E}_{\mathcal{R}}[\ell_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z})] + \mathcal{H}(\mathcal{R}), && (1.4) \end{aligned}$$

où KL est la divergence de Kullback-Leibler :

$$\mathrm{KL}(\mathcal{R} || \mathbb{P}_{\boldsymbol{\theta}}(\cdot|\mathbf{X})) = \mathbb{E}_{\mathcal{R}} \left[\log \frac{\mathcal{R}(\mathbf{Z})}{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})} \right],$$

et $\mathcal{H}$ est l'entropie de Shannon :

$$\mathcal{H}(\mathcal{R}) = -\mathbb{E}_{\mathcal{R}}[\log \mathcal{R}(\mathbf{Z})] = -\sum_{z} \mathcal{R}(z) \log \mathcal{R}(z),$$

et $\mathcal{R}$ est une distribution sur les variables latentes $\mathbf{Z}$. Dans le cas du SBM, la complexité computationelle vient de l'absence de factorisation de la loi de $\mathbf{Z}|\mathbf{X}$. Le VEM repose alors sur une approximation à champ moyen de manière à ce que $\mathcal{R}$ soit choisi dans une famille de distribution entièrement factorisable de la forme :

$$\mathcal{R}(\mathbf{Z}) = \prod_{i \in \mathcal{V}} \mathcal{R}_i(Z_i) = \prod_{i \in \mathcal{V}} \prod_{q \in \mathcal{Q}} \tau_{iq}^{Z_{iq}},$$

où $\tau_{iq} = \mathbb{E}_{\mathcal{R}}[Z_{iq}]$, avec la contrainte que $\sum_{q \in \mathcal{Q}} \tau_{iq} = 1$ pour tout $i \in \mathcal{V}$. $\boldsymbol{\tau} := \{(\tau_{i1}, \ldots, \tau_{iQ})\}_{i \in \mathcal{V}}$ sont appelés les paramètres variationnels.

Le VEM consiste alors à itérer les deux étapes suivantes :

VE**-step** On calcule

$$\boldsymbol{\tau}^{(t+1)} = \arg\max_{\boldsymbol{\tau}} \mathcal{J}(\mathcal{R}, \boldsymbol{\theta}^{(t)}),$$

cela revient à maximiser l'équation (1.3) et ainsi à minimiser la divergence de Kullback-Leibler, c'est à dire à trouver la loi variationnelle des affectations aux blocs la plus proche de $\mathbf{Z}|\mathbf{X}$ pour les paramètres courants $\boldsymbol{\theta}^{(t)}$.

M**-step** On calcule

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{J}(\mathcal{R}^{(t)}, \boldsymbol{\theta}),$$

ce qui revient à maximiser l'équation (1.4).

Pour l'implémentation de l'algorithme VEM dans le cas du SBM standard (Daudin et al., 2008), l'étape VE repose sur un algorithme de point fixe. Il n'y a pas de garantie de l'unicité de ce point fixe et donc que l'étape VE va accroître la borne variationnelle en pratique. Toutefois, cela fonctionne très bien empiriquement. Vu et al. (2012) remplacent la maximisation de l'étape VE par une étape dite de "Generalized" E-step qui se contente d'augmenter la borne variationnelle. Pour l'étape M, il existe une forme close et les estimateurs des paramètres sont facilement interprétables :

$$\hat{\pi}_q = \frac{1}{n} \sum_{i \in \mathcal{V}} \tau_{iq} \qquad \hat{\alpha}_{qr} = \frac{\sum_{(i \neq j) \in \mathcal{V}^2} \tau_{iq} \tau_{jr} X_{ij}}{\sum_{(i \neq j) \in \mathcal{V}^2} \tau_{iq} \tau_{jr}},$$

il s'agit respectivement de la proportion moyenne de nœuds dans le bloc q et la fréquence moyenne des arêtes entre les blocs q et r.

Pour proposer une partition des nœuds, on utilise le maximum a posteriori (MAP) après convergence de l'algorithme VEM en posant :

$$\hat{Z}_i^{MAP} = \arg\max_{q \in \mathcal{Q}} \hat{\tau}_{iq}, \qquad \forall i \in \mathcal{V}.$$

Des résultats théoriques asymptotiques existent, sur la convergence de la loi a posteriori des blocs vers une masse de Dirac située en les véritables appartenances

(Mariadassou and Matias, 2015), sur la consistance des estimateurs variationnels (Celisse et al., 2012) et leur normalité asymptotique (Bickel et al., 2013). Les résultats sur les estimateurs variationnels ont été étendus aux réseaux ayant des données manquantes (Mariadassou and Tabouy, 2020) et aux réseaux bipartites (Brault et al., 2020).

Quelques travaux dans la littérature ne font pas reposer leur approche variationnelle sur une famille de loi factorisable. D'autres algorithmes se basant sur la propagation des convictions (Pearl, 1982, Belief Propagation) sont populaires pour les réseaux parcimonieux (Decelle et al., 2011), car l'approximation effectuée dans cet algorithme est exacte pour les réseaux en forme d'arbre. Par ailleurs, les travaux de Yin et al. (2020) ont pour objectif de choisir $\mathcal{R}$ parmi une famille de loi plus large. Une version bayésienne de l'approche variationnelle est développée par Latouche et al. (2012).

Initialisation de l'algorithme VEM

La borne variationnelle n'est pas concave et l'algorithme VEM peut se coincer dans des maximaux locaux de la borne variationnelle. Ainsi, l'algorithme est très sensible à l'initialisation. Des stratégies d'initialisation communément utilisées incluent, d'essayer un grand nombre d'initialisation aléatoire, d'utiliser le clustering obtenu à partir d'un algorithme moins coûteux tel que le clustering spectral ou bien d'initialiser à partir de modèle voisin (des modèles tels que le nombre de blocs appartient à $\{Q-1, Q+1\}$) (Daudin et al., 2008; Leger et al., 2020). Lorsque un modèle voisin a été inféré, la classification des nœuds qu'il procure peut servir de bon point d'initialisation locale pour l'algorithme VEM. Pour obtenir cette initialisation, si le nombre de blocs initiaux est plus grand que Q, on fusionne alors 2 blocs, tandis que si ce nombre est plus petit, on sépare un bloc en deux (le bloc q par exemple) en réalisant un clustering des lignes de la matrice $(A_{ij})_{i\in q, j\in\mathcal{V}}$.

Clustering spectral Une stratégie très commune pour initialiser l'algorithme VEM est de partir de la partition des nœuds obtenus par l'absolute spectral clustering (Rohe et al., 2011). Les auteurs ont prouvé des résultats asymptotiques sur le nombre de nœuds mal classifiés pour une version du SBM où les blocs sont des paramètres et non des variables latentes. L'absolute spectral clustering est décrit dans l'algorithme 1.

Algorithm 1: Absolute Spectral Clustering

Data: A une matrice d'adjacence de taille $n \times n$, Q le nombre de clusters
Définir $L_{abs} = D^{-1/2}AD^{-1/2}$ où $D_{ii'} = \mathbb{1}_{i=i'} \sum_j A_{ij}$
Trouver les Q vecteurs propres $u_1, \ldots, u_k$ correspondant aux Q plus grandes valeurs propres en valeur absolue de L_{abs}
Définir $U = [u_1, \ldots, u_Q]$ une matrice de taille $n \times Q$
Classifier les lignes de U avec l'algorithme des k-means
return *Une partition des n nœuds en clusters $C_1, \ldots, C_Q$*

Pour plus de détails sur le spectral clustering, je renvoie vers Von Luxburg (2007).

1.2.6. Sélection de modèle

Bien que les premiers travaux sur le SBM supposent le nombre de groupes fixé (Snijders and Nowicki, 1997), de nombreuses méthodes ont été développées depuis pour estimer le nombre de blocs dans le SBM. La revue de Lee and Wilkinson (2019) procure d'avantage de détails. Certains travaux se basent sur des méthodes de validation croisée, en leave-one-out (Kawamoto and Kabashima, 2017) ou par bloc de la matrice d'adjacence (Chen and Lei, 2018). L'avantage de ces méthodes est qu'elles sont facilement généralisables à des modèles plus complexes et permettent, en plus de choisir le nombre de blocs, de comparer différents modèles entre eux par exemple un SBM et un DCSBM. Toutefois, Vallès-Català et al. (2017) mettent en garde contre les risques de surapprentissage, le modèle donnant les meilleurs prédictions n'étant pas forcément le plus parcimonieux au sens de la minimum description length (MDL). Ce critère est souvent utilisé dans les formulations bayésiennes du SBM (Peixoto, 2014b).

Ce critère est comparable, sous certaines hypothèses à l'integrated classification likelihood (Biernacki et al., 2000; Daudin et al., 2008, ICL), critère que j'adapterai dans mes travaux. Je donne ici l'idée principale, des développements plus détaillés ayant lieu dans les chapitres 2 et 3. Développé pour les modèles de mélange, l'ICL favorise des blocs bien séparés. Il s'obtient en prenant le log de la vraisemblance intégrée sur les paramètres du modèle des données complètes :

$$\log \mathcal{L}(\mathbf{X}, \mathbf{Z}|Q) = \log \int_{\boldsymbol{\alpha}} \mathcal{L}_{\boldsymbol{\alpha}}(\mathbf{X}|\mathbf{Z}, Q) p(\boldsymbol{\alpha}|Q) \mathrm{d}\alpha + \log \int_{\boldsymbol{\pi}} \mathcal{L}_{\boldsymbol{\pi}}(\mathbf{Z}|Q) p(\boldsymbol{\pi}|Q) \mathrm{d}\boldsymbol{\pi}, \quad (1.5)$$

en supposant l'indépendance des paramètres de mélange et de connexion. Le terme dépendant de $\boldsymbol{\pi}$ est calculé en choisissant une loi a priori adaptée (une loi de Dirichlet qui est conjuguée pour une loi multinomiale), dont on peut faire un développement asymptotique avec la formule de Stirling, tandis que le terme en $\boldsymbol{\alpha}$ l'est par une approximation asymptotique de type BIC (Schwarz, 1978) :

$$ICL(\mathbf{X}, Q) = \max_{\boldsymbol{\theta}} \ell_{\boldsymbol{\theta}}(\mathbf{X}, \hat{\mathbf{Z}}|Q) - pen(Q).$$

Deux approches existent pour la valeur de $\hat{\mathbf{Z}}$, Biernacki et al. (2000) proposent de remplacer $\hat{\mathbf{Z}}$ par le maximum a posteriori (MAP) de $\mathbf{Z}|\mathbf{X}$, i.e. $\hat{Z}_i = \arg\max_{q \in \mathcal{Q}} \mathbb{P}(Z_i = q|\mathbf{X})$, alors que McLachlan and Peel (2000) suggèrent d'utiliser l'estimateur de l'espérance conditionnelle $\hat{Z}_i = \mathbb{E}[Z_i|\mathbf{X}]$.

Dans le cadre de l'inférence variationnelle, $\hat{\mathbf{Z}}$ est choisie comme étant le MAP des paramètres variationnels, i.e. $\hat{Z}_i = \arg\max_{q \in \mathcal{Q}} \tau_{iq}$ ou bien par les paramètres variationnels eux-mêmes. Pour un SBM sur un réseau simple et non dirigé, la pénalité est donnée par :

$$pen(Q) = \frac{1}{2}\left((Q-1)\log(n) + \frac{Q(Q+1)}{2}\log(|\mathcal{D}|)\right).$$

La pénalité est facilement interprétable : le premier terme dépend du nombre de degré de liberté du paramètre de mélange $\boldsymbol{\pi}$ et du nombre de nœuds, tandis que le second dépend du nombre de paramètres de connexion et du nombre de dyades.

Dans le cadre du SBM, pour de nombreuses lois d'émission ayant des lois a priori conjuguées, l'approximation asymptotique n'est pas nécessaire. Latouche et al. (2012) utilisent une version non asymptotique comme critère de sélection, le couplant à une approximation de la vraisemblance observée par méthode bayésienne variationnelle. Ils argumentent que l'objectif est différent de celui de l'ICL, car leur critère (ILvb) se focalise sur l'estimation de la densité et non sur celle de la classification. On dispose d'une expression analytique pour l'équation (1.5), on parle alors d'ICL exact et ce critère devient la fonction objectif à optimiser (Côme and Latouche, 2015). Cela permet d'optimiser le nombre de blocs en même temps que le clustering, sans passer par un algorithme variationnel pour chaque taille de modèle.

Hayashi et al. (2015) adaptent et simplifient le Factorized Information Criterion(FIC) aux modèles de mélange, pour le SBM ce critère est appelé BICEM ou corrected ICL (cICL) :

$$cICL = \mathbb{E}_{\mathbf{Z}|\mathbf{X}}[\ell_{\hat{\boldsymbol{\theta}}}(\mathbf{X}, \mathbf{Z})] - pen(Q) + \mathcal{H}(p(\mathbf{Z}|\mathbf{X})),$$

cela revient à rajouter un terme d'entropie à l'ICL évalué avec la loi a posteriori des variables latentes et donc, comme ILvb, à approcher la vraisemblance des données observées plutôt que la vraisemblance classifiante. Hayashi et al. (2016) comparent cette approche avec d'autres critères, dont une implémentation directe du FIC dans des modèles bayésiens pour des graphes parcimonieux, qui permet d'intégrer la sélection de modèle à l'inférence.

D'autres approches par vraisemblance pénalisée ont été développées. Yan (2016) propose un critère de sélection bayésien permettant de comparer différents modèles (model selection) ayant un nombre de groupes différent (order selection) et l'applique au SBM et au DCSBM. Wang and Bickel (2017) regardent si des blocs peuvent être séparés en analysant le rapport de vraisemblance et en dérivent un critère de vraisemblance pénalisée consistant, de la forme

$$\beta(Q) = \sum_{\boldsymbol{\theta} \in \Theta_Q} \log \mathcal{L}_{\boldsymbol{\theta}}(X) - \lambda \frac{Q(Q+1)}{2} n \log n,$$

où λ est un paramètre de régularisation. La consistance du critère est conservée même si la vraisemblance des données observées peut-être remplacée par sa borne variationnelle $\mathcal{J}$ en utilisant un résultat de convergence de la borne variationnelle (Bickel et al., 2013). Saldaña et al. (2017) proposent un composite-likelihood BIC en considérant les $\mathbf{Z}$ fixes et inconnus (pas une variable latente) et arrivent à une pénalité en $\frac{Q(Q+1)}{2} \log n$. Hu et al. (2020) argumentent que ces deux critères sous-estiment et surestiment respectivement le nombre de blocs et proposent un corrected Bayesian information criterion (CBIC) consistant de la forme suivante :

$$\max_{z \in \mathcal{Q}^n} \sup_{\boldsymbol{\alpha} \in A_q} \log \mathcal{L}_{\boldsymbol{\alpha}}(\mathbf{X}|z) - \left(\lambda n \log Q + \tfrac{Q(Q+1)}{2} \log n\right).$$

Lorsque $\lambda = 1$, ce critère est très proche de la version asymptotique de l'ICL si l'on considère des blocs de tailles égales et un $\boldsymbol{\pi}$ uniforme.

Bien que les résultats empiriques montrent que l'ICL donne de très bons résultats pour les réseaux denses, signalons que dans les cas où le nombre de blocs trouvé n'est pas le bon, l'ICL a tendance à sous-estimer le nombre de blocs (Hu et al., 2020; Hayashi et al., 2016; Latouche et al., 2011; Mariadassou et al., 2010).

Lien entre ICL et BIC En passant en revue les différents critères de sélection de modèle ci-dessus, nous remarquons que certains se concentrent sur le clustering des données, i.e. retrouver le bon nombre de clusters, tandis que d'autres essaient de retrouver la bonne dimension du modèle, i.e. retrouver le bon nombre de paramètres. Dans le cadre du SBM ces deux approches sont liées et nous illustrons cela en comparant la version asymptotique de l'ICL (version McLachlan and Peel, 2000) au BIC (Schwarz, 1978). L'approche de sélection de modèle par classification, peut être reliée à celle par densité grâce à la relation suivante :

$$\ell(\mathbf{X}|Q) = \ell(\mathbf{X},\mathbf{Z}|Q) - \ell(\mathbf{Z}|\mathbf{X},Q),$$

que l'on applique à l'expression du BIC.

$$\begin{aligned}
\mathrm{BIC}(Q) &= \max_{\boldsymbol{\theta}} \ell_{\boldsymbol{\theta}}(\mathbf{X}) - \mathrm{pen}_{\mathrm{BIC}}(Q) \\
&= \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})}[\ell_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z}|Q) - \ell_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X},Q)] - \mathrm{pen}_{\mathrm{BIC}}(Q) \\
&= \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{X})}[\ell_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z}|Q)] + \mathcal{H}(\mathbb{P}_{\boldsymbol{\theta}}(\cdot|\mathbf{X})|Q) - \mathrm{pen}_{\mathrm{BIC}}(Q).
\end{aligned}$$

Si l'on conjecture que $\mathrm{pen}_{\mathrm{BIC}}(Q) = \mathrm{pen}_{\mathrm{ICL}}(Q)$, on peut alors voir l'ICL comme le BIC pénalisé par un terme d'entropie. Ce terme est maximal lorsque

$$\mathbb{P}_{\theta}(\mathbf{Z}_i = q|\mathbf{X}) = \frac{1}{Q} \quad \forall q \in \mathcal{Q}$$

et a contrario tend vers 0 lorsque toute la masse se concentre en un seul point, ce qui est le cas asymptotiquement pour le SBM (Mariadassou and Matias, 2015). Donc l'ICL pénalise d'avantage que le BIC les clusters mal séparés, mais sous cette conjecture, les 2 critères sont équivalents asymptotiquement.

Dans notre cas, nous ne connaissons pas la loi de $\mathbf{Z}|\mathbf{X}$ et nous avons recours à une approximation variationnelle, alors :

$$\max_{\boldsymbol{\theta}} \ell_{\boldsymbol{\theta}}(\mathbf{X}) = \max_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\hat{\mathcal{R}}}[\ell_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z}|Q)] + \mathcal{H}(\hat{\mathcal{R}}|Q)}_{\mathcal{J}(\hat{\mathcal{R}},\boldsymbol{\theta}|Q)} - \underbrace{\mathcal{H}(\hat{\mathcal{R}}|Q) + \mathbb{E}_{\hat{\mathcal{R}}}[\ell_{\theta}(\mathbf{Z}|\mathbf{X},Q)]}_{\mathrm{KL}(\hat{\mathcal{R}}\|\mathbb{P}_{\boldsymbol{\theta}}(\cdot|\mathbf{X})|Q)},$$

et sous la même conjecture sur les pénalités que précédemment, on peut obtenir une approximation du BIC à partir du calcul de l'ICL :

$$\begin{aligned}
\mathrm{BIC}(Q) &= \mathrm{ICL}(Q) + \mathcal{H}(\hat{\mathcal{R}}|Q) + \mathrm{KL}(\hat{\mathcal{R}}\|\mathbb{P}_{\boldsymbol{\theta}}(\cdot|\mathbf{X})|Q) \\
&\geq \max_{\boldsymbol{\theta}} \mathcal{J}(\hat{\mathcal{R}},\boldsymbol{\theta}|Q) - \mathrm{pen}_{\mathrm{ICL}}(Q),
\end{aligned}$$

au prix d'un terme de minoration représenté par la divergence de Kullback-Leibler. C'est la même erreur que celle faite lors de l'approximation variationnelle à taille de modèle fixée.

Je renvoie vers Baudry (2015) pour une discussion intéressante sur le critère ICL et au chapitre 8 de Celeux et al. (2018) pour une introduction au critère de sélection de modèle dans les modèles de mélange.

Je résume dans la table 1.1 les extensions de SBM aux collections de réseaux présentés en section 1.2.4 ainsi que leurs méthodes d'inférence, de sélection de modèle et le type de données sur lesquelles ces modèles sont appliqués.

Article	Type de réseau	SBM	Inférence	#Blocs	Applications	Autre
Airoldi et al. (2008)	Multiplexe	MMSBM	VEM	BIC, CV	Social, protéine	
Yang et al. (2011)	Temporel	SBM	MCMC	–	Social, co-auteur	Comparent estimation ponctuelle du VEM à l'estimation bayésienne par MCMC. Inférence en ligne et hors ligne.
Xu and Hero (2013)	Temporel	logit SBM	Changement de bloc		Email d'Enron	Inférence en ligne
Miele et al. (2014)	Spatial	SBM	VEM	ICL	Écologie	VEM régularisé
Sweet et al. (2014)	Collection	MMSBM	MCMC	–	Social	Covariables
Han et al. (2015)	Multiplexe	SBM	Spectral, VEM	ICL	Social	Consistance du clustering spectral et de l'EMV ($M \to \infty$)
Peixoto (2015)	Multiplexe	MM–DC–SBM	MCMC	MDL	Politique, Transport, Social	Arêtes valuées
Paul and Chen (2016)	Multiplexe	SBM	VEM	–	Twitter	Consistance de l'EMV, Test d'indépendance entre couches par bootstrap paramétrique
Stanley et al. (2016)	Multiplexe	SBM	VEM	–	Microbiome	Clustering des couches
Vallès-Català et al. (2016)	Multiplexe	SBM	MCMC	–	Divers	
Barbillon et al. (2017)	Multiplexe	SBM	VEM	ICL	Social	
De Bacco et al. (2017)	Multiplexe	MMSBM	VEM	–	Social	Clustering des couches
Matias and Miele (2017)	Temporel	SBM	VEM	ICL	Contact, social d'animaux	
Le et al. (2018)	Collection	SBM	tous + EM	–	IRMf	
Paul and Chen (2018)	Collection	SBM	VEM, MF	–	IRMf	Test pour changement de blocs
Matias et al. (2018)	Dynamic	ppSBM	VEM	ICL	Transport	M-step par histogramme
Bartolucci et al. (2018)	Temporel	SBM	VEM	ICL	Social	Se concentre sur la réciprocité
Tarres-Deulofeu et al. (2019)	Multiplexe	MMSBM	VEM	CV	Email, médicament	Clustering des couches
Bar-Hen et al. (2020)	Multipartite	SBM	VEM	ICL	Ethnobiologie, écologie	
Paul and Chen (2020)	Multiplexe	SBM	spectral, MF	–	–	Consistance de l'inférence
Pavlović et al. (2020)	Collection	SBM	VEM	ICL	Réseaux cérébraux	Covariables, test sur les changements de connections entre couches

TABLE 1.1 – Quelques SBM pour réseaux multicouches et collection de réseaux

1.2.7. Comparaison de clustering

La comparaison de clustering a de nombreuses applications dans le cadre de l'analyse de réseaux. Cela permet entre autre de :

1. Juger dans le cadre de simulations de la performance des algorithmes et des méthodes d'inférence,
2. Juger de la stabilité des clusterings obtenus, en relançant les algorithmes à partir de données bruitées (cf. section 1.2.8),
3. Comparer les clusterings obtenus via différentes méthodes,
4. Quantifier la similarité entre différentes couches d'un réseau multiplexe ou temporel, ou différents réseaux d'une collection impliquant les mêmes nœuds.

À propos du premier point, pour juger de l'efficacité de la modélisation et de son implémentation à retrouver les vrais clusters, il existe des benchmarks à partir de réseaux simulés (Lancichinetti and Fortunato, 2009). Toutefois le SBM est un modèle génératif et il est aisé de simuler à partir du modèle. Certains auteurs s'intéressent également à comparer les clusterings obtenus à ceux d'un *ground truth* (une vérité de terrain qui correspond à une covariable qualitative sur les nœuds) (Hric et al., 2014). Je n'adopterai pas cette solution car elle me parait inadaptée et mal posée (Peel et al., 2017; Olhede and Wolfe, 2013). En particulier, Peel et al. (2017) montrent que la partition des nœuds pouvant générer le réseau n'est pas unique et prouvent un *no free lunch theorem* avec l'adjusted mutual information (AMI) comme indice de comparaison de clustering. Par ailleurs, bien que le SBM permet de retrouver des formes de structures universelles (Young et al., 2018), quelques travaux récents restreignent l'espace des paramètres de manière à retrouver des structures facilement interprétables : communautaire (Zhang and Peixoto, 2020) ou cœur-périphérie (Zhang et al., 2015; Gallagher et al., 2021).

Je présente deux indices particulièrement utilisés dans le cadre des réseaux. L'un est basé sur la théorie de l'information, le NMI (Danon et al., 2005, Normalized Mutual Information) et l'autre, l'ARI (Hubert and Arabie, 1985, Adjusted Rand Index), sur le principe du comptage de paires. Pour une discussion sur les avantages de différents indices de comparaison de partitions, je renvoie à l'article de Fortunato and Hric (2016).

NMI Soit $(C, D) \in \mathcal{Q}^n \times \mathcal{Q}^n$ deux vecteurs aléatoires de clustering des nœuds, alors le NMI vaut :

$$\mathrm{NMI}(C, D) = \frac{\mathcal{H}(C) + \mathcal{H}(D) - \mathcal{H}(C, D)}{\frac{1}{2}(\mathcal{H}(C) + \mathcal{H}(D))}. \tag{1.6}$$

où $\mathcal{H}$ est l'entropie de Shannon $\mathcal{H}(X) = -\sum_x \mathbb{P}(X = x) \log(\mathbb{P}(X = x))$. D'autres normalisations que celle du dénominateur de l'équation (1.6) sont également utilisées, en particulier $\sqrt{\mathcal{H}(C)\mathcal{H}(D)}$ et $\max(\mathcal{H}(C), \mathcal{H}(D))$. Posons la table de contingence

des clusterings $R_{qr} = \sum_{i\in\mathcal{V}} \mathbb{1}_{C_i=q,D_i=r}$ pour tout $(q,r) \in \mathcal{Q}^2$ et $R_{+r} = \sum_{q\in\mathcal{Q}} R_{qr}$ (resp. $R_{q+} = \sum_{r\in\mathcal{Q}} R_{qr}$). Alors on peut réécrire le NMI comme suit :

$$\mathrm{NMI}(C,D) = \frac{2(\sum_{q\in\mathcal{Q}}(R_{q+}\log\frac{R_{q+}}{n} + R_{+q}\log\frac{R_{+q}}{n} - \sum_{r\in\mathcal{Q}}\frac{R_{qr}}{n}\log\frac{R_{qr}}{n^2}))}{\sum_{q\in\mathcal{Q}}(R_{q+}\log\frac{R_{q+}}{n} + R_{+q}\log\frac{R_{+q}}{n})}.$$

Le NMI vaut 1 si les partitions sont identiques et 0 si elles sont indépendantes.

Je signale que l'AMI utilisée ci-dessus par Peel et al. (2017) est une version ajustée du NMI qui prend en compte la géométrie de l'espace des clusterings. En notant $I(C,D)$ l'information mutuelle au numérateur de l'équation (1.6), elle est donnée par :

$$\mathrm{AMI}(C,D) = \frac{I(C,D) - \mathbb{E}[I(C,D)]}{\sqrt{\mathcal{H}(C)\mathcal{H}(D)} - \mathbb{E}[I(C,D)]}.$$

ARI Pour l'ARI, on regarde si des paires de nœuds sont classées dans le même bloc ou non dans les 2 clusterings. Cela revient à regarder pour chaque dyade, si elle est intra-bloc ou inter-bloc. Introduisons les valeurs suivantes pour chaque couple de clusterings :

$$a_{11} = \sum_{(i\neq j)\in\mathcal{D}} \mathbb{1}_{C_i=C_j,D_i=D_j}, \qquad a_{01} = \sum_{(i\neq j)\in\mathcal{D}} \mathbb{1}_{C_i\neq C_j,D_i=D_j},$$
$$a_{10} = \sum_{(i\neq j)\in\mathcal{D}} \mathbb{1}_{C_i=C_j,D_i\neq D_j}, \qquad a_{00} = \sum_{(i\neq j)\in\mathcal{D}} \mathbb{1}_{C_i\neq C_j,D_i\neq D_j}.$$

Le Rand Index (RI) est défini par $\mathrm{RI}(C,D) = \frac{a_{11}+a_{00}}{\binom{n}{2}}$. Cet indice reste défini même si le nombre de blocs est différent entre les 2 clusterings et permet ainsi de comparer des clusterings obtenus à partir de modèles de tailles différentes. Il est possible de faire le lien entre les termes $(a_{kl})_{(k,l)\in\{0,1\}^2}$ et la table de contingence $R_{qr} = \sum_{i\in\mathcal{V}} \mathbb{1}_{C_i=q,D_i=r}$, $(q,r) \in \mathcal{Q}_C \times \mathcal{Q}_D$, en particulier :

$$a_{11} = \sum_{(q,r)\in\mathcal{Q}_C\times\mathcal{Q}_D} \binom{R_{qr}}{2}, \quad a_{00} = \binom{n}{2} + \sum_{(q,r)\in\mathcal{Q}_C\times\mathcal{Q}_D} \binom{R_{qr}}{2} - \sum_{q\in\mathcal{Q}_C} \binom{R_{q+}}{2} - \sum_{r\in\mathcal{Q}_D} \binom{R_{+r}}{2}.$$

L'ARI corrige le RI en supposant que la table de contingence construite suit une loi hypergéométrique :

$$\mathrm{ARI}(C,D) = \frac{\mathrm{RI}(C,D) - \mathbb{E}[\mathrm{RI}(C,D)]}{1 - \mathbb{E}[\mathrm{RI}(C,D)]}, \tag{1.7}$$

où $\mathbb{E}[\mathrm{RI}(C,D)] = 1 + 2\frac{\sum_{q\in\mathcal{Q}_C}\binom{R_{q+}}{2}\sum_{r\in\mathcal{Q}_D}\binom{R_{+r}}{2}}{\binom{n}{2}^2} - \frac{\sum_{q\in\mathcal{Q}_C}\binom{R_{q+}}{2}+\sum_{r\in\mathcal{Q}_D}\binom{R_{+r}}{2}}{\binom{n}{2}}$.

On peut alors réécrire l'ARI à partir des 4 termes d'adéquation de couple de dyades définis ci-dessus :

$$\mathrm{ARI}(C,D) = \frac{a_{11} - \frac{(a_{11}+a_{10})(a_{11}+a_{01})}{a_{11}+a_{10}+a_{01}+a_{00}}}{\frac{1}{2}(a_{11}+a_{10}+a_{01}+a_{00}) - \frac{(a_{11}+a_{10})(a_{11}+a_{01})}{a_{11}+a_{10}+a_{01}+a_{00}}}. \tag{1.8}$$

L'ARI vaut 1 si les deux partitions sont les mêmes et est positif si les partitions comparées sont plus semblables qu'en espérance, i.e. $\mathrm{RI}(C,D) > \mathbb{E}[\mathrm{RI}(C,D)]$.

1.2.8. Données manquantes et bruitées

Un réseau observé peut comporter des données manquantes, sur les nœuds, les covariables ou les dyades. Dans ce qui suit je me concentrerai sur les dyades du réseau. La problématique est différente sur les réseaux d'interactions en écologie et en sciences sociales.

Concernant les réseaux d'interactions écologiques, il est difficile de faire la différence entre une arête non-existante et une arête non-observée. Notre capacité à observer des interactions dépend du temps d'observation, celui-ci étant fini, les réseaux écologiques observés sont connus pour être incomplets (Blüthgen et al., 2008). Si je prends le cas d'un réseau binaire, certains 0 sont des "faux" 0 et on peut s'intéresser à prédire les liens manquants du réseau (Clauset et al., 2008). Réciproquement certaines erreurs ont pu avoir lieu lors de la collecte des données : une espèce a été confondue avec une autre, une erreur s'est produite lors de la numérisation des données, etc... Ainsi, il est possible d'observer des arêtes fallacieuses que l'on peut également essayer de distinguer des véritables arêtes (Guimerà and Sales-Pardo, 2009).

Pour les réseaux sociaux, bien que l'on puisse également observer des liens fallacieux ou manquants, des dyades peuvent également être non observées. Cela peut se produire lorsqu'un formulaire est rempli de manière incomplète ou bien que via la méthode d'échantillonnage choisie pour observer le réseau, certains individus présents dans le réseau n'aient pas pu être interviewés, n'aient pas répondu à un questionnaire (Žnidaršič et al., 2012) ou plus généralement que certaines informations n'aient pas pu être récoltées. Dans ce cas là, nous savons où se trouvent les informations manquantes. On peut alors s'intéresser à inférer un modèle à partir du réseau partiellement observé et/ou à prédire les valeurs des dyades manquantes.

De nombreuses méthodes ont été développées pour prédire les liens ou les dyades manquants (Martínez et al., 2016, pour une revue) et le SBM et sa version degré corrigé ont montré de bonnes performances sur des réseaux réels (Ghasemian et al., 2020a,b). De plus des résultats d'identifiabilité et des résultats asymptotiques sur les estimateurs (du maximum de vraisemblance et variationnels) existent pour différentes stratégies d'échantillonage (Tabouy et al., 2019; Mariadassou and Tabouy, 2020).

Soit $\mathcal{D} = \mathcal{D}^O \cup \mathcal{D}^U$, la division des dyades en dyades observées ($\mathcal{D}^O$) et dyades non-observées ($\mathcal{D}^U$). L'objectif est de retrouver la distribution de $\mathcal{D}^U|\mathcal{D}^O$. Représentons le réseau sous forme matricielle où $X_{ij} = \texttt{NA}$ pour tout $(i,j) \in \mathcal{D}^U$. Alors dans le cadre d'un SBM pour un réseau simple binaire :

$$\hat{p}_{ij} = \hat{\mathbb{P}}(X_{ij} = 1|\mathcal{D}^O) = \sum_{(q,r)\in\hat{\mathcal{Q}}^2} \hat{Z}_{iq}\alpha_{qr}\hat{Z}_{jr}, \tag{1.9}$$

où l'on peut remplacer $\hat{Z}_{iq}$ et $\hat{Z}_{jr}$ par leurs estimations variationnelles respectives. $\hat{\mathcal{Q}}$ peut être donné ou estimé par une des méthodes données en section 1.2.6.

Les arêtes manquantes et fallacieuses peuvent être estimées de la même manière. Dans ce cas là $\mathcal{D}^O = \mathcal{D}$ et on se restreint à proposer une prédiction d'existence ou d'erreur des arêtes du sous-ensemble $\mathcal{D}^0 := \{(i,j) \in \mathcal{D} : X_{ij} = 0\}$ dans le premier cas et $\mathcal{D}^1 := \{(i,j) \in \mathcal{D} : X_{ij} = 1\}$ dans le second cas.

Utilisation des données manquantes et bruitées

Les réseaux bruités et/ou partiellement observés peuvent également être utilisés à travers des simulations comme outil de sélection de modèles, de vérification d'algorithmes ou même pour quantifier les différences entre les réseaux dans une collection de réseaux. Je donne ici quelques exemples d'utilisation ainsi que les outils dont je me servirai dans cette thèse.

Sélection de modèle

Comme indiqué en section 1.2.6, Chen and Lei (2018) proposent, pour choisir le nombre de blocs dans un SBM et pour choisir entre SBM et un DCSBM, une méthode de validation croisée par bloc. L'ensemble des nœuds $\mathcal{V}$ est divisé en une partition aléatoire $(\mathcal{V}_1, \mathcal{V}_2)$, l'objectif est de minimiser une erreur de prédiction sur les arêtes $(i \neq j) \in \mathcal{V}_2^2$ qui sont manquantes. L'avantage de cette approche est qu'elle peut se généraliser facilement à n'importe quelle extension du SBM, mais au prix d'un fort coût algorithmique.

Quantifier la dépendance entre réseaux

On peut utiliser la prédiction de dyades manquantes pour le réseau. L'idée est de dire que deux couches de réseau sont interdépendantes si et seulement si l'information contenue dans l'une des couches permet d'améliorer la prédiction des arêtes de l'autre couche. C'est ce que font De Bacco et al. (2017), dans le cadre d'un MMSBM pour réseau multiplexe, en masquant une partie d'une couche et en comparant la prédiction des dyades manquantes entre un MMSBM sur cette unique couche et un MMSBM sur toutes les couches du réseaux multiplexe. Ils utilisent l'aire sous la courbe ROC (ROC AUC) pour comparer les prédictions, indices que je définis ci-dessous.

ROC AUC Pour juger de la performance de la prédiction on utilise le ROC AUC. Soit X^T la représentation matricielle du vrai réseau, X^O le réseau observé et $\mathcal{D}^U$ l'ensemble des dyades/arêtes à prédire. Soit R_{ij} le rang de p_{ij} pour tout $(i,j) \in \mathcal{D}^U$ par ordre décroissant. Définissons alors les taux de vrais positifs (TPR) et de faux positifs (FPR) en fonction d'un seuil $k \in]0,1]$:

$$\begin{aligned} TPR(k) &= \frac{|(i,j) \in \mathcal{D}^U : R_{ij} \leq k|\mathcal{D}^U|, X^T_{ij} = 1|}{|(i,j) \in \mathcal{D}^U : X^T_{ij} = 1|}, \\ FPR(k) &= \frac{|(i,j) \in \mathcal{D}^U : R_{ij} \leq k|\mathcal{D}^U|, X^T_{ij} = 0|}{|(i,j) \in \mathcal{D}^U : X^T_{ij} = 0|}. \end{aligned}$$

On définit alors ROC AUC $:= \int_0^1 TPR(FPR^{-1}(k))\mathrm{d}k$. Le ROC AUC prend des valeurs entre 0 et 1, et la prédiction est d'autant meilleure que cette valeur est grande.

Une méthode qui place la même probabilité sur chaque arête a un ROC AUC de $\frac{1}{2}$. La définition du ROC AUC est directement applicable aux données bruitées via la prédiction d'arêtes manquantes ou fallacieuses.

Stabilité du clustering

Simuler des données manquantes permet également d'évaluer la pertinence d'une modélisation jointe d'une collection en regardant si celle-ci améliore la stabilité du clustering (Žnidaršič et al., 2012). Pour cela, on utilise une méthode de comparaison de clustering (par ex. : l'ARI défini en section 1.2.7) afin de comparer le clustering obtenu à partir du réseau où l'information est complète et celui obtenu à partir du réseau avec des données manquantes. Plus formellement, soit X^T le vrai réseau et X^O le réseau observé. Soit C une fonction de clustering, notre indice de stabilité est alors donné par $\text{ARI}(C(\mathcal{V}|X^T), C(\mathcal{V}|X^O))$. Alors, dans le cadre d'une collection de réseau $\mathbf{X}$, les réseaux m et m' sont liés si l'un permet de stabiliser le clustering de l'autre, i.e.

$$ARI(C(\mathcal{V}_m|X^{m,T}, X^{m'}), C(\mathcal{V}_m|X^{m,O}, X^{m'})) > ARI(C(\mathcal{V}_m|X^{m,T}), C(\mathcal{V}_m|X^{m,O})).$$

1.3. Contributions de la thèse

Les contributions de cette thèse sont réparties en trois travaux. Deux ont trait à la modélisation conjointe de réseaux n'ayant pas de nœuds communs à travers des extensions du SBM. Dans le chapitre 2, nous nous intéressons à la modélisation de réseaux multiniveaux. Nous introduisons une dépendance entre les nœuds de ces niveaux, sous la forme d'une dépendance entre les blocs latents, et tentons de comprendre l'influence d'un niveau sur la structure de connexion de l'autre niveau ; la structure de chaque niveau est laissée libre. Nous considérons un autre type de dépendance dans le chapitre 3, où nous modélisons une collection de réseaux et tentons de retrouver une structure de connexion communes aux différents réseaux. La dépendance entre les réseaux est introduite par l'hypothèse qu'ils partagent une structure commune ce qui se traduit par une correspondance entre les blocs latents. Enfin dans un dernier travail, présenté au chapitre 4, nous utilisons des modèles paramétriques adaptés à des réseaux bipartites et en particulier le SBM pour dériver une expression exacte sous ce modèle d'un indice communément utilisé en écologie pour quantifier la résilience d'un écosystème face à la disparition d'espèce, appelé la *robustesse*. Cette nouvelle expression de la robustesse permet entre autres via la renormalisation des paramètres du SBM de comparer la robustesse d'une collection de réseaux. Ces trois travaux comportent des librairies complètement documentées pour le logiciel `R` permettant d'appliquer facilement ces méthodes à de nouveaux jeux de données. Je résume les contributions majeures des différents chapitres ci-dessous, les notations sont celles de l'article et peuvent différer de celles définies dans le début de cette introduction.

1.3.1. Un modèle à blocs stochastiques pour les réseaux multiniveaux

Dans le chapitre 2, un réseau multiniveau est défini comme la collection d'un réseau (niveau) inter-individuel, un réseau (niveau) inter-organisationnel et un réseau d'affiliation des individus aux organisations. Nous faisons l'hypothèse qu'un individu est affilié à une unique organisation. Dans le cadre d'un réseau multiniveau non dirigé avec n_I individus et n_O organisations, les niveaux inter-individuels et inter-organisationnels sont représentés par les matrices d'adjacences binaires $X^I \in \{0,1\}^{n_I \times n_I}$ et $X^O \in \{0,1\}^{n_O \times n_O}$, tandis que les affiliations le sont par une matrice A de taille $n_I \times n_O$ tel que :

$$A_{ij} = \begin{cases} 1 & \text{si l'individu } i \text{ est affilié à l'organisation } j, \\ 0 & \text{sinon} \end{cases}.$$

A est tel que $\forall i = 1, \ldots, n_I$, $\sum_{j=1}^{n_O} A_{ij} = 1$ car chaque individu est affilié à une et une seule organisation.

Nous proposons une modélisation jointe des réseaux inter-individuel et inter-organisationnel à partir d'une extension du SBM appelé MLVSBM. Plus précisément, supposons que les n_O organisations sont réparties dans Q_O blocs et que les n_I individus sont répartis dans Q_I blocs. Soient $Z^O = (Z_1^O, \ldots, Z_{n_O}^O)$ et $Z^I = (Z_1^I, \ldots, Z_{n_I}^I)$ tels que $Z_j^O = l$ si l'organisation j appartient au cluster l ($l \in \{1, \ldots, Q_O\}$) et $Z_i^I = k$ si l'individu i appartient au cluster k ($k \in \{1, \ldots, Q_I\}$). Sachant les blocs, nous supposons que les interactions entre organisations et entre individus sont indépendantes et distribuées comme suit :

$$\begin{aligned} \mathbb{P}(X_{jj'}^O = 1 | Z_j^O, Z_{j'}^O) &= \alpha^O_{Z_j^O Z_{j'}^O} \\ \mathbb{P}(X_{ii'}^I = 1 | Z_i^I, Z_{i'}^I) &= \alpha^I_{Z_i^I Z_{i'}^I}. \end{aligned}$$

Par conséquent, les blocs regroupent des nœuds partageant le même profil de connectivité. Afin de prendre en compte le fait que les organisations puissent structurer les comportements indivuels, nous supposons que l'appartenance au bloc des individus (Z^I) dépend du bloc des organisations (Z^O) auxquels ils sont affiliés. Plus précisément, nous posons :

$$\mathbb{P}(Z_i^I = k | Z_j^O, A_{ij} = 1) = \gamma_{kZ_j^O} \quad \forall i \in \{1, \ldots, n_I\} \quad \forall k \in \{1, \ldots, Q_I\}$$

où γ est une matrice de taille $Q_I \times Q_O$ tel que $\sum_{k=1}^{Q_I} \gamma_{kl} = 1$ pour tout $l \in \{1, \ldots, Q_O\}$. Les (Z_j^O) sont des variables aléatoires *i.i.d.* :

$$\mathbb{P}(Z_j^O = l) = \pi_l^O, \qquad \forall j \in \{1, \ldots, n_O\} \quad \forall l \in \{1, \ldots, Q_O\}$$

tel que $\sum_{l=1}^{Q_O} \pi_l^O = 1$.

Un petit réseau multiniveau est décrit dans la Figure 1.5.

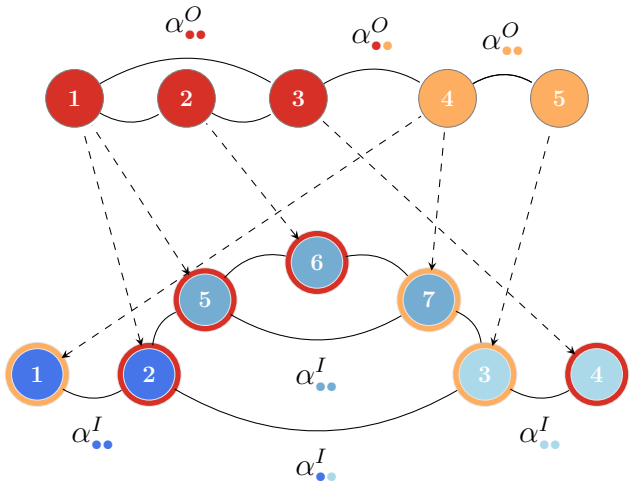


FIGURE 1.5 – Représentation d'un réseau multiniveau avec le niveau inter-organisationnel en haut et le niveau inter-individuel en bas.

Résultats théoriques Nous montrons l'identifiabilité du modèle et dérivons des conditions sur l'indépendance structurelle entre les niveaux en terme d'égalité entre paramètres. Cette proposition est fondamentale car elle permet de réécrire le modèle pour un réseau multiniveau comme le produit de deux SBMs indépendants, un pour chaque réseau, et de vérifier si l'hypothèse de dépendance multiniveau est adapté à ce réseau.

Proposition 1.1. *Dans le MLVSBM, les deux propositions sont équivalentes :*

1. *Z^I est indépendant de Z^O*
2. $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1, \dots, Q_O\}, \forall k \in \{1, \dots, Q_I\}$

   et implique :
3. *X^I et X^O sont indépendants.*

L'inférence du MLVSBM s'effectue via des méthodes variationnelles (cf. section 1.2.5) à travers un algorithme VEM adapté à ce modèle, tandis que nous développons un critère ICL pour sélectionner le nombre de blocs. La proposition 1.1 permet également d'utiliser l'ICL pour juger qu'il y a indépendance des deux niveaux si :

$$\max_{\{Q_I, Q_O\}} \mathrm{ICL}_{\mathrm{MLVSBM}}(Q_I, Q_O) \leq \max_{Q_I} \mathrm{ICL}^I_{\mathrm{SBM}}(Q_I) + \max_{Q_O} \mathrm{ICL}^O_{\mathrm{SBM}}(Q_O).$$

1.3.2. Structures communes d'une collection de réseau

Dans le chapitre 3, nous nous intéressons à une collection de réseaux unipartites que nous supposons être de même sorte (arêtes binaires ou valuées, dirigées ou non dirigées...). L'objectif de ce travail est de retrouver une strucutre méso-échelle commune entre les réseaux de cette collection et d'en déterminer la pertinence. En notant $\mathrm{SBM}_{n_m}(Q_m, \pi^m, \alpha^m)$, la loi d'un SBM à n_m nœuds, Q_m blocs et de paramètre (α^m, π^m), nous pouvons modéliser une collection de M réseaux, de

matrices d'adjacences $\mathbf{X} = \{X^1, \dots, X^M\}$, par M SBMs indépendants (separated SBM) de paramètres spécifiques à chaque réseau :

$$\mathbf{X} \sim \prod_{m=1}^{M} \mathrm{SBM}_{n_m}(Q_m, \pi^m, \alpha^m). \qquad (sepSBM)$$

Pour retrouver une structure commune, nous allons contraindre certains paramètres du SBM à être les mêmes d'un réseau à l'autre. Le modèle le plus contraint, appelé *iid-colSBM* suppose que chaque nœud de la collection a la même probabilité d'appartenir à chacun des blocs, et que les paramètres de la loi d'émission conditionnelle sont les mêmes :

$$\mathcal{X} \sim \prod_{m=1}^{M} \mathrm{SBM}_{n_m}(Q, \pi, \alpha). \qquad (iid\text{-}colSBM)$$

Nous introduisons deux mécanismes pour relâcher ces contraintes, le premier est d'autoriser les proportions de blocs à varier suivant les réseaux et de permettre aux blocs de n'être représentés que dans un sous-ensemble des réseaux. Cela permet de modéliser des structures imbriquées entre réseaux ou des structures qui se recoupent partiellement :

$$\mathbf{X} \sim \prod_{m=1}^{M} \mathrm{SBM}_{n_m}(Q, \pi^m, \alpha), \qquad (\pi\text{-}colSBM)$$

$\sum_{q \in \mathcal{Q}} \pi_q^m = 1$ avec $\pi_q \geq 0$.

Le second mécanisme est d'autoriser la densité du réseau à varier tout en gardant les mêmes rapports relatifs de connexions entre blocs, pour cela nous introduisons un paramètre de densité spécifique à chaque réseau δ^m :

$$\mathbf{X} \sim \prod_{m=1}^{M} \mathrm{SBM}_{n_m}(Q, \pi, \delta^m \alpha). \qquad (\delta\text{-}colSBM)$$

Enfin, nous proposons un dernier modèle regroupant les deux mécanismes :

$$\mathbf{X} \sim \prod_{m=1}^{M} \mathrm{SBM}_{n_m}(Q, \pi^m, \delta^m \alpha), \qquad (\delta\pi\text{-}colSBM)$$

$\sum_{q \in \mathcal{Q}} \pi_q^m = 1$ avec $\pi_q \geq 0$.

Chaque modèle repose sur un support que nous représentons sous la forme d'une matrice $S \in \{0,1\}^{M \times Q}$ indiquant les blocs autorisés pour chaque réseau :

$$S_{mq} = \mathbb{1}_{\pi_q^m > 0}, \quad \forall q = \{1, \dots Q\}, m = \{1, \dots, M\}.$$

Nous dérivons des conditions d'identifiabilité pour chacun des quatre *colSBM*s, celle du *sepSBM* étant immédiate. Nous proposons un algorithme variationnel pour l'inférence des paramètres du modèle et des clusterings. Pour la sélection du support, nous proposons une approximation du BIC via un critère ICL. Ce critère a de nombreuses autres utilités. Il permet de choisir entre les cinq modèles introduits ci-dessus, avec pour conséquence de donner une règle de décision pour juger de la pertinence de la structure commune. Cela permet également de proposer une partition des réseaux en regroupant ceux ayant une structure similaire.

1.3.3. Estimation de la robustesse de réseaux d'interactions écologiques bipartites

Dans le chapitre 4, nous nous intéressons à relier la robustesse, un indice communément utilisé en écologie pour quantifier la résilience d'un écosystème à la disparition d'espèces, à la structure du réseau représentant cet écosystème. Nous nous focalisons sur les réseaux d'interactions bipartites, où nous regardons l'influence des extinctions des espèces en ligne (extinctions primaires) sur les extinctions d'espèces en colonnes (extinctions secondaires). Pour cela, on se donne une loi sur les séquences d'extinctions primaires $\mathbb{S}$, la *statistique de robustesse* pour un réseau A de taille $n_r \times n_c$ sous cette loi d'extinction est donnée par

$$\overline{R}_{\mathbb{S}}(A) = \frac{1}{n_r}\sum_{m=0}^{n_r} R_{\mathbb{S}}(A,m) = \frac{1}{n_r}\sum_{m=0}^{n_r} \mathbb{E}_S\left[R(A,S,m)\right] \quad \text{with} \quad S \sim \mathbb{S}\,,$$

où

$$R(A,S,m) = 1 - \frac{1}{n_c}\sum_{j=1}^{n_c} \mathbf{1}_{\{\sum_{i=m+1}^{n_r} A_{S(i)j}=0\}}$$

est la proportion de nœuds en colonne restant connectés après m extinctions. Cette espérance, comme fonction de m, s'appelle la *fonction de robustesse* et est généralement approchée par une intégration Monte Carlo sur $\mathbb{S}$.

Afin de relier la structure du réseau à sa robustesse, nous proposons de mettre une loi $\mathbb{A}$ sur A et d'intégrer la robustesse sur le couple $(A,S) \sim (\mathbb{A},\mathbb{S})$. Si la loi du réseau est paramétrique, et que ces paramètres expliquent en partie la structure du réseau, nous pouvons alors relier la structure du réseau à sa robustesse.

Nous dérivons les premiers moments de la fonction de robustesse lorsque $\mathbb{S} = \mathbb{U}$ est une loi uniforme sur les permutations du groupe symétrique $\mathfrak{S}_{n_r}$, indépendante de la loi du réseau, pour A qui suit un SBM bipartite ($biSBM$) de paramètres $\boldsymbol{\theta} = (\pi,\rho,\delta)$, où δ est le paramètre de connectivité. En particulier, nous obtenons une expression analytique de l'espérance :

$$\mathbb{E}_{(A,S)}[R(A,S,m)] = 1 - \sum_{q=1}^{Q_c} \rho_q (1-\delta_{+q})^{n_r-m},$$

où $\delta_{+q} = \sum_{k=1}^{Q_r} \pi_k \delta_{kq}$, et de la variance $\mathbb{V}_A(\mathbb{E}_{S|A}[R(A,S,m)|A])$.

Afin d'obtenir des scénarios d'extinctions plus réalistes et de s'intéresser à des cas plus ou moins favorables à la robustesse de l'écosystème, nous obtenons également une expression de l'espérance de la robustesse suivant une loi des séquences d'extinctions primaires dépendantes des blocs du $biSBM$. Dans ce cas, la loi du couple $(A,S) \sim (\mathbb{A},\mathbb{B})$ n'est pas factorisable :

$$R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}(m) = 1 - \sum_{q=1}^{Q_c} \rho_q \sum_{n_1+\cdots+n_{Q_r}=n_r} \frac{n_r!}{n_1!\ldots n_{Q_r}!} \prod_{k=1}^{Q_r} \pi_k^{n_k} \left(1-\delta_{kq}\right)^{\min^+(n_k,\sum_{l\leq k} n_l - m)},$$

où $\min^+$ est la partie positive de la fonction minimum : $\min^+(x,y) = \max(0,\min(x,y))$.

Nous dérivons également des propriétés en terme de borne supérieure pour la robustesse ainsi que de monotonie par rapport au nombre de nœuds, du nombre d'extinctions et des paramètres. En particulier, à nombre de nœuds et densité $d = \sum_{(k,q)\in(\mathcal{Q}_r\times\mathcal{Q}_c)} \pi_k\rho_q\delta_{kq}$ fixés, l'ensemble des paramètres qui majorent l'espérance de la robustesse d'un biSBM pour $\mathbb{S} = \mathbb{U}$ comprend ceux d'un Erdős-Rényi (de robustesse $\mathbb{E}_{(A,S)}[R(A,S,m)] = 1-(1-d)^{nr-m}$). Et parmi cet ensemble, c'est le paramètre de l'Erdős-Rényi qui minimise la variance de la robustesse.

1.3.4. Package R

Les développements méthodologiques de la thèse ont été implémentés dans 3 packages.

- Le package `MLVSBM` (Chabert-Liddell, 2021a), disponible sur le CRAN (https://CRAN.R-project.org/package=MLVSBM), implémente les méthodes décrites au chapitre 2. Il permet la simulation et la prédiction d'arêtes de réseaux multiniveaux, ainsi que l'estimation des paramètres, des blocs latents et la sélection de modèle pour un MLVSBM. Le package repose sur de la programmation orientée objet à travers l'utilisation de classes `R6`. Une vignette illustrant son utilisation ainsi qu'une autre illustrant l'intêret de la modélisation multiniveau via une étude de simulation sont disponibles en appendice du chapitre 2. Ces vignettes sont reproduites à partir de celles disponibles sur le site dédié au package, qui propose une documentation complète : https://Chabert-Liddell.github.io/MLVSBM.
- Le package `colSBM` qui a servi à faire les simulations et les applications du chapitre 3 est disponible sur github (https://Chabert-Liddell.github.io/colSBM). Il repose également sur de la programmation orientée objet à travers l'utilisation de classes `R6`. Son utilisation est illustrée à travers une application à des données de réseaux de conseils en appendice du chapitre 3.
- Le package `robber` (Chabert-Liddell, 2021b), disponible sur le CRAN (https://CRAN.R-project.org/package=robber), propose différentes méthodes pour calculer la robustesse de réseaux écologiques bipartites. Il implémente entre autres les méthodes empiriques et par biSBM pour calculer la robustesse décrite au chapitre 4. Une démonstration du package à travers l'analyse et la comparaison de la robustesse d'une collection de réseaux décrivant différents types d'interactions (pollinisation, dispersion de graines, parasitisme) est disponible en appendice du chapitre 4. Une documentation détaillée et une vignette permettant de reproduire les analyses topologiques de l'article sont disponibles sur le site internet dédié https://Chabert-Liddell.github.io/robber

Chapter 2

A Stochastic Block Model Approach for the Analysis of Multilevel Networks: an Application to the Sociology of Organizations

Motivation Le travail qui suit est motivé par l'analyse des réseaux multiniveaux. L'idée émerge des travaux de Barbillon et al. (2017) qui traitent un réseau multiniveau comme un réseau multiplexe en considérant les interactions entre organisations comme des interactions entre individus via leurs organisations. Une approche qui n'est raisonnable que si le nombre d'individu par organisation est proche de 1. Pour pallier ce problème, il est nécessaire de développer un modèle qui prenne en compte la structure particulière de dépendance entre les niveaux des réseaux multiniveaux habituels. Ces travaux ont fait l'objet d'une collaboration avec le sociologue Emmanuel Lazega.

Résumé Un réseau multiniveau est défini comme la jonction de deux réseaux d'interaction, un niveau représentant les interactions entre individus et l'autre les interactions entre organisations. Les niveaux sont liés par une relation d'affiliation, chaque individu appartenant à une organisation unique. Un nouveau modèle à blocs stochastiques est proposé comme cadre probabiliste unifié, adapté aux réseaux multiniveaux. Ce modèle contient des blocs latents qui représentent de l'hétérogénéité dans les profils de connexion au sein de chaque niveau et qui introduisent des dépendances entre les niveaux. Les profils de connexion recherchés ne sont pas spécifiés a priori, ce qui rend cette approche flexible. Des méthodes variationnelles sont utilisées pour l'inférence du modèle et un critère de vraisemblance classifiante intégrée est développé pour choisir le nombre de blocs et aussi pour décider si les deux niveaux sont dépendants ou non. Une étude de simulation complète montre l'avantage de considérer cette approche, illustre la robustesse du clustering et met en évidence la fiabilité du critère utilisé pour la sélection du modèle. Cette approche est appliquée sur un ensemble de données sociologiques collectées lors de salons professionnels d'échange de programmes de télévision, le niveau inter-organisationnel étant le réseau économique entre les entreprises et le niveau inter-individuel étant le réseau informel entre leurs représentants. Elle apporte une représentation synthétique des deux réseaux en démêlant leur structure entrelacée et confirme la *coopétition* en jeu.

Diffusion Le contenu de ce chapitre ainsi que les quatre premières annexes ont fait l'objet d'un article (Chabert-Liddell et al., 2021) publié dans le journal Computational Statistics & Data Analysis. Un package `R` nommé `MLVSBM`, disponible sur le CRAN (Chabert-Liddell, 2021a), dont la documentation est disponible à l'adresse suivante https://Chabert-Liddell.github.io/MLVSBM/, permet d'appliquer les méthodes développées ci-dessous. Ce package fait l'objet des annexes 2.F montrant un tutoriel et un exemple d'application illustrant l'intérêt de l'approche multiniveau.

Notations for this chapter

n_I	The number of individuals
n_O	The number of organizations
X^I	The inter-individual level, an $n_I \times n_I$ adjacency matrix,
X^O	The inter-organizational level, an $n_O \times n_O$ adjacency matrix
A	The affiliation relationship, a $n_I \times n_O$ matrix
$\mathbf{X}$	The observed variables $\{X^I, X^O\}$
Z^I	The block membership of individuals
Z^O	The block membership of organizations
$\mathbf{Z}$	The latent variables $\{Z^I, Z^O\}$
Q_I	The number of individual blocks
Q_O	The number of organizational blocks
α^I	The inter-individual connectivity parameters, a $Q_I \times Q_I$ matrix
α^O	The inter-organizational connectivity parameters, a $Q_O \times Q_O$ matrix
γ	The mixture parameters of individuals, a $Q_I \times Q_O$ matrix
π^O	The mixture parameters of organizations
θ	The set of parameters $(\pi^O, \gamma, \alpha^O, \alpha^I)$
τ^I	The variational parameters for individual block memberships
τ^I	The variational parameters for organizational block memberships

Abstract A multilevel network is defined as the junction of two interaction networks, one level representing the interactions between individuals and the other the interactions between organizations. The levels are linked by an affiliation relationship, each individual belonging to a unique organization. A new Stochastic Block Model is proposed as a unified probabilistic framework tailored for multilevel networks. This model contains latent blocks accounting for heterogeneity in the patterns of connection within each level and introducing dependencies between the levels. The sought connection patterns are not specified a priori which makes this approach flexible. Variational methods are used for the model inference and an Integrated Classified Likelihood criterion is developed for choosing the number of blocks and also for deciding whether the two levels are dependent or not. A comprehensive simulation study exhibits the benefit of considering this approach, illustrates the robustness of the clustering and highlights the reliability of the criterion used for model selection. This approach is applied on a sociological dataset collected during a television program trade fair, the inter-organizational level being the economic network between companies and the inter-individual level being the informal network between their representatives. It brings a synthetic representation of the two networks unraveling their intertwined structure and confirms the *coopetition* at stake.

2.1. Introduction

The statistical analysis of network data has been a hot topic for the last decade. The last few years witnessed a growing interest for multilayer networks (see Kivelä et al., 2014; Bianconi, 2018; Giordano et al., 2019). A particular case of multilayer networks are multilevel networks where each level is a layer and an affiliation relationship represents the inter-layer. Multilevel networks are used across many fields such as sociology (Lazega and Snijders, 2015) or environmental science (Hileman and Lubell, 2018). In particular they arise in the sociology of organizations and collective action when willing to study jointly the social network of individuals and the interaction network of organizations the individuals belong to. Indeed, the individuals not only interact with each others but are also members of interacting organizations. This approach is quite generic in the social sciences and all the phenomena of *coopetition* and the maintenance of social inequalities can fall within the scope of this approach (Lazega and Jourda, 2016). It is also gaining attention as a way to articulate social network analysis and the life course studies (Vacchiano et al., 2020). Following Lazega and Snijders (2015), one might think that these two types of interactions (between individuals and between organizations) are interdependent, the individuals shaping their organizations and the organizations having an influence on the individuals. We aim to propose a statistical model for multilevel networks in order to understand how the two levels are intertwined and how one level impacts the other.

In what follows, a multilevel network is defined as the collection of an inter-individual network, an inter-organizational network and the affiliation of the individuals to

the organizations. Besides, we assume that the individuals belong to a unique organization. Such a dataset is studied by Lazega et al. (2008), some researchers in cancerology being the individuals and their laboratories the organizations. Brailly et al. (2016) deal with another dataset concerned with the economic network of audiovisual firms and the informal network of their sales representatives during a trade fair. This latter dataset will be analyzed in this paper.

In the last years, the Stochastic Block Model (SBM developed by Holland et al., 1983; Snijders and Nowicki, 1997) has become a popular tool to model the heterogeneity of connection in a network, assuming that the actors at stake are divided into blocks (clusters) and that the members of a same block share a similar profile of connectivity. Compared to other graph clustering methods such as modularity maximization, hierarchical clustering or spectral clustering (see Kolaczyk, 2009, and references therein), the SBM is a generative model, it shares with the generalized blockmodeling (Doreian et al., 2005) that they can both fit to a wide range of topologies since they gather into blocks the nodes that are structurally equivalent. However, contrary to the generalized blockmodeling which seeks a pre-specified structure in the network with given ideal blocks, the SBM is agnostic and is aimed to unravel any kind of block structure which may shape the data. This includes but is not restricted to the detection of assortative communities where the probability of connection within a block is higher than the probability of connection between the blocks. Moreover, the probabilistic generative model allows the modeler to have a unified framework for model selection and natural extensions such as dealing with non binary dyads and link prediction. The SBMs have been extended to particular types of multilayer networks : Barbillon et al. (2017) propose an SBM for multiplex networks and Matias and Miele (2017) an SBM for time-evolving networks. In this paper, we propose an SBM suited to multilevel networks (MLVSBM).

Our contribution In a few words, we model the heterogeneity in the inter-individual and inter-organizational connections by introducing blocks of individuals and blocks of organizations, the blocks containing homogeneous groups of actors (individuals or organizations) with respect to their connectivity. The two levels are assumed to be interdependent through their latent blocks. More specifically, the latent blocks of the inter-individual level depend on the latent blocks of the inter-organizational level and the affiliation. This bi-clustering approach allows us to determine how groups of organizations influence the connectivity patterns of their individuals. Note that the hierarchical model does not assume a causal effect of the blocks of organizations on the blocks of individuals but an interdependence between the two sets of blocks.

Due to the latent variables, the estimation of the parameters is a complex task. We resort to a variational version of the Expectation-Maximization (EM) algorithm. For the SBM, the variational approach (Jordan et al., 1999; Blei et al., 2017) has proven its efficiency for deriving maximum likelihood estimates (Daudin et al., 2008; Mariadassou et al., 2010; Barbillon et al., 2017) and for Bayesian inference (Latouche et al., 2012; Côme and Latouche, 2015). In the latent block model which

is suited for bipartite network, the variational estimates have also been successfully applied (Govaert and Nadif, 2008). In this paper, we obtain approximate maximum likelihood estimates by an ad-hoc version of the variational EM algorithm.

Another important task is the choice of the number of blocks. We propose an adapted version of the Integrated Complete Likelihood (ICL) criterion. First developed by Biernacki et al. (2000) for mixture models as an alternative to the Bayesian Information Criterion (BIC), it was then adapted by Daudin et al. (2008) to the SBM. The ICL has since illustrated its efficiency and relevance for various SBMs and their extensions such as multiplex network (Barbillon et al., 2017), dynamic SBM (Matias and Miele, 2017) or degree corrected SBM (Yan, 2016). A further reference for dynamic SBMs is Bartolucci et al. (2018). Besides, a critical issue in sociology is to verify the multilevel interdependence hypothesis in a multilevel network, i.e. if the two levels (inter-individual and inter-organizational) should be analyzed jointly or if a separate analysis is sufficient. We thus propose a criterion to decide whether the two levels are independent or not.

Related works The term *multilevel network* arises in the statistical literature for a wide variety of complex networks. For instance, Zijlstra et al. (2006) adapt the p2-model to handle multiple observations of a network, Sweet et al. (2014) extend the Mixed Membership Stochastic Block Model (Airoldi et al., 2008) to the hierarchical network model framework (Sweet et al., 2013) for the same type of data. Snijders (2017) discusses the use of the stochastic actor-oriented model (Snijders, 2001) for temporal and multivariate networks.

When dealing with the multilevel networks we defined before, Wang et al. (2013) adopt an exponential random graph model (ERGM) strategy that is used in applications across many fields such as environmental science (Hileman and Lubell, 2018) or sociology (Lazega and Snijders, 2015, chapter 10-11, 13-14). When focusing on a clustering approach, Žiberna (2014) develops three general approaches for blockmodeling multilevel networks. First, the separate analysis consists in clustering the levels separately or using the clustering of one level on the other. Second, the conversion approach converts the level of the organizations into a new kind of interaction between individuals, the interactions are then aggregated into a single layer network; this is close to the approach taken by Barbillon et al. (2017) who transform the inter-organizational network into an inter-individual network thus adopting a multiplex network approach (the individuals interconnect directly or through the organizations they belong to). The third approach is called the true multilevel approach and is the closest to the one we propose on this paper.

Žiberna (2014) and the extensions in Žiberna (2019, 2020) to a more general set of multilayer networks (called linked networks) use a generalized blockmodeling framework (Doreian et al., 2005). Contrary to this deterministic approach, we resort to a probabilistic generative model for all the reasons stated above. The MLVSBM additionally provides us with a natural criterion for detecting the interdependence between the two levels. Furthermore, we explicitly take into account the constraint of having a unique affiliation per individual inherent to these multilevel datasets

and do not consider the affiliation as a bipartite network.

Also, note that the multiplex SBM approach applied to a multilevel network suggested by Barbillon et al. (2017) is only applicable when the numbers of individuals and organizations are close. Indeed it requires to duplicate the data of the inter-organizational level to fit the size of the inter-individual level. Furthermore, it only provides a clustering of the individuals and not two clusterings, one of the individuals and one of the organizations. In contrast, our MLVSBM does not need to transform the data into a multiplex network and is able to obtain a clustering of the nodes within each level.

If we release the constraint of the unique affiliation, then the inter-level can be modeled by a latent block model and we obtain a particular case of the multipartite SBM of Bar-Hen et al. (2020). However, the interactions between individuals and organizations are considered at the same level as the affiliations, and the clustering might be strongly influenced by the number of individuals in each organization.

Finally, our work is also different from the SBM with edges covariates (Mariadassou et al., 2010) with the individuals as nodes and the inter-organizational network as edges covariates. Indeed, in that case, the clustering obtained for the individuals is the remaining structure of the inter-individual level once the effect of the covariates has been taken into account. In addition, this model does not provide a clustering of the organizations.

Outline of the paper The paper is organized as follows. The SBM adapted to multilevel networks (MLVSBM) is defined in Section 2.2. We also give conditions guaranteeing the independence between levels and the identifiability of the parameters. The inference strategy and the model selection criterion are provided in Section 2.3. The proof of the independence between levels, of the identifiability and the details on the variational EM and the ICL criterion are postponed to the Appendix sections. In Section 2.4, we present an extensive simulation study illustrating the relevance of our inference method, model selection criterion and procedure. Section 2.5 is dedicated to the analysis of a sociological dataset by our MLVSBM. Finally we discuss our contribution and future works in Section 2.6.

2.2. A multilevel stochastic block model

Dataset Let us consider n_I individuals involved in n_O organizations. We encode the networks into adjacency matrices as follows. Let X^I be the binary $n_I \times n_I$ matrix representing the inter-individual network. X^I is such that : $\forall (i, i') \in \{1, \dots, n_I\}^2$:

$$X^I_{ii'} = \begin{cases} 1 & \text{if there is an interaction from individual } i \text{ to individual } i', \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

X^O is the binary $n_O \times n_O$ matrix representing the inter-organizational network, $\forall (j, j') \in \{1, \dots, n_O\}^2$:

$$X^O_{jj'} = \begin{cases} 1 & \text{if there is an interaction from organization } j \text{ to organization } j', \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Remark. *In general, no self-loop are considered in the network, thus the interactions are defined for $i \neq i'$ and $j \neq j'$. Moreover, if the interactions are undirected then*

$$X^I_{ii'} = X^I_{i'i} \quad \forall (i, i') \qquad \textit{or/and} \qquad X^O_{jj'} = X^O_{j'j} \quad \forall (j, j').$$

In what follows, we present the methodology for undirected networks. However, all the results can be adapted to directed networks without any difficulty.

Let A be the affiliation matrix. A is a $n_I \times n_O$ matrix such that:

$$A_{ij} = \begin{cases} 1 & \text{if individual i belongs to organization j,} \\ 0 & \text{otherwise} \end{cases}.$$

A is such that $\forall i = 1, \dots, n_I$, $\sum_{j=1}^{n_O} A_{ij} = 1$ since we assume that any individual belongs to a unique organization. A synthetic view of a generic dataset is provided in Figure 2.1.

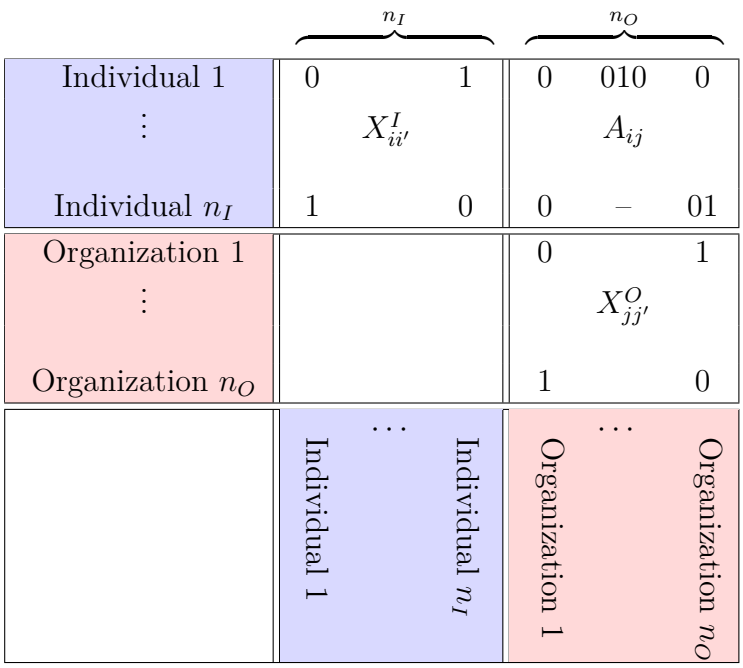


Figure 2.1 – Matrix representation of a multilevel network

We propose a joint modeling of the inter-individual and inter-organizational networks based on an extension of the SBM. More precisely, assume that the n_O organizations are divided into Q_O blocks and that the n_I individuals are divided into Q_I blocks. Let $Z^O = (Z^O_1, \dots, Z^O_{n_O})$ and $Z^I = (Z^I_1, \dots, Z^I_{n_I})$ be such that $Z^O_j = l$ if organization

j belongs to block l ($l \in \{1, \ldots, Q_O\}$) and $Z_i^I = k$ if individual i belongs to block k ($k \in \{1, \ldots, Q_I\}$).

Given these clusterings, we assume that the interactions between organizations and the interactions between individuals are independent and distributed as follows:

$$\begin{aligned} \mathbb{P}(X_{jj'}^O = 1 | Z_j^O, Z_{j'}^O) &= \alpha_{Z_j^O Z_{j'}^O}^O, \\ \mathbb{P}(X_{ii'}^I = 1 | Z_i^I, Z_{i'}^I) &= \alpha_{Z_i^I Z_{i'}^I}^I. \end{aligned} \tag{2.3}$$

As a consequence, the blocks gather nodes (blocks of individuals on the one hand and blocks of organizations on the other hand) sharing the same profiles of connectivity. In order to take into account the fact that organizations may shape the individual behaviors, we assume that the memberships of the individuals (Z^I) depend on the blocks of the organizations (Z^O) they are affiliated to. More precisely, we set:

$$\mathbb{P}(Z_i^I = k | Z_j^O, A_{ij} = 1) = \gamma_{kZ_j^O} \quad \forall i \in \{1, \ldots, n_I\} \quad \forall k \in \{1, \ldots, Q_I\}, \tag{2.4}$$

where γ is a $Q_I \times Q_O$ matrix such that $\sum_{k=1}^{Q_I} \gamma_{kl} = 1 \ \forall l \in \{1, \ldots, Q_O\}$. The (Z_j^O) are assumed to be independent random variables distributed as

$$\mathbb{P}(Z_j^O = l) = \pi_l^O, \qquad \forall j \in \{1, \ldots, n_O\} \quad \forall l \in \{1, \ldots, Q_O\}, \tag{2.5}$$

with $\sum_{l=1}^{Q_O} \pi_l^O = 1$.
Equations (2.4) and (2.5) state that the clustering of an individual is not completely driven by his/her behavior but is also shaped by the clustering of the organization he/she belongs to. In particular, if $Q_O = Q_I$ and γ is equal to the identity matrix (up to a reordering of the rows) then, the clustering of the individuals is completely determined by the clustering of the organizations. At the opposite, if all the columns of γ are equal, then the clustering of the individuals is independent on the clustering of the organizations. This point will be developed hereafter.

Equations (2.3), (2.4) and (2.5) define a joint modeling of X^I and X^O. In what follows, we set $\theta = \{\pi^O, \gamma, \alpha^O, \alpha^I\}$ the vector of the unknown parameters, $\mathbf{X} = \{X^I, X^O\}$ are the observed variables and $\mathbf{Z} = \{Z^I, Z^O\}$ the latent variables. The DAG of the MLVSBM is plotted in Figure 2.2. An illustration of the MLVSBM for a small multilevel network is represented in Figure 2.3.

Likelihood From Equations (2.3), (2.4) and (2.5), we derive the complete log-likelihood for an undirected MLVSBM:

$$\begin{aligned} \log \ell_\theta \left(X^I, X^O, \mathbf{Z} | A\right) &= \log \ell_{\pi^O}(Z^O) + \log \ell_\gamma(Z^I | Z^O, A) + \log \ell_{\alpha^I}(X^I | Z^I) + \log \ell_{\alpha^O}(X^O | Z^O) \\ &= \sum_{j,l} \mathbb{1}_{Z_j^O = l} \log \pi_l^O + \sum_{i,k} \mathbb{1}_{Z_i^I = k} \sum_{j,l} A_{ij} \mathbb{1}_{Z_j^O = l} \log \gamma_{kl} \\ &+ \frac{1}{2} \sum_{i' \neq i} \sum_{k,k'} \mathbb{1}_{Z_i^I = k} \mathbb{1}_{Z_{i'}^I = k'} \log \phi(X_{ii'}^I, \alpha_{kk'}^I) + \frac{1}{2} \sum_{j' \neq j} \sum_{l,l'} \mathbb{1}_{Z_j^O = l} \mathbb{1}_{Z_{j'}^O = l'} \log \phi(X_{jj'}^O, \alpha_{ll'}^O), \end{aligned} \tag{2.6}$$

where $\phi(x, a) = a^x (1 - a)^{1-x}$.

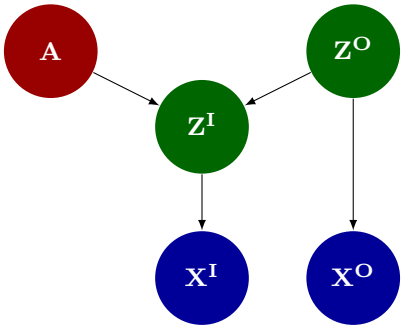


Figure 2.2 – DAG of the stochastic block model for multilevel network (MLVSBM)

Remark. *Note that the factors* $1/2$ *in Equation* (2.6) *derive from the fact that we consider undirected networks. If one or both of the networks are directed, then the corresponding* $1/2$ *disappears.*

The log-likelihood of the observations $\ell_\theta(\mathbf{X}|A)$ is obtained by integrating out the latent variables $\mathbf{Z}$ in Equation 2.6. As soon as n_O, n_I, Q_O, or Q_I increase, this summation over all the possible clusterings Z^I and Z^O cannot be performed within a reasonable computational time. As a consequence, we will resort to the variational EM algorithm to maximize this likelihood (see Section 2.3).

Independence We now derive conditions for the structural independence between levels in terms of parameters equality.

Proposition 2.1. *In the MLVSBM, the two following properties are equivalent:*

1. Z^I *is independent on* Z^O,
2. $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1, \ldots, Q_O\}$,

and imply that:

3. X^I *and* X^O *are independent.*

This proposition is proved in 2.A. Proposition 2.1 can be interpreted as follows: in the case where the clustering of the individuals does not depend on the clustering of the organizations, all column vectors of γ are identical. Hence, under this restriction on γ, the model for multilevel network can be rewritten as the product of two independent SBMs, one for each level. Conversely, in the case of a strong dependence between the levels, each column of γ will have one coefficient close to one, the others being close to 0. Therefore, the individuals affiliated to organizations belonging to the same block of organizations will be affiliated to one block of individuals. Even if the γ's imply a dependent relationship between the two levels, the connections of the corresponding blocks at the two levels may have different connectivity patterns since there is no constraint on the corresponding connection parameters α^O and α^I.

Identifiability The identifiability conditions for the MLVSBM are given in the following proposition.

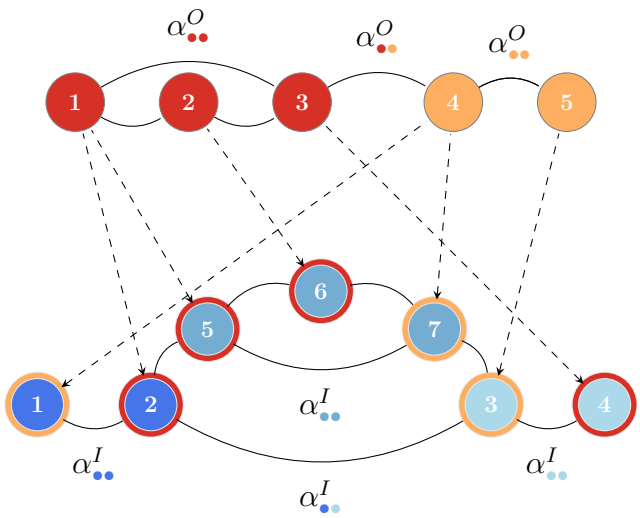


Figure 2.3 – MLVSBM with inter-organizational level on the top and inter-individual level on the bottom. The various shades of blue depict the clustering of the individuals and the various shades of red depict the clustering of the organizations. The parameters α over the plain links between nodes are the probabilities of connections given the nodes colors (clustering/blocks). The outer circles around the nodes of the individuals represent the blocks of the organizations they are affiliated to. The dashed links stand for the affiliations.

Proposition 2.2. *The MLVSBM is identifiable up to label switching under the following assumptions:*

$\mathcal{A}1$. *All coefficients of* $\alpha^I \cdot \gamma \cdot \pi^O$ *are distinct and all coefficients of* $\alpha^O \cdot \pi^O$ *are distinct.*

$\mathcal{A}2$. $n_I \geq 2Q_I$ *and* $n_O \geq \max(2Q_O, Q_O + Q_I - 1)$.

$\mathcal{A}3$. *At least* $2Q_I$ *organizations contain one individual or more.*

The set of parameters that does not verify assumption $\mathcal{A}1$ has null Lebesgue measure.

Assumption $\mathcal{A}2$ is very weak in practice. Assumption $\mathcal{A}3$, on the affiliation, means that at least some organizations must not be empty and enough individuals belong to different organizations. The proof of this proposition is provided in 2.B and results from an extension of the proof given in Celisse et al. (2012).

2.3. Statistical Inference

We now present a maximum likelihood procedure and a criterion for model selection.

2.3.1. Variational method for maximum likelihood estimation

As said before, $\ell_\theta(\mathbf{X}|A)$ is obtained by integrating out the latent variables $\mathbf{Z}$ in the complete data likelihood (2.6). However, this calculus becomes not computationally tractable as the numbers of nodes and blocks increase.

The Expectation-Maximization algorithm (EM) (Dempster et al., 1977) is a popular solution to maximize the likelihood of models with latent variables. However it requires the computation of $\mathbb{P}_\theta(\mathbf{Z}|\mathbf{X}, A)$ which is also not tractable in our case. The variational version of the EM algorithm is a powerful solution for such cases. It was first used for the SBM by Daudin et al. (2008).

In a few words, the variational EM algorithm maximizes the so-called variational bound i.e. a lower bound of the log-likelihood denoted $\mathcal{J}_\theta(\mathcal{R}(\mathbf{Z}|A))$ and defined as follows:

$$\begin{aligned}\mathcal{J}_\theta(\mathcal{R}(\mathbf{Z}|A)) &:= \mathbb{E}_{\mathcal{R}}\left[\ell_\theta\left(\mathbf{Z}, \mathbf{X}|A\right)\right] + \mathcal{H}\left(\mathcal{R}(\mathbf{Z}|A)\right) \\ &= \ell_\theta\left(\mathbf{X}|A\right) - \mathrm{KL}\left(\mathcal{R}(\mathbf{Z}|A)\|\mathbb{P}_\theta(\mathbf{Z}|\mathbf{X}, A)\right) \leq \ell_\theta(\mathbf{X}|A),\end{aligned} \tag{2.7}$$

where KL is the Kullback-Leibler divergence, $\mathcal{H}$ is the Shannon entropy: $\mathcal{H}(P) = \mathbb{E}_P[-\log(P)]$ and $\mathcal{R}(\mathbf{Z}|A)$ is an approximation of the true distribution $\mathbb{P}_\theta(\mathbf{Z}|\mathbf{X}, A)$. In our context, and following Daudin et al. (2008), we propose to choose $\mathcal{R}(\mathbf{Z}|A)$ in a family of factorized distributions, resulting into a mean field approximation $\mathcal{R}(\mathbf{Z}|A)$ defined as:

$$\mathcal{R}(\mathbf{Z}|A) = \prod_{i=1}^{n_I}\prod_{k=1}^{Q_I}(\tau_{ik}^I)^{\mathbb{1}_{Z_i^I=k}}\prod_{j=1}^{n_O}\prod_{l=1}^{Q_O}(\tau_{jl}^O)^{\mathbb{1}_{Z_j^O=l}}, \tag{2.8}$$

where $\tau_{ik}^I = \mathbb{P}_{\mathcal{R}}(Z_i^I = k)$ and $\tau_{jl}^O = \mathbb{P}_{\mathcal{R}}(Z_j^O = l)$.
Inputting Equations (2.6) and (2.8) into Equation (2.7), the variational bound for the MLVSBM can be written as follows:

$$\begin{aligned}\mathcal{J}_\theta(\mathcal{R}(\mathbf{Z}|A)) &= \sum_{j,l}\tau_{jl}^O\log\pi_l^O + \sum_{i,k}\tau_{ik}^I\sum_{j,l}A_{ij}\tau_{jl}^O\log\gamma_{kl} \\ &+ \frac{1}{2}\sum_{i'\neq i}\sum_{k,k'}\tau_{ik}^I\tau_{i'k'}^I\log\phi\left(X_{ii'}^I, \alpha_{kk'}^I\right) + \frac{1}{2}\sum_{j'\neq j}\sum_{l,l'}\tau_{jl}^O\tau_{j'l'}^O\log\phi\left(X_{jj'}^O\alpha_{ll'}^O\right) \\ &- \sum_{i,k}\tau_{ik}^I\log\tau_{ik}^I - \sum_{j,l}\tau_{jl}^O\log\tau_{jl}^O.\end{aligned}$$

The variational EM algorithm consists in iterating two steps. Step VE maximizes the variational bound with respect to the parameters of the approximate distribution defined in Equation (2.8). This is equivalent to minimizing the Kullbach-Leibler divergence term. Step M maximizes the variational bound with respect to the model parameters θ. The procedure is given in Algorithm 2 and details of the calculus and algorithm are developed in 2.C. Algorithm 2 can be slightly modified to handle missing data (dyads which are not observed in any of the two levels) by summing up on observed dyads only. An interesting feature of the MLVSBM is to make use of one level to help the prediction of missing dyads of the other level.

Remark. *Although the family of the variational distributions does not consider the affiliation matrix A, the minimization of the Kullback-Leibler divergence between the variational distribution and $\mathbb{P}_\theta(\mathbf{Z}|\mathbf{X}, A)$ induces an indirect dependence on A in the variational distribution. One may consider more complex distributions but the simulation studies show that the inference algorithm is able to retrieve properly the dependence between the Z^Is and the Z^Os in this family of distributions.*

Algorithm 2: Variational EM algorithm

Data: $\{\mathbf{X}, \mathbf{Z}, A\}$, a multilevel network with an initial clustering of size (Q_I, Q_O).

Procedure:

- Set $\{\tau^I, \tau^O\}$ from the initial clustering.

while $\mathcal{J}_\theta(\mathcal{R}(\mathbf{Z}|A))$ *is increasing* **do**

- **M step** compute

$$\theta^{(t+1)} = \arg\max_\theta \mathcal{J}_\theta(\mathcal{R}^{(t+1)}(Z^I, Z^O|A)),$$

by updating the model parameters as follows:

$$\widehat{\pi_l^O} = \frac{1}{n_O}\sum_j \widehat{\tau_{jl}^O} \qquad \widehat{\alpha_{ll'}^O} = \frac{\sum_{j'\neq j}\widehat{\tau_{jl}^O} X_{jj'}^O \widehat{\tau_{j'l'}^O}}{\sum_{j'\neq j}\widehat{\tau_{jl}^I}\widehat{\tau_{j'l'}^I}}$$

$$\widehat{\gamma}_{kl} = \frac{\sum_{i,j}\widehat{\tau_{ik}^I} A_{ij}\widehat{\tau_{jl}^O}}{\sum_{i,j} A_{ij}\widehat{\tau_{jl}^O}} \qquad \widehat{\alpha_{kk'}^I} = \frac{\sum_{i'\neq i}\widehat{\tau_{ik}^I} X_{ii'}^I \widehat{\tau_{i'k'}^I}}{\sum_{i'\neq i}\widehat{\tau_{ik}^I}\widehat{\tau_{i'k'}^I}}.$$

- **VE step** compute

$$\{\tau^I, \tau^O\}^{(t+1)} = \arg\max_{\tau^I,\tau^O} \mathcal{J}_{\theta^{(t)}}(\mathcal{R}(Z^I, Z^O|A))$$

by updating the variational parameters with the following fixed points relationships:

$$\widehat{\tau_{jl}^O} \propto \pi_l^O \prod_{i,k}\gamma_{kl}^{A_{il}\widehat{\tau_{ik}^I}} \prod_{j'\neq j}\prod_{l'}\phi(X_{jj'}^O, \alpha_{ll'}^O)^{\widehat{\tau_{j'l'}^O}}$$

$$\widehat{\tau_{ik}^I} \propto \prod_{j,l}\gamma_{kl}^{A_{il}\widehat{\tau_{jl}^O}} \prod_{i'\neq i}\prod_{k'}\phi(X_{ii'}^I, \alpha_{kk'}^I)^{\widehat{\tau_{i'k'}^I}}.$$

return $\mathcal{J}_\theta(\mathcal{R}(\mathbf{Z}|A))$, $\widehat{\theta}$ *and* $\{\widehat{\tau^I}, \widehat{\tau^O}\}$

2.3.2. Model selection

Selection of the number of blocks

Following Biernacki et al. (2000) and Daudin et al. (2008), we propose a model selection criterion to choose the unknown number of blocks Q_I and Q_O. The ICL criterion is an integrated version of BIC applied to the complete likelihood. In other words, it is an asymptotic approximation of the complete likelihood integrated over its parameters and latent variables, it values both goodness of fit and classification sharpness (Mariadassou et al., 2010).

Our criterion is equal to:

$$\mathrm{ICL}_{\mathrm{MLVSBM}}(Q_I, Q_O) = \log \ell_{\widehat{\theta}}(X^I, X^O, \widehat{Z^I}, \widehat{Z^O} | A, Q_I, Q_O) - \mathrm{pen}_{\mathrm{MLVSBM}}(Q_I, Q_O), \tag{2.9}$$

where

$$\begin{aligned}\mathrm{pen}_{\mathrm{MLVSBM}}(Q_I, Q_O) =& \frac{1}{2}\frac{Q_I(Q_I+1)}{2}\log\frac{n_I(n_I-1)}{2} + \frac{Q_O(Q_I-1)}{2}\log n_I + \\ & \frac{1}{2}\frac{Q_O(Q_O+1)}{2}\log\frac{n_O(n_O-1)}{2} + \frac{Q_O-1}{2}\log n_O,\end{aligned} \tag{2.10}$$

where $\widehat{Z^O}$ and $\widehat{Z^I}$ are the imputed latent variables using the maximum a posteriori (MAP) of $\mathbb{P}_{\hat{\theta}}(\mathbf{Z}|\mathbf{X}, A; Q_I, Q_O)$. The calculus is provided in 2.D. As for the variational inference, $\mathbb{P}_{\hat{\theta}}(\mathbf{Z}|\mathbf{X}, A; Q_I, Q_O)$ is unknown and, in practice, we replace it by its mean-field approximation $\mathcal{R}_{\hat{\theta}}(\mathbf{Z}|A; Q_I, Q_O)$.

Remark. *Once again, note that the penalty* (2.10) *is adapted to undirected networks. For instance, the term* $\frac{Q_I(Q_I+1)}{2}\log\frac{n_I(n_I-1)}{2}$ *would become* $Q_I^2 \log n_I(n_I-1)$ *if* X^I *were not symmetric.*

Remark. *We recall that the penalty of the ICL for a (unilevel) SBM is given by*

$$\mathrm{pen}_{\mathrm{SBM}}(Q) = \frac{1}{2}\frac{Q(Q+1)}{2}\log\frac{n(n-1)}{2} + \frac{Q-1}{2}\log n. \tag{2.11}$$

The penalty term in Equation (2.10) for the inter-organizational level is the same as the one given in Equation (2.11). For the inter-individual network, the factor in front of $\log n_I$ is $Q_O(Q_I-1)$ instead of $Q_I - 1$ for the SBM as in Equation (2.11), that is the penalty term which corresponds to the degree of freedom of γ.

Determining the independence between levels

The ICL criterion can also be used to assess whether the two levels of interactions are independent or not. If γ is forced to have all its columns identical, then the penalty term on γ becomes $\frac{1}{2}(Q_I-1)\log n_I$ and, as a consequence:

$$\mathrm{ICL}_{\mathrm{Ind}}(Q_I, Q_O) = \mathrm{ICL}^I_{\mathrm{SBM}}(Q_I) + \mathrm{ICL}^O_{\mathrm{SBM}}(Q_O). \tag{2.12}$$

The ICL criterion favors independence if

$$\max_{\{Q_I,Q_O\}} \mathrm{ICL}_{\mathrm{MLVSBM}}(Q_I,Q_O) \leq \max_{Q_I} \mathrm{ICL}^I_{\mathrm{SBM}}(Q_I) + \max_{Q_O} \mathrm{ICL}^O_{\mathrm{SBM}}(Q_O).$$

If this is the case, then the gain in terms of likelihood does not compensate the gain $\frac{1}{2}(Q_O-1)(Q_I-1)\log n_I$ in the penalty. This criterion focuses on the dependence between levels given by the inter-level.

Remark. *If $Q_I = 1$ or $Q_O = 1$, the MLVSBM is the product of two independents SBM, as such* $\mathrm{ICL}_{\mathrm{Ind}}(Q_I,Q_O) = \mathrm{ICL}_{\mathrm{MLVSBM}}(Q_I,Q_O)$.

Procedure for model selection

We now provide a procedure for model selection which seeks for the optimal number of blocks at a reasonable cost. As a by-product, it states whether the two levels are independent or not.

The practical choice of the model and the estimation of its parameters are computationally intensive tasks. Indeed, we should compare all the possible models – one model corresponding to a given (Q_I,Q_O) – through the ICL criterion. Furthermore, for each model, the variational EM algorithm should be initialized at a large number of initialization points (due to its sensitivity to the starting point), resulting in an unreasonable computational cost. Instead, we propose to adopt a stepwise strategy, resulting in a faster exploration of the model space, combined with efficient initializations of the variational EM algorithm. The procedure we suggest is given in Algorithm 3.

Algorithm 3: Model selection algorithm

Data: $\{X^I, X^O, A\}$, a multilevel network.

Procedure:

- Infer independent SBMs on X^I and X^O for a respective range of Q_I and Q_O. Deduce

$$\widehat{Q_I}^{\mathrm{Ind}} = \arg\max_{Q_I} \mathrm{ICL}^I_{\mathrm{SBM}}(Q_I) \quad \text{and} \quad \widehat{Q_O}^{\mathrm{Ind}} = \arg\max_{Q_O} \mathrm{ICL}^O_{\mathrm{SBM}}(Q_O).$$

  Compute $\mathrm{ICL}_{\mathrm{Ind}} = \mathrm{ICL}^I_{\mathrm{SBM}}(\widehat{Q_I}^{\mathrm{Ind}}) + \mathrm{ICL}^O_{\mathrm{SBM}}(\widehat{Q_O}^{\mathrm{Ind}})$.

- Start at $Q_I = \widehat{Q_I}^{\mathrm{Ind}}$ and $Q_O = \widehat{Q_O}^{\mathrm{Ind}}$.

while *ICL is increasing* **do**

  - Fit an MLVSBM on every model of size $(Q_I \pm 1, Q_O \pm 1)$ initialized by merging 2 blocks or splitting a block with hierarchical clustering.
  - Among all estimated models, keep the one with the highest ICL.

return $(\widehat{Q_I}, \widehat{Q_O}) = \arg\max \mathrm{ICL}(Q_I,Q_O)$, $\widehat{\theta}_{(\widehat{Q_I},\widehat{Q_O})}$ *and* $\widehat{\mathbf{Z}}$.

Each step of the algorithm requires $O(\max\{Q_I, Q_O\}^2)$ variational EM algorithms which converge in a few iterations as a result of the local initialization. Inferring an independent SBM on each level beforehand is a fast way to start with good initialization and allows us to state on the independence of the model at the same time as we just need to compare the sum of the ICL_{Ind} and $\text{ICL}_{\text{MLVSBM}}(\widehat{Q_I}, \widehat{Q_O})$.

Package All the codes are available as an `R` package at https://chabert-liddell.github.io/MLVSBM/. It features the simulation and inference of multilevel networks with symmetric and/or asymmetric adjacency matrices, model and independence selection. It also handles missing at random data (Rubin, 1976) on the adjacency matrices of one or both levels and link prediction.

2.4. Illustration on simulated data

In this section, we study the performances of the inference procedure for the MLVSBM including the ability to recover blocks, the selection of the numbers of blocks and the independence detection.

Remark. *In order to evaluate the ability to recover blocks, we resort to the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) which is a comparison index between two clusterings with a correction for chance. This index is close to* 0 *when the two clusterings are independent and is* 1 *when the clusterings are identical (up to label switching).*

Remark. *In our results, we focus on the ability to recover blocks rather than on the quality of the model parameter estimates since it is the hardest task. Indeed, once the blocks are recovered (ARI=1), the estimation of the model parameters boils down to the computation of the proportions of observed links between blocks which is a consistent estimator in a Bernoulli i.i.d. model.*

2.4.1. Experimental design

In what follows, we set $Q_O = Q_I = 3$. The networks are of sizes: $n_O = 60$ and $n_I = 180$.

Let d be a density parameter: the lower d, the sparser the network and the harder the inference. ϵ (≥ 1) is a parameter tuning the strength of the communities; when ϵ is high, the communities are easily separable. In the simulation study, we focus on the three following standard topologies.

- *Assortative communities.* The probability of connection within communities is higher than the probability of connection between communities: $\alpha^I = d * \begin{bmatrix} \epsilon & 1 & 1 \\ 1 & \epsilon & 1 \\ 1 & 1 & \epsilon \end{bmatrix}$.

- *Disassortative communities.* The probability of connection within communities is lower than the probability of connection between communities: $\alpha^I = d * \begin{bmatrix} 1 & \epsilon & \epsilon \\ \epsilon & 1 & \epsilon \\ \epsilon & \epsilon & 1 \end{bmatrix}$.
- *Core-periphery.* A core block is highly connected to the whole network while the probability of connection in the periphery is low: $\alpha^I = d * \begin{bmatrix} \epsilon & \epsilon & 1 \\ \epsilon & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.

We fix the topology of the inter-organizational level X^O to be an assortative communities with $d = 0.1$, $\epsilon = 5$ and of communities of equal size on average. We expect this topology to be easy to infer and to obtain a perfect recovery of the clustering with high probability.

For the inter-individual level, d is set to 0.01, 0.05 or 0.1 while ϵ ranges from 1 to 10 by stepize of 0.5. $\epsilon = 1$ corresponds to an Erdős-Rényi graph and the communities should be indistinguishable.
The affiliation matrix A is generated from a power-law distribution in order to get different sizes of organizations. Other distributions were tried but the results (not reported here) show that their impact on the inference is weak.

Finally, δ is a parameter for the strength of the dependence between levels, ranging from 0 to 1. More precisely, we set:

$$\gamma = \begin{bmatrix} \delta & \frac{1}{2}(1-\delta) & \frac{1}{2}(1-\delta) \\ \frac{1}{2}(1-\delta) & \delta & \frac{1}{2}(1-\delta) \\ \frac{1}{2}(1-\delta) & \frac{1}{2}(1-\delta) & \delta \end{bmatrix}$$

where γ has been defined in Equation (2.4). $\delta = 1/Q_I$ corresponds to the case of independence between levels. The further δ is from $1/Q_I$, the stronger the dependence between levels. $\delta = 1$ implies a deterministic link between the clustering of the two levels, ie. the block of an individual is fully determined by the block of his/her organization. With this experimental design we aim to exhibit how the inference is improved by applying the MLVSBM rather than the SBM when the two levels are intertwined.

2.4.2. Simulation results

During the inference procedure, the number of blocks is unknown for both levels. We run the model selection for $\widehat{Q_I} \in \{1, \ldots, 10\}$ and $\widehat{Q_O} \in \{1, \ldots, 10\}$.

First, we fix $\delta = 0.8$ and make ϵ vary. Each situation is simulated 50 times. We test the ability of our model to recover the true clustering of Z^I from (X^I, X^O). We compare our performances to the ones obtained by applying a standard (unilevel) SBM on X^I. Because (Q_I, Q_O) are assumed to be unknown, two types of error may occur: one for not selecting the right Q_I and one for assigning nodes to the wrong blocks. The results are displayed in Figure 2.4.

In Figure 2.4 A, we plot – for 3 values of density d and the 3 topologies (assortative, core-periphery and disassortative) – the ARI when using MLVSBM (plain line) and SBM (dashed line) as ε varies. We observe that, for any topology, the MLVSBM starts to recover perfectly the clustering for a lower value of ϵ than the SBM because in the MLVSBM, the inter-individual level benefits from the information held in the inter-organizational level through the dependence of their blocks. The difficulty of the inference increases as ϵ decreases: as can be seen in Figure 2.4 A, MLVSBM still performs well ($ARI > 0$) for small values of ϵ while the SBM is unable to recover the clustering.

In Figures 2.4 B and C, we plot the number of blocks chosen by the MLVSBM (B) and the SBM (C) for 3 values of density (rows) and 3 topologies (columns) (the true value being $Q_I = 3$). We observe that using the MLVSBM allows to recover more precisely Q_I than using the SBM. $\widehat{Q_I}$ varies from 1, when no structure is detected to 3 which is the true number of blocks. The procedure never selects more blocks than expected, which is coherent with prior knowledge that the ICL for the SBM tends to select models of smaller size (Hayashi et al., 2016; Brault, 2014).

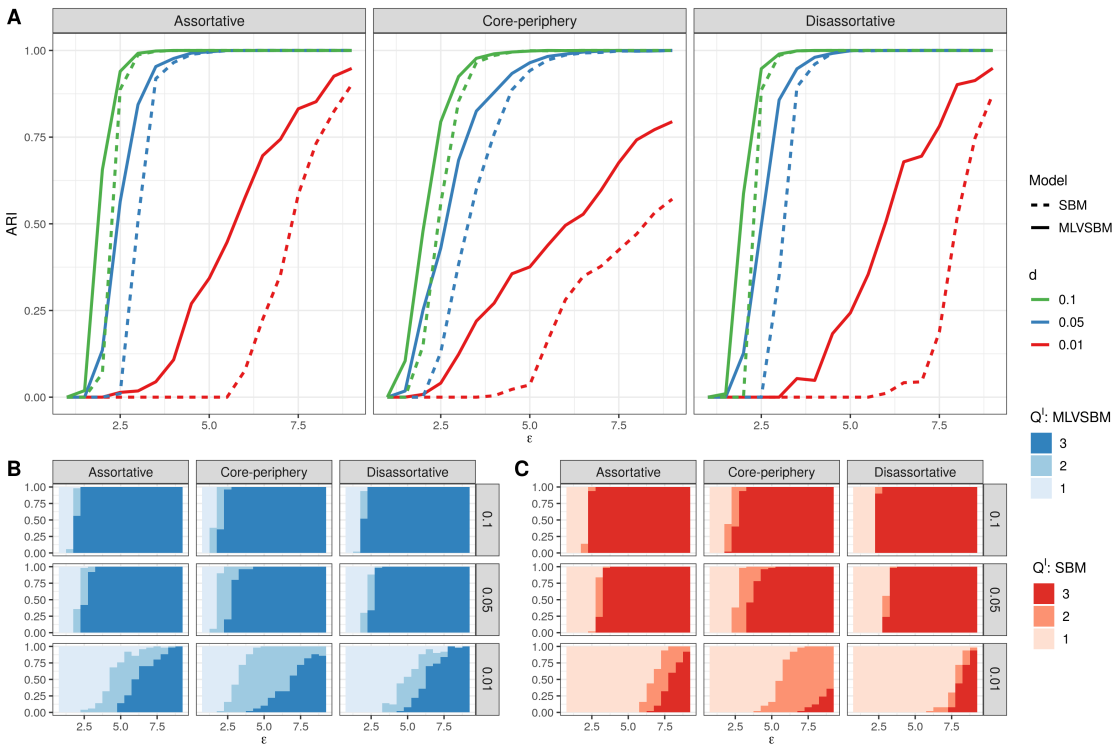


Figure 2.4 – Clustering and model selection for 3 different topologies on the inter-individual level, varying ϵ and density d. Each situation is simulated 50 times. **A**: ARI for the inter-individual level, comparing the model used for inference. **B**: Stacked frequency barplot of the selected number of blocks for the inter-individual level in the MLVSBM (in blue). **C**: Stacked frequency barplot of the selected number of blocks for the inter-individual level chosen in the SBM (in red).

On the three topologies with $\epsilon = 3$, depending on the density d, MLVSBM and

SBM supply Z^I either a perfect recovery of the clustering or a random clustering or something in between. In order to understand better this phenomenon, we fix ϵ to 3 and make δ – which quantifies the dependency between the two levels – vary. The results are reported in Figure 2.5 for 50 simulations of each situation.

When $\delta = 1/3$ (yellow vertical line in Figure 2.5 A), the two levels are independent and the results in terms of clustering are the same for the MLVSBM and the SBM on X^I (see ARI in Figure 2.5 A). As soon as δ departs from this value, the MLVSBM is able to recover some of the structure of the inter-individual level thanks to the inter-organizational level and this ability is observed even for very low density when δ gets closer to 1 (see Figure 2.5 A and B).

Figure 2.5.C depicts the performances of the ICL criterion to state on the independence between the two levels. For $d = 0.01$, X^I is very sparse, $\widehat{Q_I} = 1$ (no structure is detected on the inter-individual level) leading to $\text{ICL}_{ind} = \text{ICL}_{\text{MLVSBM}}$ and preventing us from detecting any dependency. For higher densities, we see as expected, that if $\delta \approx 1/3$, the independent SBM will be preferred. On the contrary the further δ departs from 1/3 the more the MLVSBM will be selected, even-though the MLVSBM and the independent SBM may provide the same clusterings. This phenomenon occurs faster for higher density d. In our simulation, the MLVSBM is never selected when $\delta = 1/3$. This is a consequence of the conservative nature of ICL, requiring strong evidence from the likelihood to select a more complex model.

We chose not to present results concerning the inter-organizational level since its structure was selected to be “easy-to-infer”. Hence, the SBM and the MLVSBM perform well for selecting the true number of blocks Q_O and recovering the block structure. Simulations gave similar results (not reported here) when we inverse the topologies on X^I and X^O, showing that information on structure transits in both ways. Moreover, when the number of nodes of the "easy-to-infer" level increases, it facilitates the recovering of the clustering on the "hard-to-infer" level. When both levels are "hard-to-infer", the inference of each level benefits from one another if the dependence between the two levels is strong enough. One can exhibit cases where the unilevel SBM is unable to recover the clustering of any of the two levels but where the MLVSBM succeeds in recovering the true blocks for both. Detailed results for such a simulation study are available on the MLVSBM `R` package website https://chabert-liddell.github.io/MLVSBM/articles/hard_to_infer.html.

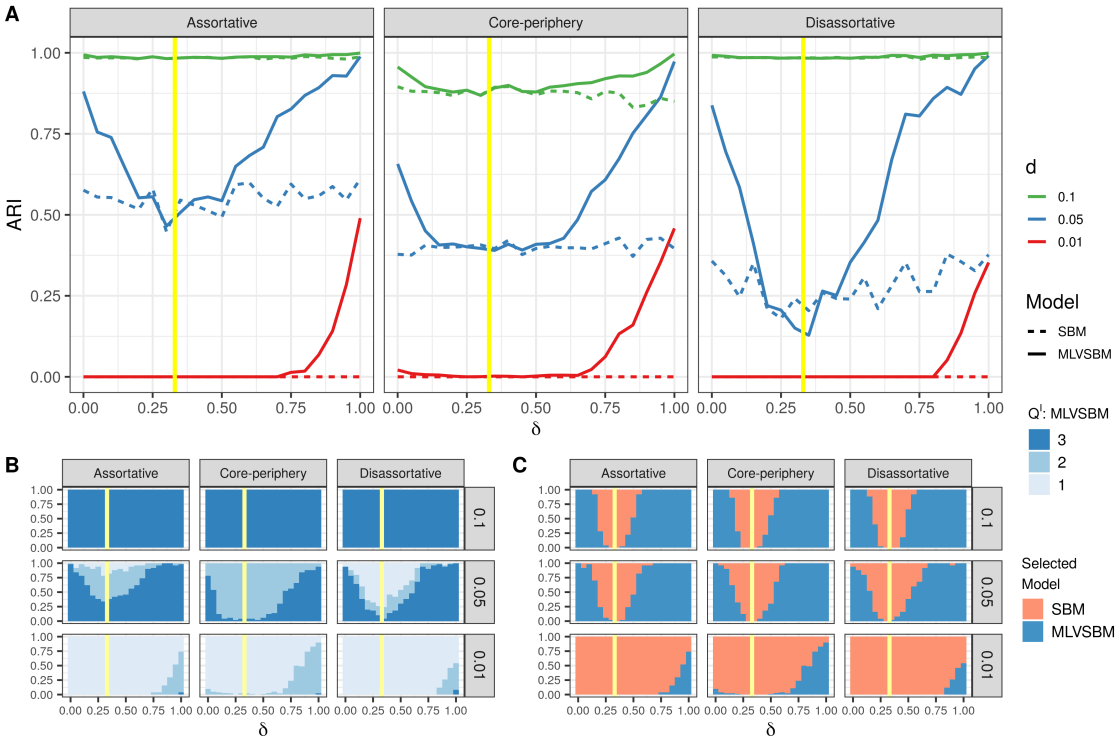


Figure 2.5 – Clustering and model selection for 3 different topologies on the inter-individual level, as function of δ and density d. Each situation is simulated 50 times. The yellow vertical lines represent a $\delta = 1/3$ (i.e. a γ with uniform coefficients, resulting into independence between the two levels). **A**: ARI for the inter-individual level, comparing the model used for inference. **B**: Stacked frequency barplot of the selected number of blocks for the inter-individual level in the MLVSBM. **C**: Stacked frequency barplot of the selected model with respect to inter-level dependence.

2.4.3. Computational costs

Inferring the blocks and the parameters of a multilevel network is a challenging task which can be time consuming. As a guideline for readers, we present in Table 2.1 the average computation time using the `R` package MLVSBM on two cores of a desktop computer with 32GB of RAM and a Intel® Xeon(R) CPU E5-1650 v4 @ 3.60GHz × 12 processor running on Ubuntu 18.04.5 LTS for the inference of simulated networks including model selection for different network sizes and different numbers of blocks.

Network Size		Running time (mean ± sd) in seconds		
n_I	n_O	$Q_I = Q_O = 2$	$Q_I = Q_O = 4$	$Q_I = Q_O = 8$
150	50	9.87 ± 4.13	-	-
600	200	443 ± 205	1794 ± 1287	-
1500	500	1093 ± 900	2583 ± 1226	7050 ± 2670

Table 2.1 – Average running time for the inference of the MLVSBM for different network sizes and different numbers of blocks.

2.5. Application to the multilevel network issued from a television programs trade fair

We apply our model to the data set (Brailly et al., 2016) described below.

2.5.1. Context and Description of the data set

Promoshow East is a television programs trade fair for Eastern Europe. Sellers from Western Europe and the USA come to sell audiovisual products to regional and local buyers such as broadcasting companies. The data gather observations on one particular audiovisual product, namely animation and cartoons. From a sociological perspective, reconstituting and analyzing multilevel (inter-individual and inter-organizational) networks in this industry is important. In economic sociology, it helps redefine the nature of markets (Brailly et al., 2016, 2017; Lazega and Mounier, 2002). In the sociology of culture, it helps understand, from a structural perspective, the mechanisms underlying contemporary globalization and standardization of culture (Brailly et al., 2016; Favre et al., 2016). In the sociology of organizations and collective action, it helps understand the importance of multilevel relational infrastructures for the management of tense competition and cooperation dilemmas by various categories of actors (Lazega, 2020), in this case the (sophisticated) sales representatives of cultural industries.

The data were collected by face-to-face interviews. At the individual level, people were asked to select from a list the individuals from which they obtain advice or information during or before the trade fair. The level consists of 128 individuals and 710 directed interactions (density = 0.044). The individuals were affiliated to 109 organizations, each one containing from one to six individuals. At the inter-organizational level, two kinds of interactions were collected: a deal network (deals signed since the last trade fair) and a meeting network (derived from the aggregation at the inter-organizational level of the meetings planned by individuals on the trade fair's website). Both networks are symmetric with respective densities 0.067 and 0.059.

2.5.2. Statistical analysis

The MLVSBM is inferred on the two datasets (one dataset corresponding to the deal network at the inter-organizational level, the other dataset to the meeting network at the inter-organizational level). In both cases the ICL criterion favors dependence between the two levels and chooses $\widehat{Q_I} = 4$ blocks of individuals. $\widehat{Q_O}$ is equal to 3 for the deal network and 4 for the meeting network.

In order to determine which is the most relevant inter-organizational network, we test the ability of the MLVSBM to predict dyads or links in the inter-individual network when the deal or the meeting networks are considered. To do so, we choose uniformly dyads and links to remove and try to predict them. More precisely, we set $X^I_{ii'} = \texttt{NA}$ for a certain percentage of (i, i') (this percentage ranging from 5% to 40% by step-size of 5%). We also propose to remove existing links (ie. forcing $X^I_{ii'} = 0$ when $X^I_{ii'} = 1$ was observed, for some randomly chosen (i, i')). The percentage of removed existing links varies from 5% to 95% (with step-size of 5%). We repeat the following procedure 100 times:

1. Remove dyads or links uniformly at random
2. Infer the newly obtained network from scratch in order to obtain the probability of a link $\mathbb{P}(X^I_{ii'} = 1; \widehat{\theta})$ for each missing dyad or for each dyad such that $X^I_{ii'} = 0$
3. Predict link among all missing dyads or among all dyads such that $X^I_{ii'} = 0$.

Missing data are handled as Missing At Random (Tabouy et al., 2019) and the probability of existence of an edge is given by: $\mathbb{P}(X^I_{ii'} = 1; \widehat{\theta}) = \sum_{k,k'} \widehat{\tau^I_{ik}} \widehat{\alpha^I_{kl}} \widehat{\tau^I_{i'k'}}$. Since the result of our procedure is equivalent to a binary classification problem, we assess the performance through the area under the ROC curve (AUC) (a random classification corresponding to AUC = 0.5).

Figure 2.6 shows that using the MLVSBM compared to a single level SBM improves a lot the recovery of the inter-individual level for this dataset. This confirms the dependence between levels detected by the ICL. Moreover, using the deal network gives better predictions for both missing dyads and missing links than the meeting network. We also considered a merged network at the inter-organizational level by making the union of links of the deal and the meeting network, i.e. for all $j, j' \in n_O$, $X^{O,\text{merged}}_{jj'} = \max\{X^{O,\text{deal}}_{jj'}, X^{O,\text{meeting}}_{jj'}\}$. The improvement in terms of prediction over the deal network is not very significant and this composite network is much harder to analyze sociologically.

Remark. *Another way to simulate missing data is to consider actor non-response like in (Žnidaršič et al., 2012). In our case, it corresponds to selecting a portion of the individuals at random and putting all their out-going dyads to* `NA` *(i.e.* $X^I_{ii'} = \texttt{NA}$ *for all* i' *if individual* i *did not respond). Then we look at the stability of the clustering as in Žnidaršič et al. (2012, 2019) (the ARI between the clustering of the individuals with the full data and the one with the missing data). By doing so, we notice in simulations (not reported here) that the clustering of the individuals is more stable when considering an MLVSBM on* $(X^I, X^{O,deal})$ *than when considering a unilevel*

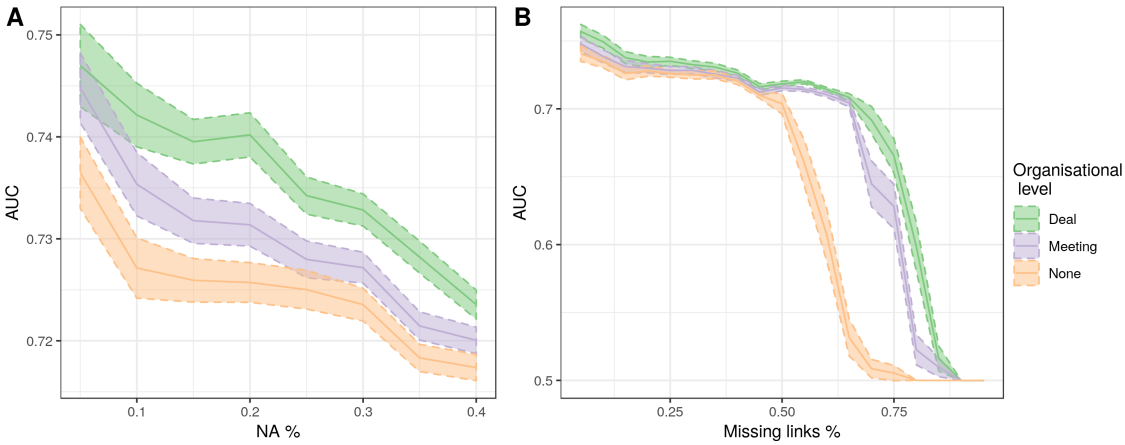


Figure 2.6 – AUC of the prediction for **A**: missing dyads, **B**: missing links, in function of the missing proportion for the inter-individual level. Colors represent different network at the inter-organizational level. None (beige) is equivalent to a single layer SBM on the individuals. The confidence interval is given by $mean \pm stderror$.

SBM on X^I. This is one more clue in favor of the dependence between the two levels.

Remark. *Žiberna (2019) and Žiberna (2020) also deals with this dataset from Brailly et al. (2016). However, Žiberna (2019) uses the dataset collected in 2012 and Žiberna (2020) gathers the datasets collected in 2011 and 2012 while we only use the 2011 dataset. Moreover, different choices were made on the individuals and organizations to include or not. Thus, a direct comparison does not make sense. Applying Žiberna's method on the dataset we consider provides us with clusterings that somewhat agree on both levels (ARIs>0.6). We have checked that the difference derives from the fact that the two methods do not seek the same patterns.*

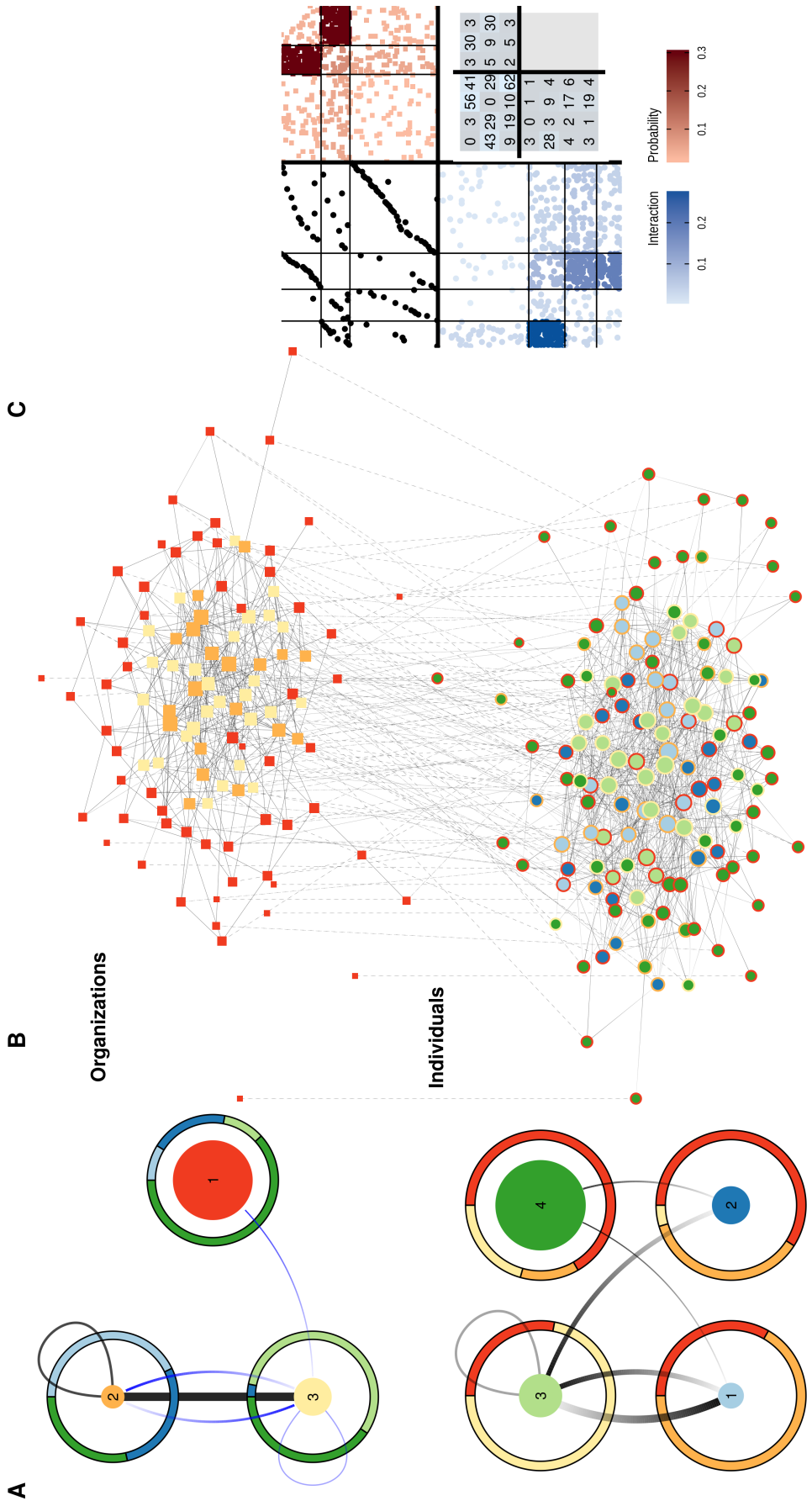


Figure 2.7 – Multilevel network of the Promoshow East trade fair 2011. Above: the deal network for the organizations and below: the advice network for the individuals. **A**: Mesoscopic view of the multilevel network. Nodes stand for the blocks, donut charts show the relation between Z^O and Z^I. Black edges are the probabilities of connection α^I and α^O, blue edges stand for $\mathbb{P}(X^I_{ii'} = 1 | Z^O_{A_i}, Z^O_{A_{i'}})$, i.e. the probability of interaction between organizations through their individuals. For sake of clarity only edges with probabilities above the density are shown. **B**: View of the network. The size of a node is proportional to its in-degree. Colors represent the clustering obtained with the MLVSBM. **C**: Matrix representation of the multilevel network. At the bottom-left, the adjacency matrix of the advice network between individuals, at the top-right, the deal network between organizations, at the top-left, the affiliation matrix of the individuals to the organizations. Entries are reordered by block from left to right and bottom to top. Blocks are separated by thin lines and levels by thick lines. The entries of the bottom-right matrix are the parameters α^I, γ and α^O multiplied by 100.

2.5.3. Analysis and comments

For the analysis, we use the MLVSBM inferred from the deal network. We select $\widehat{Q_O} = 3$ and $\widehat{Q_I} = 4$ blocks and the ICL is in favor of a dependence between the two levels. This network is plotted in Figure 2.7 B and we reordered the adjacency matrices of both levels by blocks in Figure 2.7 C. In Figure 2.7 A, we plot a synthetic view of the blocks of this multilevel network. The size of each node is proportional to the cardinal of each block. For the inter-organizational level, we link blocks of organizations by α^O (plain black edges) and by the probability of interactions of their individuals $\mathbb{P}(X^I_{ii'} = 1|Z^O_{A_i}, Z^O_{A_{i'}})$ (gradual blue edges). The donut charts around the nodes is the parameter γ. For the inter-individual level, blocks of individuals are linked by α^I and the donut chart for a given block is the apportionment of each block of organizations in the individuals' affiliation.

We can now interpret the block with respect to the actors' covariates shown in Table 2.2. At the inter-organizational level, block 1 (in red) is a residual group composed of 61 organizations that are weakly connected to the rest of the organizations. Block 2 (in orange) consists of customers: broadcasters that come to the trade fair to buy programs and independent buyers who buy programs, planning to sell them later to broadcasters. We observe a non-null intra-block connection, but deals are mainly done between organizations of the blocks 2 and 3 (block 3 in yellow), the latter mostly containing distributors.

At the inter-individual level, blocks 1 and 2 consist of buyers (exclusively for block 1). They differ in their affiliations, both are affiliated to the second block of organizations but a larger proportion of the individuals of block 2 are affiliated to the residual block of organizations. They also differ in the way they connect to blocks 3 and 4. Block 4 is a residual group consisting of roughly half of the individuals. It does not exhibit any particular pattern in its affiliations and is weakly connected, mainly inward connection from block 2. Block 3 consists of sellers giving advices to individuals of block 2 and has reciprocal relationship with individuals of block 1. They are mainly affiliated to producing and distributing companies of block 3 of organizations. It is also the block that has the strongest intra-block connections.

The blue edges in Figure 2.7 A show that the organizations of blocks 2 and 3 and their respective individuals follow the same pattern for their inter-block connections but differ in their intra-block connections. Individuals affiliated to organizations of block 3 have above average intra-block connections while few contracts are signed between their organizations (mainly distributors).

These results confirm neo-structural insights into the functioning of markets. Competition between producers/distributors is strong: they all need to find broadcasting companies and distributors on the buying side. However, most of them arrive to the trade fair without updated information about the products in which buyers are interested in that year, their available budgets for each category of product, their willingness to negotiate, etc. The value of multilevel network analysis that is used

Organizations		Covariates				
Block	Size	Producer	Distributor	Media group	Independent buyer	Broadcaster
1	61	14	16	9	14	8
2	20	1	0	2	7	10
3	28	3	19	5	1	0

Individuals		Covariates		Affiliation		
Block	Size	Buyer	Seller	1	2	3
1	18	18	0	6	12	0
2	22	16	6	13	8	1
3	25	2	23	7	0	18
4	63	15	48	42	8	13

Table 2.2 – Contingency table of covariates and clustering for the organizations (top) and the individuals (bottom)

here is to show that inter-individual personal relationships between individuals affiliated with competing organizations help manage the tensions between these directly competing organizations (Lazega et al., 2016; Lazega, 2009). This is where personal ties between individuals affiliated in these companies – especially among sellers and buyers, but also less visibly among sellers – are important: they help manage the strong tensions between companies by creating *coopetition*, i.e. cooperation among their competing firms. Here, social/advice ties between buyers (blocks 1 and 2 of individuals) affiliated to buying companies in block 2 of organizations (broadcasting companies and distributors) exchange advice from sellers of block 3 representing production and distribution companies: this is the normal, stabilized, overlapping, commercial ties between companies embedded in social ties between representatives.

As seen above, block 3 has strong intra-block connections which may signal discreet coordination efforts between sellers as shown by Brailly (2016); Brailly et al. (2016). When a seller has closed a deal with a buyer, he/she can advise and update another seller – i.e. a coopetitor in terms of affiliation to a competing company – about other products in which this buyer is interested, what budget is left in his/her pocket, i.e. precious information for the next sellers. This kind of personal service is expected to be reciprocated over the years; otherwise the relationship decays. This is the most unexpected phenomenon from an orthodox economic perspective and should lead to new perspectives in neo-structural economic sociology (Lazega and Mounier, 2002).

This cross-level interdependence between inter-organizational ties and inter-individual ties is strong enough for companies to be unable to lay off its sales representatives. Having long tried to replace costly trade fairs with online websites

and catalogues, companies realized that they still need the service that real persons and their personal relational capital provide in terms of multilevel management of coopetition (Lazega, 2020).

2.6. Discussion

In this paper, we propose an SBM for multilevel networks. We develop variational methods for the inference of the model and a criterion that allows us to choose the number of blocks and to state on the independence between the levels at the same time. There are clear advantages at considering a joint modeling of the two levels over an independent model for each level. Indeed, we show on some simulation studies that when we detect dependence between levels, it helps us to recover the block structure of a level with low signal thanks to the structure of the other level and also to improve the prediction of missing links or dyads. On the trade fair dataset, this joint modeling brought us a synthetic representation of the two networks unraveling their intertwined structure and provide new insights on the social organization.

In lieu of a Bernoulli distribution, the edge distribution of any level may be extended to a valued distribution and/or to include edge covariates in a similar way as for the SBM (Mariadassou et al., 2010). One way to account for the degree distribution would be to use nodes degrees as covariates, another would be to rewrite the edge distribution as the Degree Corrected SBM (Karrer and Newman, 2011). Our choice to model the interaction levels given the affiliations (A being fixed) is driven by the fact that, in a lot of applications, these affiliations are known and the object of the analysis is the interactions. We choose to consider a unique affiliation per individual since this was the case on the datasets available to us, but this approach could be extended to a less restricted number of affiliations (this model is implemented in our R package). We could even consider any hierarchical structure such as multi-scale networks to model the levels given the hierarchy or more generally multilayer networks by modeling the layers given the inter-layers.

Furthermore, our model is able to decide about the independence of the structure of connections of the two levels. This is done by a model selection criterion. It would be interesting to test (in a statistical meaning) this independence but we know that the variance of our estimators is underestimated because of the variational approach (see Blei et al. (2017) for a review). Besides, sociological studies stated that some individuals benefit more than others from their organization's interactions (Lazega and Snijders, 2015), which could lead us to consider more local independence between levels.

For multiplex networks, De Bacco et al. (2017) use dyad predictions as a way to define interdependence between layers while Stanley et al. (2016) make a clustering of layer by aggregating the most similar. Our work considers multilevel networks where each level has nodes of different natures and Figure 2.6 shows that the dependence between levels leads to a better recovery of missing information. This can be used to help data collection or to correct spurious information on existing data as suggested

in Clauset et al. (2008) or Guimerà and Sales-Pardo (2009). Indeed, one might imagine that the data of one level may be easier to collect or to verify than the other one (for instance because it is public, already exists or is cheaper to collect). Thus, we think that this approach could be used to leverage the interdependence in a multilevel network in order to compensate for some missing or spurious information on a given level which is known to be difficult to observe.

Acknowledgements

The authors would like to thank Julien Brailly for providing the dataset. This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. This work was partially supported by the grant ANR-18-CE02-0010-01 of the French National Research Agency ANR (project EcoNet). This project received financial support from INRAE and CIRAD as part of the SEARS project funded by the GloFoods metaprogram. This work was presented and discussed within the framework of working days organized by the MIRES group (with the financial support of INRAE) and the GDR RESODIV (with the financial support of CNRS).

2.A. Proof of Proposition 2.1

Proposition 2.1. *In the MLVSBM, the two following properties are equivalent:*

1. *Z^I is independent on Z^O,*
2. *$\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1, \dots, Q_O\}$,*

and imply that:

3. *X^I and X^O are independent.*

Proof. We first derive an expression for $\ell_\gamma(Z^I) = \ell_\gamma(Z^I|A)$:

$$\begin{aligned}
\ell_\gamma(Z^I|A) &= \int_{Z^O} \ell_\gamma(Z^I|A, Z^O) d\mathbb{P}(Z^O) \\
&= \sum_{l_1,\dots,l_{n_O}} \ell_\gamma(Z^I|A, Z_1^O = l_1, \dots, Z_{n_O}^O = l_{n_O}) \mathbb{P}(Z_1^O = l_1, \dots, Z_{n_O}^O = l_{n_O}) \\
&= \sum_{l_1,\dots,l_{n_O}} \prod_j \left(\prod_i \ell_\gamma(Z_i^I|A, Z_{A_i}^O = l_{A_i}) \right) \mathbb{P}(Z_j^O = l_j) \\
&= \sum_{l_1,\dots,l_{n_O}} \prod_j \left(\prod_{i,k} \gamma_{kl_j}^{\mathbb{1}_{Z_i^I = k} A_{ij}} \right) \pi_{l_j}^O = \prod_j \sum_l \prod_{i,k} \gamma_{kl}^{A_{ij} \mathbb{1}_{Z_i^I = k}} \pi_l^O
\end{aligned}$$

where $A_i = \{j : A_{ij} = 1\}$.

2. $\Rightarrow$ 1.: Assume that $\gamma_{kl} = \gamma_{kl'} \quad \forall l, l' \in \{1, \dots, Q_O\}$, then:

$$\begin{aligned}
\ell_\gamma(Z^I|Z^O, A) &= \prod_{k,l} \gamma_{kl}^{\sum_{i,j} A_{ij} \mathbb{1}_{Z_i^I = k} \mathbb{1}_{Z_j^O = l}} = \prod_k \gamma_{k1}^{\sum_{i,j} A_{ij} \mathbb{1}_{Z_i^I = k} \sum_l \mathbb{1}_{Z_j^O = l}} \\
&= \prod_{i,k} \gamma_{k1}^{\mathbb{1}_{Z_i^I = k}},
\end{aligned}$$

and

$$\begin{aligned}
\ell_\gamma(Z^I|A) &= \prod_j \sum_l \prod_{i,k} \gamma_{kl}^{A_{ij} \mathbb{1}_{Z_i^I = k}} \pi_l^O \\
&= \prod_j \prod_{i,k} \gamma_{k1}^{A_{ij} \mathbb{1}_{Z_i^I = k}} \sum_l \pi_l^O = \prod_{i,k} \gamma_{k1}^{\mathbb{1}_{Z_i^I = k}},
\end{aligned}$$

hence $\ell_\gamma(Z^I|Z^O, A) = \ell_\gamma(Z^I|A)$.

1. $\Rightarrow$ 2.: Assume that $\ell_\gamma(Z^I|Z^O, A) = \ell_\gamma(Z^I|A)$ for any values of Z^I, Z^O, then in particular $\ell_\gamma(Z_1^I|Z^O, A) = \ell_\gamma(Z_1^I|A)$. Assuming that individual 1 belongs to organization j, we can write, for any k:

$$\mathbb{P}(Z_1^I = k | Z_j^O, A_{ij} = 1) = \gamma_{kZ_j^O}.$$

However, this quantity does not depend on Z_j^O so $\gamma_{kZ_j^O} = \gamma_{k\cdot}$ for any value of k and Z_j^O. And so we have $\gamma_{kl} = \gamma_{kl'}$ for any (l, l').

1. $\Rightarrow$ 3.:

$$\begin{aligned}
\ell_{\alpha^I,\alpha^O}(X^I, X^O|A) &= \int_{z^I,z^O} \ell_{\alpha^I,\alpha^O}(X^I, X^O|A, Z^I = z^I, Z^O = z^O)\mathbb{P}(Z^I = z^I, Z^O = z^O)\mathrm{d}z^I\mathrm{d}z^O \\
&= \int_{z^I,z^O} \ell_{\alpha^I}(X^I|Z^I = z^I)\mathbb{P}(Z^I = z^I|A, Z^O = z^O)\ell_{\alpha^O}(X^O|Z^O = z^O)\mathbb{P}(Z^O = z^O)\mathrm{d}z^I\mathrm{d}z^O \\
&= \int_{z^I} \ell_{\alpha^I}(X^I|Z^I = z^I)\mathbb{P}(Z^I = z^I)\mathrm{d}z^I \int_{z^O} \ell_{\alpha^O}(X^O|Z^O = z^O)\mathbb{P}(Z^O = z^O)\mathrm{d}z^O \\
&= \ell_{\alpha^I}(X^I)\ell_{\alpha^O}(X^O)
\end{aligned}$$

which is the definition of the independence. □

2.B. Proof of Proposition 2.2

Proposition 2.2. *The stochastic block model for multilevel networks is identifiable up to label switching under the following assumptions:*

- $\mathcal{A}1$. *All coefficients of $\alpha^I \cdot \gamma \cdot \pi^O$ are distinct and all coefficients of $\alpha^O \cdot \pi^O$ are distinct.*
- $\mathcal{A}2$. *$n_I \geq 2Q_I$ and $n_O \geq \max(2Q_O, Q_O + Q_I - 1)$.*
- $\mathcal{A}3$. *At least $2Q_I$ organizations contain one individual or more.*

Proof. Let $\theta = \{\pi^O, \gamma, \alpha^I, \alpha^O\}$ be the set of parameters and $\mathbb{P}_X$ the distribution of the observed data. We will prove that there is a unique θ corresponding to $\mathbb{P}_X$. More precisely, in what follows, we will compute the probabilities of some particular events, from which we will derive a unique expression for the unknown parameters. The beginning of the proof –identifiability of π^O and α^O– is mimicking the one given in Celisse et al. (2012). The last steps of the proof are original work.

Notations. For the sake of simplicity, in what follows, we use the following shorten notation:

$$x_{i:k} := (x_i, \ldots, x_k), \quad X_{j,i:k} = (X_{ji}, \ldots, X_{jk}).$$

Moreover, $\{X_{j,i:k} = 1\}$ stands for $\{X_{ji} = 1, \ldots, X_{jk} = 1\}$.

Identifiability of π^O For any $l = 1, \ldots, Q_O$, let τ_l be the following probability:

$$\tau_l = \mathbb{P}(X_{ij}^O = 1|Z_i^O = l) = \sum_{l'} \alpha_{ll'}^O \pi_{l'}^O = (\alpha^O \cdot \pi^O)_l, \quad \forall (i, j). \tag{2.B.13}$$

Moreover, a quick computation proves that

$$\mathbb{P}(X_{i,j:(j+k)}^O = 1|Z_i^O = l) = \tau_l^{k+1} \tag{2.B.14}$$

According to Assumption $\mathcal{A}1$, the coordinates of vector $(\tau_1, \ldots, \tau_{Q_O})$ are all different. Hence, the Vandermonde matrix R^O of size $Q_O \times Q_O$ such that

$$R^O_{il} = (\tau_l)^{i-1}, \quad 1 \leq i \leq Q_O, \quad 1 \leq l \leq Q_O$$

is invertible. We define u^O_i as follows:

$$\begin{array}{rcl} u^O_i & = & \mathbb{P}_{\mathbf{X},\theta}(X^O_{1,2:(i+1)} = 1) \quad \text{for } 1 \leq i \leq 2Q_O - 1 \\ u^O_0 & = & 1. \end{array}$$

The existence of $(u^O_i)_{i=0,\ldots,2Q_O-1}$ comes from Assumption $\mathcal{A}2$ ($n_O \geq 2Q_O$). Moreover, the $(u^O_i)_{i=0,\ldots,2Q_O-1}$ are calculated from the marginal distribution $\mathbb{P}_X$. We will use these quantities to identify the parameters (π^O, α^O).

First we have, for $1 \leq i \leq 2Q_O - 1$:

$$u^O_i = \sum_{l=1}^{Q_O} \mathbb{P}(X^O_{1,2:(i+1)} = 1 | Z^O_1 = l)\mathbb{P}(Z^O_1 = l) = \sum_{l=1}^{Q_O} \tau^i_l \pi^O_l,$$

using Equation (2.B.14). Now, let us define M^O a $(Q_O + 1) \times Q_O$ matrix such that:

$$M^O_{ij} = u^O_{i+j-2} = \sum_{l=1}^{Q_O} \tau^{i-1}_l \pi^O_l \tau^{j-1}_l, \quad 1 \leq i \leq Q_O + 1, \quad 1 \leq j \leq Q_O. \tag{2.B.15}$$

For $k \in \{1, \ldots, Q_O + 1\}$, we define δ_k as $\delta_k = \mathrm{Det}(M^O_{-k})$ where M^O_{-k} is the square matrix corresponding to M^O without the k-th row. Let B^O be the polynomial function defined as:

$$B^O(x) = \sum_{k=0}^{Q_O} (-1)^{k+Q_O} \delta_{k+1} x^k. \tag{2.B.16}$$

- B^O is of degree Q_O. Indeed, $\delta_{Q_O+1} = \det(M^O_{-(Q_O+1)})$ and $M_{-(Q_O+1)} = R^O D_{\pi^O} R^{O'}$ where $D_{\pi^O} = \mathrm{diag}(\pi^O)$. As a consequence, $M^O_{-(Q_O+1)}$ is the product of invertible matrices then $\delta_{Q_O+1} \neq 0$ and we can conclude.
- Moreover, $\forall l = 1, \ldots, Q_O$, $B^O(\tau_l) = 0$. Indeed, $B^O(\tau_l) = \det(N^O_l)$ where N^O_l is the concatenated matrix $N^O_l = \left(M^O \,|\, V_l\right)$ with $V_l = [1, \tau_l, \ldots, \tau^{Q_O}_l]'$ (computation of the determinant development against the last column). However, from Equation (2.B.15), we have $M^O_{\bullet j} = \sum_l \tau^{j-1}_l \pi^O_l V_l$, i.e. each column vector of M^O is a linear combination of $V_1, \ldots, V_{Q_O}$. As a consequence, $\forall l = 1, \ldots, Q_O$, N^O_l is of rank $< Q_O + 1$, and so $B^O(\tau_l) = 0$.

The $(\tau_l)_{l=1,\ldots,Q_O}$ being the roots of B, they can be expressed in a unique way (up to label switching) as functions of $(\delta_k)_{k=0,\ldots,Q_O}$, which themselves are derived from $\mathbb{P}_{\mathbf{X},\theta}$. As a consequence, the identifiability of R^O is derived from the identifiability of $(\tau_l)_{l=1,\ldots,Q_O}$. Using the fact that $D_{\pi^O} = R^{O^{-1}} M^O_{-Q_O} R^{O'^{-1}}$, we can identify π^O in a unique way.

Identifiability of α^O For $1 \leq i, j \leq Q_O$, we define U_{ij} as follows:

$$U_{ij}^O = \mathbb{P}(X_{1,2:(i+1)}^O = 1, X_{2,(n_O-j+2):n_O}^O = 1)$$

with $U_{i1}^O = \mathbb{P}(X_{1,2:(i+1)}^O = 1)$.

$$U_{i,j}^O = \sum_{l_1,l_2} \tau_{l_1}^{i-1} \pi_{l_1}^O \alpha_{l_1 l_2}^O \pi_{l_2}^O (\tau_{l_2})^{j-1}, \quad \forall 1 \leq i, j \leq Q_O,$$

and as consequence $U^O = R^O D_{\pi^O} \alpha^O D_{\pi^O} R^{O'}$. D_{π^O} and R^O being invertible, we get: $\alpha^O = D_{\pi^O}^{-1} R^{O^{-1}} U^O R^{O'^{-1}} D_{\pi^O}^{-1}$. And so U_O is uniquely derived from $\mathbb{P}_X$, so α^O is identified.

Identifiability of α^I To identify α^I, we have to take into account the affiliation matrix A. Without loss of generality, we reorder the entries of both levels such that the affiliation matrix A has its $2Q_I \times 2Q_I$ top left block being an identity matrix (Assumption $\mathcal{A}3$).

- For any $k = 1 \ldots, Q_I$ and for $i = 2, \ldots, 2Q_I$, let σ_k be the probability $\mathbb{P}(X_{1i}^I = 1 | Z_1^I = k, A)$, A being such that $A_{jj} = 1, \forall j = 1, \ldots, 2Q_I$.

$$\begin{aligned}\sigma_k &= \mathbb{P}(X_{1i}^I = 1 | Z_1^I = k, A) \\ &= \sum_{k'} \mathbb{P}(X_{1i}^I = 1 | Z_1^I = k, Z_i^I = k') \mathbb{P}(Z_i^I = k' | Z_1^I = k, A)\,.\end{aligned}$$

  Moreover,

$$\begin{aligned}\mathbb{P}(Z_i^I = k' | Z_1^I = k, A) &= \sum_l \mathbb{P}(Z_i^I = k' | Z_i^O = l, Z_1^I = k, A) \mathbb{P}(Z_i^O = l | Z_1^I = k, A) \\ &= \sum_l \gamma_{kl} \mathbb{P}(Z_i^O = l | Z_1^I = k, A)\,. \qquad (2.B.17)\end{aligned}$$

  However, by Bayes' formula

$$\mathbb{P}(Z_i^O = l | Z_1^I = k, A) = \frac{\mathbb{P}(Z_1^I = k | Z_i^O = l, A) \mathbb{P}(Z_i^O = l)}{\mathbb{P}(Z_1^I = k, A)}\,.$$

  Taking into the fact that $i \neq 1$ and A is such that 1 belongs to organization 1 and i to organization i, we have: $\mathbb{P}(Z_1^I = k | Z_i^O = l, A) = \mathbb{P}(Z_1^I = k | A)$. And so

$$\mathbb{P}(Z_i^O = l | Z_1^I = k, A) = \mathbb{P}(Z_i^O = l | A) = \pi_l^O.$$

  Consequently, from Equation (2.B.17), we have:

$$\mathbb{P}(Z_i^I = k' | Z_1^I = k, A) = \sum_l \gamma_{k'l} \pi_k^O$$

and so:

$$\begin{aligned}\sigma_k &= \sum_{k'} \mathbb{P}(X^I_{1i} = 1 | Z^I_1 = k, Z^I_i = k') \sum_l \gamma_{k'l} \pi^O_k \\ &= \sum_{k'l} \alpha^I_{kk'} \gamma_{k'l} \pi^O_l = (\alpha^I \cdot \gamma \cdot \pi^O)_k \\ &= (\alpha^I \cdot \pi^I)_k, \qquad \text{where } \pi^I = \gamma \cdot \pi^O.\end{aligned}$$

- Now, we prove that $\forall i = 1, \ldots, 2Q_I - 1$,

$$\mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_1 = k, A) = \sigma^i_k. \tag{2.B.18}$$

Indeed,

$$\begin{aligned}&\mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_1 = k, A) \\ &= \sum_{k_{2:(i+1)}} \mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_{1:(i+1)} = (k, k_{2:(i+1)}), Z^I_1 = k) \mathbb{P}(Z^I_{2:(i+1)} = k_{2:i+1} | Z^I_1 = k, A) \\ &= \sum_{k_{2:(i+1)}} \mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_{1:(i+1)} = (k, k_{2:(i+1)})) \mathbb{P}(Z^I_{2:(i+1)} = k_{2:i+1} | A) \\ &= \sum_{k_{2:(i+1)}} \mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_{1:(i+1)} = (k, k_{2:(i+1)})) \\ &\qquad \sum_{l_{2:(i+1)}} \mathbb{P}(Z^I_{2:(i+1)} = k_{2:(i+1)}, Z^O_{2:(i+1)} = l_{2:(i+1)}, A).\end{aligned}$$

Note that, to go from line 2 to line 3, we used the fact that $\mathbb{P}(Z^I_{2:(i+1)} = k_{2:i+1} | Z^I_1 = k, A) = \mathbb{P}(Z^I_{2:(i+1)} = k_{2:i+1} | A)$, which is due the the particular structure of A (left diagonal block of size at least $2Q_I$, i.e. for any $i' = 1, \ldots, 2Q_I$, individual i' belongs to organization i'). Moreover, we can write:

$$\begin{aligned}&\mathbb{P}(Z^I_{2:(i+1)} = k_{2:(i+1)}, Z^O_{2:(i+1)} = l_{2:i+1} | A) \\ &= \left[\prod_{\lambda=2,\ldots i+1} \mathbb{P}(Z^I_\lambda = k_\lambda | Z^O_\lambda = l_\lambda) \mathbb{P}(Z^O_\lambda = l_\lambda) \right] \\ &= \left[\prod_{\lambda=2,\ldots i+1} \gamma_{k_\lambda l_\lambda} \pi^O_\lambda \right].\end{aligned}$$

Moreover, by conditional independence of the entries of the matrix X^I given the clustering we have:

$$\mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_1 = k, Z^I_{2:(i+1)} = k_{2:(i+1)}) = \prod_{\lambda=2,\ldots i+1} \alpha^I_{kk_\lambda}.$$

As a consequence,

$$\begin{aligned}\mathbb{P}(X^I_{1,2:(i+1)} = 1 | Z^I_1 = k, A) &= \sum_{k_{2:(i+1)}, l_{2:(i+1)}} \prod_{\lambda=2,\ldots i+1} \alpha^I_{kk_\lambda} \gamma_{k_\lambda l_\lambda} \pi^O_\lambda \\ &= \prod_{\lambda=2,\ldots i+1} \sum_{k_\lambda, l_\lambda} \alpha^I_{kk_\lambda} \gamma_{k_\lambda l_\lambda} \pi^O_\lambda = \sigma^i_k\end{aligned}$$

- Then we define $(u_i^I)_{i=0,\dots,2Q_I-1}$, such that $u_0^I = 1$ and $\forall 1 \leq i \leq 2Q_I - 1$:

$$\begin{aligned} u_i^I &= \mathbb{P}(X^I_{1,2:(i+1)} = 1|A) \\ &= \sum_{k,l} \mathbb{P}(X^I_{1,2:(i+1)} = 1|Z_1^I = k)\mathbb{P}(Z_1^I = k|Z_1^O = l, A)\mathbb{P}(Z_1^O = l) \\ &= \sum_k \sigma_k^i \underbrace{\sum_l \gamma_{kl}\pi_l^O}_{=\pi_k^I} \\ &= \sum_k \sigma_k^i \pi_k^I. \end{aligned}$$

  Note that the (u^I)'s can be defined because $n_I \geq 2Q_I$ (assumption $\mathcal{A}2$).

- To conclude we use the same arguments as the ones used for the identifiability of α^O, i.e. we define M^I a $(Q_I+1) \times Q_I$ matrix such that $M^I_{ij} = u^I_{i+j-2}$ together with the matrices M^I_{-k} and the polynomial function B^I (see Equation (2.B.16)). Let R^I be a $Q_I \times Q_I$ matrix such that $R^I_{ik} = \sigma_k^{i-1}$. R^I is an invertible Vandermonde matrix because of assumption $\mathcal{A}1$ on $\alpha^I \cdot \gamma \cdot \pi^O$. As before, R^I can be identified in unique way from B^I. Then, noting that $M^I_{-(Q_I+1)} = R^I D_{\pi^I} R^{I'}$ where $D_{\pi^I} = \mathrm{diag}(\pi^I) = \mathrm{diag}(\gamma \cdot \pi^O)$, we obtain: $D_{\pi^I} = (R^I)^{-1} M^I_{-Q_I} (R^{I'-1})$, which is uniquely defined by $\mathbb{P}_X$. Now, let us introduce

$$U^I_{ij} = \mathbb{P}(X^I_{1,2:(i+1)} = 1, X^I_{2,(n_I-j+2):n_I} = 1)$$

  with $U^I_{i1} = \mathbb{P}(X^I_{1,2:(i+1)} = 1)$. Then we have $U^I = R^I D_{\pi^I} \alpha^I D_{\pi^I} R^{I'}$ and so $\alpha^I = D^{-1}_{\pi^I}(R^I)^{-1} U^I (R^I)'^{-1} D^{-1}_{\pi^I}$. As a consequence, α^I is uniquely identified from $\mathbb{P}_X$.

Identifiability of γ For any $2 \leq i \leq Q_I$ and $2 \leq j \leq Q_O$, let $U^{IO}_{i,j}$ be the probability that $X^I_{1,2:i} = 1$ and $X^O_{1,(i+1):(i+j-1)} = 1$. Note that the $U^{IO}_{i,j}$ can be defined because $n_O \geq Q_I + Q_O - 1$ and $n_I \geq Q_I$ (assumption $\mathcal{A}2$).

- Then, for all $2 \leq i \leq Q_I$ and $2 \leq j \leq Q_O$,

$$\begin{aligned} U^{IO}_{ij} &= \mathbb{P}(X^I_{1,2:i} = 1, X^O_{1,(i+1):(i+j-1)} = 1|A) \\ &= \sum_{k,l} \mathbb{P}(X^I_{1,2:i} = 1, X^O_{1,(i+1):(i+j-1)} = 1|A, Z_1^I = k, Z_1^O = l) \\ &\quad \times \mathbb{P}(Z_1^I = k, Z_1^O = l, A)\,. \end{aligned} \tag{2.B.19}$$

- We first prove that :

$$\mathbb{P}(X^I_{1,2:i} = 1, X^O_{1,i+1:i+j-1} = 1|A, Z_1^I = k, Z_1^O = l) = \sigma_k^{i-1}\tau_l^{j-1}\,. \tag{2.B.20}$$

Indeed,

$$
\begin{aligned}
&\mathbb{P}(X^I_{1,2:i} = 1, X^O_{1,(i+1):(i+j-1)} = 1 | A, Z^I_1 = k, Z^O_1 = l) = \\
&= \sum_{k_{2:i}, l_{2:n_O}} \mathbb{P}(X^I_{1,2:i} = 1, X^O_{1,(i+1):(i+j-1)} = 1 | Z^I_{1:i} = (k, k_{2:i}), Z^O = (l, l_{2:n_O}), A) \\
&\quad \times \mathbb{P}(Z^I_{2:i} = k_{2:i}, Z^O_{2:n_O} = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A) \\
&= \sum_{k_{2:i}, l_{2:n_O}} \mathbb{P}(X^I_{1,2:i} = 1 | Z^I_{1:i} = (k, k_{2:i})) \\
&\quad \times \mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l, Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)}) \\
&\quad \times \mathbb{P}(Z^I_{2:i} = k_{2:i}, Z^O_{2:n_O} = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A) \,. \qquad (2.B.21)
\end{aligned}
$$

Moreover, let us have a look at $\mathbb{P}(Z^I_{2:i} = k_{2:i}, Z^O = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A)$:

$$
\begin{aligned}
&\mathbb{P}(Z^I_{2:i} = k_{2:i}, Z^O_{2:n_O} = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A) \\
= \; &\mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^O_{2:n_O} = l_{2:n_O}, Z^I_1 = k, Z^O_1 = l, A) \times \mathbb{P}(Z^O_{2:n_O} = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A) \,.
\end{aligned}
$$

Because A has a diagonal block of size $\geq Q_I$, we have, for any $i = 1, \ldots, Q_I$, $A_{ij} = 1$ if $j = i$, 0 otherwise, we have

- $\mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^O_{2:n_O} = l_{2:n_O}, Z^I_1 = k, Z^O_1 = l, A) = \mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^O_{2:i} = l_{2:i})$,

- $\mathbb{P}(Z^O_{2:n_O} = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A) = \mathbb{P}(Z^O_{2:n_O} = l_{2:n_O}) \,.$

As a consequence,

$$
\begin{aligned}
&\mathbb{P}(Z^I_{2:i} = k_{2:i}, Z^O_{2:n_O} = l_{2:n_O} | Z^I_1 = k, Z^O_1 = l, A) = \\
&\quad \mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^O_{2:i} = l_{2:i}) \mathbb{P}(Z^O_{2:i} = l_{2:i}) \mathbb{P}(Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)}) \\
&\times \mathbb{P}(Z^O_{(i+j):n_O} = l_{(i+j):n_O}) \,.
\end{aligned}
$$

Going back to Equation (2.B.21) and decomposing the summation we obtain:

$$
\begin{aligned}
& \mathbb{P}(X^I_{1,2:i} = X^O_{1,(i+1):(i+j-1)} = 1 | A, Z^I_1 = k, Z^O_1 = l) \\
= & \sum_{k_{2:i}, l_{2:n_O}} \mathbb{P}(X^I_{1,2:i} = 1 | Z^I_{1:i} = (k, k_{2:i})) \\
& \quad \times \mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l, Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)}) \\
& \quad \times \mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^O_{2:i} = l_{2:i}) \mathbb{P}(Z^O_{2:i} = l_{2:i}) \mathbb{P}(Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)}) \\
& \quad \times \mathbb{P}(Z^O_{(i+j):n_O} = l_{(i+j):n_O}) \\
= & \sum_{k_{2:i}} \mathbb{P}(X^I_{1,2:i} = 1 | Z^I_{1:i} = (k, k_{2:i})) \sum_{l_{2:i}} \mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^O_{2:i} = l_{2:i}) \mathbb{P}(Z^O_{2:i} = l_{2:i}) \\
& \sum_{l_{(i+1):(i+j-1)}} \mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l, Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)})) \\
& \times \underbrace{\mathbb{P}(Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)})}_{=\mathbb{P}(Z^O_{(i+1):(i+j-1)} = l_{(i+1):(i+j-1)} | Z^O_1 = l)} \underbrace{\sum_{l_{(i+j):n_O}} \mathbb{P}(Z^O_{(i+j):n_O} = l_{(i+j):n_O})}_{=1} \\
= & \sum_{k_{2:i}} \mathbb{P}(X^I_{1,2:i} = 1 | Z^I_1 = k, Z^I_{2:i} = k_{2:i}) \mathbb{P}(Z^I_{2:i} = k_{2:i} | A) \times \mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l) \\
= & \sum_{k_{2:i}} \mathbb{P}(X^I_{1,2:i} = 1 | Z^I_1 = k, Z^I_{2:i} = k_{2:i}) \mathbb{P}(Z^I_{2:i} = k_{2:i} | Z^I_1 = k, A) \\
& \times \mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l) \\
= & \mathbb{P}(X^I_{1,2:i} = 1 | Z^I_1 = k, A) \mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l) \,.
\end{aligned}
$$

Finally, we have :

$$
\begin{aligned}
\mathbb{P}(X^I_{1,2:i} = 1 | Z^I_1 = k, A) &= \sigma_k^{i-1}, \quad \text{from Equation (2.B.18)} \\
\mathbb{P}(X^O_{1,(i+1):(i+j-1)} = 1 | Z^O_1 = l) &= \tau_l^{j-1},
\end{aligned}
$$

and so, we have proved equality (2.B.20).

- Now, $A_{11} = 1$ implies $\mathbb{P}(Z^I_1 = k, Z^O_1 = l | A) = \gamma_{kl} \pi^O_l$ and combining this result with Equations (2.B.20) and (2.B.19) leads to: $U^{IO}_{ij} = \sum_{k,l} \sigma_k^{i-1} \gamma_{kl} \pi^O_l \tau_l^{j-1}$. Setting

$$
\begin{aligned}
U^{IO}_{1j} &= \mathbb{P}(X^O_{1,i+1} = 1, \ldots, X^O_{1,i+j-1} = 1 | A) = \sum_{k,l} \gamma_{kl} \pi^O_l \tau_l^{j-1}, \quad \text{for } j > 1 \\
U^{IO}_{i1} &= \mathbb{P}(X^I_{12} = \cdots = X^I_{1,i} = 1 | A) = \sum_{k,l} \gamma_{kl} \pi^O_l, \quad \text{for } i > 1 \\
U^{IO}_{11} &= 1
\end{aligned}
$$

we obtain the following matrix expression for U^{IO}: $U^{IO} = R^I \gamma D_{\pi^O} R^{O'}$ where U^{IO} is completely defined by $\mathbb{P}_{X,\theta}$ and the other terms have been identified before. Thus $\gamma = (R^I)^{-1} U^{IO} (R^{O'})^{-1} D^{-1}_{\pi^O}$ and γ is identified.

□

2.C. Details of the Variational EM

The variational bound for the stochastic block model for multilevel network can be written as follows:

$$\begin{aligned}\mathcal{I}_\theta(\mathcal{R}(Z^I, Z^O|A)) &= \sum_{j,l} \tau^O_{jl} \log \pi^O_l + \sum_{i,k} \tau^I_{ik} \sum_{j,l} A_{ij} \tau^O_{jl} \log \gamma_{kl} \\ &+ \frac{1}{2} \sum_{i' \neq i} \sum_{k,k'} \tau^I_{ik} \tau^I_{i'k'} \log \phi\left(X^I_{ii'}, \alpha^I_{kk'}\right) + \frac{1}{2} \sum_{j' \neq j} \sum_{l,l'} \tau^O_{jl} \tau^O_{j'l'} \log \phi\left(X^O_{jj'} \alpha^O_{ll'}\right) \\ &- \sum_{i,k} \tau^I_{ik} \log \tau^I_{ik} - \sum_{j,l} \tau^O_{jl} \log \tau^O_{jl}\end{aligned}$$

The variational EM algorithm then consists on iterating the two following steps. At iteration $(t+1)$:

VE step compute

$$\begin{aligned}\{\tau^I, \tau^O\}^{(t+1)} &= \arg\max_{\tau^I, \tau^O} \mathcal{I}_{\theta^{(t)}}(\mathcal{R}(Z^I, Z^O|A)) \\ &= \arg\min_{\tau^I, \tau^O} \mathrm{KL}\left(\mathcal{R}(Z^I, Z^O|A) \| \mathbb{P}_{\theta^{(t)}}(Z^I, Z^O|X^I, X^O, A)\right) .\end{aligned}$$

M step compute

$$\theta^{(t+1)} = \arg\max_\theta \mathcal{I}_\theta(\mathcal{R}^{(t+1)}(Z^I, Z^O|A)).$$

The variational parameters are sought by solving the equation:

$$\Delta_{\tau^I, \tau^O}\left(\mathcal{I}_\theta(\mathcal{R}(Z^I, Z^O|A) + L(\tau^I, \tau^O)\right) = 0,$$

where $L(\tau^I, \tau^O)$ are the Lagrange multipliers for τ^I_i, τ^O_j for all $i \in \{1, \ldots, n_I\}$, $j \in \{1, \ldots, n_I\}$. There is no closed-form formula but when computing the derivatives, we obtain that the variational parameters follow the fixed point relationships:

$$\begin{aligned}\widehat{\tau^O_{jl}} &\propto \pi^O_l \prod_{i,k} \gamma_{kl}^{A_{ij} \widehat{\tau^I_{ik}}} \prod_{j' \neq j} \prod_{l'} \phi(X^O_{jj'}, \alpha^O_{ll'})^{\widehat{\tau^O_{j'l'}}} \\ \widehat{\tau^I_{ik}} &\propto \prod_{j,l} \gamma_{kl}^{A_{ij} \widehat{\tau^O_{jl}}} \prod_{i' \neq i} \prod_{k'} \phi(X^I_{ii'}, \alpha^I_{kk'})^{\widehat{\tau^I_{i'k'}}} ,\end{aligned}$$

which are used in the VE step to update the τ^I_i's and τ^O_j's.

On each update, the variational parameters of a certain level depend on both the parameter γ and the variational parameters of the other level, which emphasizes the dependency structure of this multilevel model and the role of γ as the dependency parameter of the model. Notice also that when $\gamma_{kl} = \gamma_{kl'} = \pi^I_k$ for all l, l', that is the case of independence between the two levels then we can rewrite the fixed point relationships as follows:

$$\widehat{\tau^O_{jl}} \propto \pi^O_l \prod_{j' \neq j} \prod_{l'} \phi(X^O_{jj'}, \alpha^O_{ll'})^{\widehat{\tau^O_{j'l'}}} \quad \text{and} \quad \widehat{\tau^I_{ik}} \propto \pi^I_k \prod_{i' \neq i} \prod_{k'} \phi(X^I_{ii'}, \alpha^I_{kk'})^{\widehat{\tau^I_{i'k'}}},$$

which is exactly the expression of the fixed point relationship of two independent SBMs. Then, for the M step, we derive the following closed-form formulae:

$$\widehat{\pi_l^O} = \frac{1}{n_O}\sum_j \widehat{\tau_{jl}^O} \qquad \widehat{\alpha_{ll'}^O} = \frac{\sum_{j'\neq j}\widehat{\tau_{jl}^O} X_{jj'}^O \widehat{\tau_{j'l'}^O}}{\sum_{j'\neq j}\widehat{\tau_{jl}^I}\widehat{\tau_{j'l'}^I}}$$

$$\widehat{\gamma}_{kl} = \frac{\sum_{i,j}\widehat{\tau_{ik}^I} A_{ij}\widehat{\tau_{jl}^O}}{\sum_{i,j} A_{ij}\widehat{\tau_{jl}^O}} \qquad \widehat{\alpha_{kk'}^I} = \frac{\sum_{i'\neq i}\widehat{\tau_{ik}^I} X_{ii'}^I \widehat{\tau_{i'k'}^I}}{\sum_{i'\neq i}\widehat{\tau_{ik}^I}\widehat{\tau_{i'k'}^I}}$$

for which the gradient

$$\Delta_\theta\left(\mathcal{I}_\theta(\mathcal{R}(Z^I, Z^O|A)) + L(\pi^O, \gamma)\right),$$

is null. The term $L(\pi^O, \gamma)$ contains the Lagrange multipliers for π^O and $\gamma_{k\cdot}$ for all $k \in \{1, \ldots, Q_I\}$.

Model parameters have natural interpretations. π_l^O is the mean of the posterior probabilities for the organizations to belong to block l. $\alpha_{kk'}^I$ (resp. $\alpha_{ll'}^O$) is the ratio of existing links over possible links between blocks k and k' (resp. l and l'). γ_{kl} is the ratio of the number of individuals in block k that are affiliated to any organization of block l on the number of individuals that are affiliated to any organization of block l. If γ is such that the levels are independent, then any column of γ represents the proportion of individuals in the different blocks:

$$\pi_k^I = \gamma_{k1} = \frac{1}{n_I}\sum_i \widehat{\tau_{ik}^I}.$$

2.D. Details of the ICL criterion

We now derive an expression for the Integrated Complete Likelihood (ICL) model selection criterion. Following Daudin et al. (2008), the ICL is based on the integrated complete likelihood i.e. the likelihood of the observations and the latent variables where the parameters have been integrating out against a prior distribution. The latent variables (Z^I, Z^O) being unobserved, they are imputed using the maximum a posteriori (MAP) or $\hat{\tau}$. We denote by $\widehat{Z^O}$ and $\widehat{Z^I}$ the inputed latent variables. After imputation of the latent variables, an asymptotic approximation of this quantity leads to the ICL criterion given in the paper (Equation (2.9)) and recalled here:

$$\begin{aligned} ICL(Q_I, Q_O) = {} & \log \ell_{\widehat{\theta}}(X^I, X^O, \widehat{Z^I}, \widehat{Z^O}|A, Q_I, Q_O) \\ & - \frac{1}{2}\frac{Q_I(Q_I+1)}{2}\log\frac{n_I(n_I-1)}{2} - \frac{Q_O(Q_I-1)}{2}\log n_I \\ & - \frac{1}{2}\frac{Q_O(Q_O+1)}{2}\log\frac{n_O(n_O-1)}{2} - \frac{Q_O-1}{2}\log n_O. \end{aligned}$$

Let $\Theta = \Pi^O \times \mathcal{A}^I \times \mathcal{A}^O \times \Gamma$ be the space of the model parameters. We set a prior distribution on θ:

$$p(\theta|Q_I, Q_O) = p(\gamma|Q_I, Q_O)p(\pi^O|Q_O)p(\alpha^I|Q_I)p(\alpha^O|Q_O)$$

where $p(\pi^O|Q_O)$ is a Dirichlet distribution of hyper-parameter $(1/2, \cdots, 1/2)$ and $p(\alpha^I|Q_I)$ and $p(\alpha^O|Q_O)$ are independent Beta distributions.
The marginal complete likelihood is written as follows:

$$\begin{aligned}
\log \ell_\theta(\mathbf{X}, \mathbf{Z}|A, Q_I, Q_O) &= \log\left(\int_\Theta \ell_\theta(X^I, X^O, Z^I, Z^O|\theta, A, Q_I, Q_O)p(\theta|Q_I, Q_O)\mathrm{d}\theta\right) \\
&= \log \ell_{\alpha^I}(X^I|Z^I, Q_I) && (2.D.22)\\
&\quad + \log \ell_\gamma(Z^I|A, Z^O, Q_I, Q_O) && (2.D.23)\\
&\quad + \log \ell_{\alpha^O,\pi^O}(X^O, Z^O|Q_O)\,. && (2.D.24)
\end{aligned}$$

The quantity defined in (2.D.24) evaluated at $Z^O := \widehat{Z^O}$ is approximated as in Daudin et al. (2008) by

$$\begin{aligned}
\log \ell_{\alpha^O}(X^O, \widehat{Z^O}, Q_O) &\approx_{n_O\to\infty} \log \ell_{\widehat{\alpha^O},\widehat{\pi^O}}(X^O, \widehat{Z^O}|Q_O) - \mathrm{pen}(\pi^O, \alpha^O, Q_O) \\
\mathrm{pen}(\pi^O, \alpha^O, Q_O) &= \tfrac{Q_O-1}{2}\log n_O + \tfrac{1}{2}\tfrac{Q_I(Q_I+1)}{2}\log\tfrac{n_I(n_I-1)}{2}
\end{aligned} . \qquad (2.D.25)$$

This approximation results from a BIC-type approximation of $\log \ell_{\widehat{\alpha^O}}(X^O|\widehat{Z^O}, Q_O)$ and a Stirling approximation of $\log \ell_{\pi^O}(\widehat{Z^O}, Q_O)$.
The same BIC-type approximation on $\log \ell_{\alpha^I}(X^I|\widehat{Z^I}, Q_I)$ (Equation (2.D.22)) leads to:

$$\begin{aligned}
\log \ell_{\alpha^I}(X^I|\widehat{Z^I}, Q_I) &= {}_{n_I\to\infty} \log \ell_{\widehat{\alpha^I}}(X^I|\widehat{Z^I}, Q_I) + \mathrm{pen}(\alpha^I, Q_I) \\
\text{with } \mathrm{pen}(\alpha^I, Q_I) &= \tfrac{1}{2}\tfrac{Q_I(Q_I+1)}{2}\log\tfrac{n_I(n_I-1)}{2}
\end{aligned} . \qquad (2.D.26)$$

For quantity (2.D.23) depending on γ and Z^I given (Q_I, Q_O), we have to adapt the calculus. Let us set independent Dirichlet prior distributions of order Q_I $\mathcal{D}(1/2, \ldots, 1/2)$ on the columns $\gamma_{\cdot l}$. We are able to derive an exact expression of $\log \ell_\gamma(Z^I|A, Z^O, Q_I, Q_O)$:

$$\begin{aligned}
\ell_\gamma(Z^I|A, Z^O, Q_I, Q_O) &= \int \ell(Z^I|A, Z^O, \gamma, Q_I, Q_O)p(\gamma, Q_I, Q_O)\mathrm{d}\gamma \\
&= \prod_i \int \prod_{j,k,l} \gamma_{kl}^{A_{ij}Z^I_{ik}Z^O_{jl}} p(\gamma_{kl})\mathrm{d}\gamma_{kl} \\
&= \prod_l \int \prod_k \gamma_{kl}^{N_{kl}} p(\gamma_{kl})\mathrm{d}\gamma_{kl}, \quad \text{where} \quad N_{kl} = \sum_{ij} A_{ij}Z^I_{ik}Z^O_{jl} \\
&= \prod_l \int \prod_k \gamma_{k,l}^{N_{kl}+a-1} \frac{\Gamma(1/2\cdot Q_I)}{\Gamma(1/2)^{Q_I}}\mathrm{d}\gamma_{kl} \\
&= \frac{\Gamma(1/2Q_I)^{Q_O}}{\Gamma(1/2)^{Q_O+Q_I}} \prod_l \frac{\prod_k \Gamma(N_{kl}+1/2)}{\Gamma(1/2Q_I + \sum_k N_{kl})}\,.
\end{aligned}$$

Now, using the fact that $\log \Gamma(n+1) \overset{n\to\infty}{\sim} (n+1/2)\log n + n$, we obtain:

$$\begin{aligned}\log \ell_\gamma(Z^I|A, Z^O, Q_I, Q_O) \quad &\approx \quad {}_{(n_O,n_I)\to\infty} \textstyle\sum_{k,l}(N_{kl}\log N_{kl} + N_{kl}) \\ &\quad - \textstyle\sum_l \left(\frac{Q_I-1}{2} + \sum_k N_{kl}\right)\log\left(\sum_k N_{kl}\right) - \sum_{k,l} N_{kl}.\end{aligned} \tag{2.D.27}$$

The quantity (2.D.27) evaluated at $(Z^I, Z^O) := (\widehat{Z^I}, \widehat{Z^O})$ can be reformulated in the following way:

$$\begin{aligned}\log \ell_\gamma(\widehat{Z^I}|A, \widehat{Z^O}, Q_I, Q_O) \quad &\approx \quad {}_{(n_O,n_I)\to\infty} \log \ell_{\hat\gamma}(\widehat{Z^I}|A, \widehat{Z^O}, Q_I, Q_O) - \tfrac{Q_I-1}{2}\textstyle\sum_l \log \sum_{i,j} A_{ij}\widehat{Z^O_{jl}} \\ \text{with } \hat\gamma_{kl} \quad &= \quad \frac{\sum_{i,j}\widehat{Z^I_{ik}}A_{ij}\widehat{Z^O_{jl}}}{\sum_{i,j}A_{ij}\widehat{Z^O_{jl}}}\end{aligned}$$

Noticing that $\log \sum_{i,j} A_{ij}\widehat{Z^O_{jl}} = \log n_I + \log \frac{\sum_{i,j}A_{ij}\widehat{Z^O_{jl}}}{n_I} = O(\log n_I)$ leads to

$$\log \ell_\gamma(\widehat{Z^I}|A, \widehat{Z^O}, Q_I, Q_O) \approx_{(n_O,n_I)\to\infty} \log \ell_{\hat\gamma}(\widehat{Z^I}|A, \widehat{Z^O}, Q_I, Q_O) - \frac{Q_I-1}{2}Q_O \log n_I\,. \tag{2.D.28}$$

Combining Equations (2.D.25), (2.D.26) and (2.D.28) we obtain the given expression.

2.E. Stochastic Block Model for Generalized Multilevel Network

In this section, I propose a generalization of MLVSBM to networks with more than 2 levels and any number of affiliations. The idea is to generalize the hierarchical model of the interactions given the affiliations developed before.

2.E.1. Description of the generative model

Let L bet he number of levels, n_l the number of nodes in level l, and Q_l the number of blocks in level l, $l \in \{1, \dots, L\}$. Level L corresponds to the highest level in the hierarchy (the organizational level in a 2 level multilevel network). Let A^l be a matrix of size $n_l \times n_{l+1}$ such that for all $l \in \{1, \dots, L-1\}$:

$$\sum_{j=1}^{n_{l+1}} A^l_{ij} \in \{0,1\}, \quad A^l_{ij} \geq 0.$$

(A^l_{ij}) can be seen as the proportion of time actor i of level l dedicates to actor j of level $l+1$. Each row of $A^{l-1,l}$ sums either to 1 if the actor of the row i has one or multiple affiliations or to 0 if it has none.

I define the generative model as follows:

1. Draw Z^L, the latent blocks of level L under the following iid distribution:

$$\mathbb{P}(Z^L_i = q) = \pi^L_q, \quad i \in \{1, \dots, n_L\}, \quad q \in \{1, \dots, Q_L\}.$$

2. Draw for each $i \in \{1, \dots, n_{L-1}\}$, the latent affiliation $W^{L-1} \in \{0,1\}^{n_{L-1}, n_L+1}$ that will be used to determine the block memberships for level $L-1$. For all $j \in \{0, 1, \dots, n_L\}$:
$$\mathbb{P}(W_i^{L-1} = j) = A_{ij}^{L-1}$$

3. Draw Z_i^{L-1} the blocks of level $L-1$ independently for all $i \in \{1, \dots, n_{L-1}\}$ with probability:
$$\mathbb{P}(Z_i^{L-1} = q | W_i^{L-1}, Z^L) = \prod_{r=1}^{Q_L} (\gamma_{kr}^{L-1})^{\sum_{j \geq 1} W_{ij}^{L-1} Z_{jr}^L},$$

4. Iterate step 2 and 3 until W^1 et Z^1 are drawn

5. Draw the intra-level link for all $(i \neq j) \in \{1, \dots, n_l\}^2$, $l \in \{1, \dots, L\}$ independently under the following distribution:
$$\mathbb{P}(X_{ij}^l = x | Z_i^l, Z_j^l) = f^l(x; \alpha_{Z_i^l Z_j^l}^l). \tag{2.E.29}$$

A DAG depicting the the generative model is given in Figure 2.8

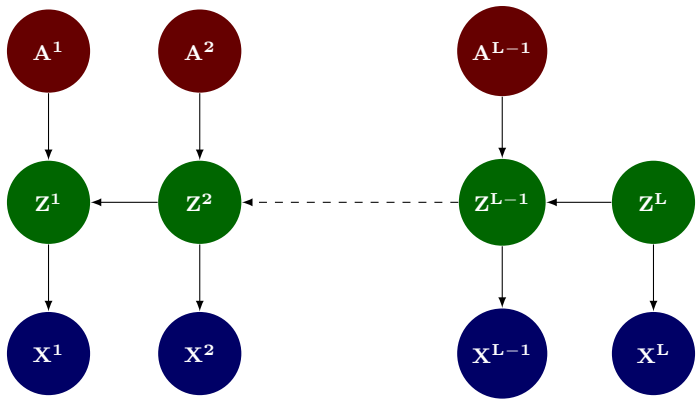


Figure 2.8 – DAG of the generalized Stochastic Block Model for Multilevel Network (gMLVSBM)

The MLVSBM corresponds to the above model with $L = 2$, $A^{1,2}$ with just one 1 per row and f^l a Bernoulli, $l \in \{1, 2\}$.

Dealing with actors with no affiliation It is possible to extend the model, so that it handles actors with no affiliation, ie. $(i, l) : \sum_j A_{ij}^l = 0$. To do so, we add another column to the matrix of affiliation, such that $A_{i,n_{l+1}+1}^l := \mathbb{1}_{\sum_j A_{ij}^l = 0}$. Then, each row of the matrix A^l sums to one.

We do the same with the mixture parameter γ^l, so that it becomes for $l \in \{1, \dots, L-1\}$, a $Q_l \times Q_{l+1} + 1$ matrix with each column summing to one. The row $Q_{l+1} + 1$ of γ^l is a probability vector of length Q_l, giving the mixture parameter for actor of level l with no affiliation.

Likelihood of the model From now on, we will assume that the levels are undirected. To ease the notations, we will not consider actors with no affiliation on level $l < L$, but the extension is straightforward.

The likelihood of the model can be written as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{X};\mathbf{A}) &= \int_{\mathbf{Z}}\int_{\mathbf{W}} \mathcal{L}(\mathbf{X}|\mathbf{Z},\mathbf{W})\mathcal{L}(\mathbf{Z},\mathbf{W};\mathbf{A})\mathrm{d}\mathbf{W}\mathrm{d}\mathbf{Z} \\ &= \int_{\mathbf{Z}}\int_{\mathbf{W}} \prod_{l=1}^{L}\mathcal{L}(X^l|Z^l)\prod_{l=1}^{L-1}\mathcal{L}(Z^l|W^l,Z^{l+1})\mathcal{L}(W^l;A^l)\mathcal{L}(Z^L)\mathrm{d}\mathbf{W}\mathrm{d}\mathbf{Z},\end{aligned}$$

where:

$$\begin{aligned}\mathcal{L}(X^l|Z^l) &= \prod_{i\neq j}\prod_{q,r}(\alpha^l_{qr})^{X^l_{ij}Z^l_{iq}Z^l_{jr}}(1-\alpha^l_{qr})^{(1-X^l_{ij})Z^l_{iq}Z^l_{jr}} \\ \mathcal{L}(Z^l|W^l,Z^{l+1}) &= \prod_{i}\prod_{q,r}\gamma^l_{qr}{}^{\sum_{j\geq 1}W^l_{ij}Z^l_{iq}Z^{l+1}_{jr}} \\ \mathcal{L}(W^l;A^l) &= \prod_{i,j}A^l_{ij}{}^{W^l_{ij}} \\ \mathcal{L}(Z^L) &= \prod_{i,q}\pi^L_q{}^{Z^L_{iq}}\end{aligned}$$

2.E.2. Variational inference

As in the SBM and the MLVSBM, this integral is not tractable. To estimate the maximum likelihood, we will rely on a variational EM algorithm. The trick is to use a variational distribution on W^l, for all $l \in \{1,\dots,L-1\}$, to make it then independent on (Z^l, Z^{l+1}). As A^l is already known, there is no need to infer those parameters.

$$\begin{aligned}\ell(\mathbf{X};\mathbf{A}) &\geq \ell(\mathbf{X};\mathbf{A}) - \mathrm{KL}(\mathcal{R}(\mathbf{Z},\mathbf{W};\mathbf{A})\|p(\mathbf{Z},\mathbf{W}|\mathbf{X};\mathbf{A})) \\ &= \mathbb{E}_{\mathcal{R}}\left[\ell\left(\mathbf{Z},\mathbf{X},\mathbf{W};\mathbf{A}\right)\right] + \mathcal{H}\left(\mathcal{R}(\mathbf{Z},\mathbf{W};\mathbf{A})\right) \\ &=: \mathcal{J}(\mathcal{R}(\mathbf{Z},\mathbf{W};\mathbf{A})),\end{aligned}$$

where we choose $\mathcal{R}$ in a family of factorizable distribution of the type:

$$\mathcal{R}(\mathbf{Z},\mathbf{W};\mathbf{A}) = \prod_{l=1}^{L}\prod_{i=1}^{n_l}\mathcal{R}(Z^l_i)\prod_{l=1}^{L-1}\prod_{i}^{n_l}\mathcal{R}(W^l_i),$$

with

$$\mathbb{E}_{\mathcal{R}}[Z^l_{iq}] = \tau^l_{iq} \quad \text{and} \quad \mathbb{E}_{\mathcal{R}}[W^l_{ij}] = \omega^l_{ij}.$$

The complete expression of the variational bound is the following:

$$\begin{aligned}\mathcal{J}(\mathcal{R}(\mathbf{Z},\mathbf{W};\mathbf{A})) &= \frac{1}{2}\sum_{l=1}^{L}\sum_{i\neq i'}\sum_{q,r}\tau_{iq}^{l}\tau_{jr}^{l}\log f^{l}(X_{ii'}^{l},\alpha_{qr}^{l}) \\ &+ \sum_{l=1}^{L-1}\sum_{i}\sum_{q}\tau_{iq}^{l}(\sum_{j\geq 1}\sum_{r}\omega_{ij}^{l,l+1}\tau_{jr}^{l+1}\log\gamma_{qr}^{l}) \\ &+ \sum_{i,q}\tau_{iq}^{L}\log\pi_{q}^{L}-\sum_{l=1}^{L}\sum_{i,k}\tau_{iq}^{l}\log\tau_{iq}^{l}\end{aligned}$$

where we choose $\omega_{ij}^{l}=A_{ij}^{l}$ for all $l\in\{1,\ldots,L-1\}$, $i\in\{1,\ldots,n_l\}$, $j\in\{1,\ldots,n_{l+1}\}$.

Parameters update

VE-Step We need 3 different formulae: one for $l=1$, the other for $l\in\{2,\ldots,L-1\}$ and a last one for $l=L$. Maximizing the gradient, using Lagrange multiplier to consider the constrain on τ, we obtain the following fixed point equation:

$$\begin{aligned}\widehat{\tau_{iq}^{l}} \propto & \prod_{j\neq i}\prod_{r} f^{l}(X_{ij}^{l},\alpha_{qr}^{l})^{\widehat{\tau_{jr}^{l}}} \\ & \times\left(\prod_{r}\gamma_{qr}^{l\,\sum_{j\geq 1}A_{ij}^{l}\widehat{\tau_{jr}^{l+1}}}\right)^{\mathbb{1}_{l<L}} \\ & \times\left(\prod_{k}\gamma_{kq}^{l-1\,\sum_{j\geq 1}A_{ji}^{l-1}\widehat{\tau_{jk}^{l}}}\right)^{\mathbb{1}_{l>1}} \\ & \times\pi_{q}^{l\,\mathbb{1}_{l=L}},\end{aligned}$$

hence the variational block membership parameters of level l, depends on the varational block membership parameters of level $l-1$, l and $l+1$.

M-step The update of α^l, $l\in\{1,\ldots,L\}$ are the same as the standard SBM or MLVSBM. It depends on the emission distribution, while we obtain new formulae for γ^l, $l<L$ and π^L. They are the following:

$$\widehat{\pi_{q}^{L}}=\frac{1}{n_L}\sum_{i=1}^{n_L}\widehat{\tau_{iq}^{L}} \qquad \widehat{\gamma_{qr}^{l}}=\frac{\sum_{i=1}^{n_l}\sum_{j=1}^{n_l+1}\widehat{\tau_{iq}^{l}}A_{ij}^{l,l+1}\widehat{\tau_{jq}^{l+1}}}{\sum_{i=1}^{n_l}\sum_{j=1}^{n_l+1}A_{ij}^{l,l+1}\widehat{\tau_{jq}^{l+1}}} \qquad \text{if } l<L$$

The multi-affiliation part of this work is implemented in the `R` package `MLVSBM` for multilevel networks with 2 levels.

2.F. MLVSBM package Tutorial

```
library(MLVSBM)
```

The package deals with multilevel networks defined as the junction of two interaction network (adjacency matrices) linked by an affiliation relationship (affiliation matrix). The notations used in the package are the one from the research paper (Chabert-Liddell et al., 2021).

First, we simulate a multilevel network with 100 individuals and 3 clusters of individuals for the lower level; 50 organizations and 3 clusters for the upper level. The inter-organizational level will have an assortative structure and will be undirected, the inter-individual's one a core-periphery structure and will be directed. Affiliation matrix will be generated by a power law and the dependency between the latent blocks of the two levels will be strong.

```
set.seed(123)
my_mlvsbm <- MLVSBM::mlvsbm_simulate_network(
  n = list(I = 60, O = 40), # Number of nodes for the  lower level
                            # and the upper level
  Q = list(I = 3, O = 3), # Number of blocks for the lower level a
                          # and the upper level
  pi = c(.5, .3, .2), # Block proportion for the upper level, must sum to one
  gamma = matrix(c(.8, .1, .1,  # Block proportion for the lower level,
                   .1, .8, .1,
                   .1, .1, .8), # each column must sum to one
                 nrow = 3, ncol = 3, byrow = TRUE),
  alpha = list(I = matrix(c(.1, .1, .3,
                            .1, .2, .5,
                            .1, .5, .5),
                          nrow = 3, ncol = 3, byrow = TRUE), # Connection matrix
               O = matrix(c(.5, .1, .1,
                            .1, .5, .1,
                            .1, .1, .5),
                          nrow = 3, ncol = 3, byrow = TRUE)),# between blocks
  directed = list(I = TRUE, O = FALSE), # Are the upper and lower level
                                        # directed or not ?
  affiliation = "preferential", # How the affiliation matrix is generated
  no_empty_org = FALSE) # May the affiliation matrix have column suming to 0
```

The network is stocked in an `R6` object of type `MLVSBM`.

Now, we create a multilevel network object from 2 existing adjacency matrix and an affiliation matrix. The lower level correspond to the inter-individual level while the upper level is the inter-organizational level:

```
lower_level <- my_mlvsbm$adjacency_matrix$I # matrix of size nI x nI
upper_level <- my_mlvsbm$adjacency_matrix$O # matrix of size nO x nO
affiliation <- my_mlvsbm$affiliation_matrix # matrix of size nI x nO
my_mlvsbm2 <- mlvsbm_create_network(X = list(I = lower_level, O = upper_level),
                                    A = affiliation)
```

We can now infer the parameters, blocks and edge probabilities of our network

by using the `mlvlsbm_estimate_network()` function on an `MLVSBM` object. It will return the best model for this network as another R6 object of type `FitMLVSBM`.

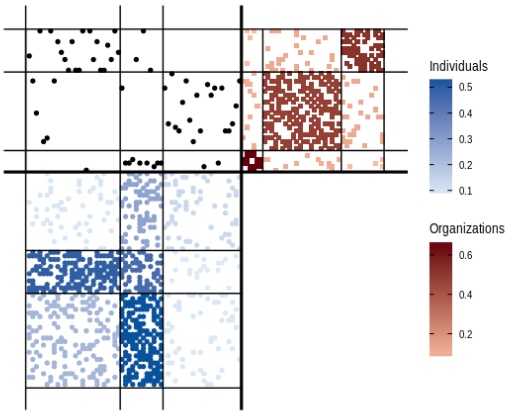
```
fit <- MLVSBM::mlvsbm_estimate_network(my_mlvsbm, nb_cores = 1L)
#>
#> [1] "Infering lower level :"
#> [1] "# blocks: 3, ICL = -1753.45 !"
#>
#> [1] "Infering upper level :"
#> [1] "# blocks: 3, ICL = -429.33 !"
#> [1] "=== # Individual blocks: 3 , # Organizational blocks: 3,  ICL : -2161.44==="
#> [1] "=== # Individual blocks: 3 , # Organizational blocks: 3,  ICL : -2161.44==="
#> [1] "ICL for independent levels : -2182.78"
#> [1] "ICL for interdependent levels : -2161.44"
#> [1] "=====Interdependence is detected between the two levels!====="
```

2.F.1. Generic functions

Generic functions are provided to print, plot, extract the model parameters and predict the existence of a dyad for the fitted network.

```
print(fit)
#> Multilevel Stochastic Block Model -- bernoulli variant
#> ======================================================================
#> Dimension = ( 60 40 ) - ( 3 3 ) blocks.
#> ======================================================================
#> * Useful fields
#>   $independent, $distribution, $nb_nodes, $nb_clusters, $Z
#>   $membership, $parameters, $ICL, $vbound, $X_hat
# plot(fit, type = "matrix")
coef(fit)
#> $alpha
#> $alpha$I
#>            [,1]      [,2]       [,3]
#> [1,] 0.23072817 0.5382933 0.09444543
#> [2,] 0.49659089 0.4545821 0.08339372
#> [3,] 0.09614086 0.2840693 0.13421325
#>
#> $alpha$O
#>            [,1]       [,2]      [,3]
#> [1,] 0.66670837 0.09845986 0.0557177
#> [2,] 0.09845986 0.48056388 0.1137974
#> [3,] 0.05571770 0.11379738 0.5299402
#>
#>
#> $pi
#> $pi$O
#> [1] 0.1499768 0.5496535 0.3003697
#>
#>
#> $gamma
#>            [,1]       [,2]          [,3]
#> [1,] 0.09064248 0.19243672 8.692712e-01
#> [2,] 0.72734754 0.03843131 1.307266e-01
```

```
#> [3,] 0.18200998 0.76913196 2.105228e-06
pred <- predict(fit)
pred$nodes$O
#>O1  O2  O3  O4  O5  O6  O7  O8  O9 O10 O11 O12 O13 O14 O15 O16 O17 O18 O19 O20
#> 3   2   2   2   1   2   1   1   3   2   3   3   3   2   2   2   2   3   2   2
#>O21 O22 O23 O24 O25 O26 O27 O28 O29 O30 O31 O32 O33 O34 O35 O36 O37 O38 O39 O40
#>  2   3   2   3   2   2   1   1   1   2   2   3   2   3   2   2   3   2   2   3
pred$dyads$I[1:2,1:12]
#>       I1    I2    I3    I4    I5    I6    I7    I8    I9   I10   I11   I12
#> I1 0.000 0.094 0.094 0.231 0.094 0.538 0.231 0.538 0.231 0.094 0.231 0.094
#> I2 0.096 0.000 0.134 0.096 0.134 0.284 0.096 0.284 0.096 0.134 0.096 0.134
plot(fit, type = "matrix"))
```



2.F.2. Other useful output

Output of the algorithm are stocked in the `MLVSBM` and `FitMLVSBM` objects. The `MLVSBM` object stocks information of the observed or simulated network and a list of all the fitted SBM and MLVSBM models.

```
# A data frame of the inferred models
my_mlvsbm$ICL
#>   index Q_I Q_O       ICL
#> 1     1   3   3 -2161.445
# The fitted model with index  the highest ICL
my_fit <- my_mlvsbm$fittedmodels[[which.max(my_mlvsbm$ICL$ICL)]]

# A fitted SBM for the lower level with 3 blocks
my_sbm_lower <- my_mlvsbm$fittedmodels_sbm$lower[[3]]
# A fitted SBM for the upper level with 2 blocks
my_sbm_upper <- my_mlvsbm$fittedmodels_sbm$upper[[2]]
```

You can also get the parameters and the clustering of the fitted model from the `FitMLVSBM` object as follows:

```
fit$parameters # The connectivity and membership parameters of the model
#> $alpha
#> $alpha$I
```

```
#>            [,1]      [,2]       [,3]
#> [1,] 0.23072817 0.5382933 0.09444543
#> [2,] 0.49659089 0.4545821 0.08339372
#> [3,] 0.09614086 0.2840693 0.13421325
#>
#> $alpha$O
#>            [,1]       [,2]      [,3]
#> [1,] 0.66670837 0.09845986 0.0557177
#> [2,] 0.09845986 0.48056388 0.1137974
#> [3,] 0.05571770 0.11379738 0.5299402
#>
#>
#> $pi
#> $pi$O
#> [1] 0.1499768 0.5496535 0.3003697
#>
#>
#> $gamma
#>            [,1]       [,2]         [,3]
#> [1,] 0.09064248 0.19243672 8.692712e-01
#> [2,] 0.72734754 0.03843131 1.307266e-01
#> [3,] 0.18200998 0.76913196 2.105228e-06
fit$Z # The block membership of each nodes
#> $I
#>  I1  I2  I3  I4  I5  I6  I7  I8  I9 I10 I11 I12 I13 I14 I15 I16 I17 I18 I19 I20
#>   1   3   3   1   3   2   1   2   1   3   1   3   3   1   3   3   1   1   3   1
#> I21 I22 I23 I24 I25 I26 I27 I28 I29 I30 I31 I32 I33 I34 I35 I36 I37 I38 I39 I40
#>   1   1   3   1   3   3   1   1   1   1   3   1   2   2   1   3   2   3   3   3
#> I41 I42 I43 I44 I45 I46 I47 I48 I49 I50 I51 I52 I53 I54 I55 I56 I57 I58 I59 I60
#>   2   1   1   1   2   2   1   3   3   1   2   1   3   1   2   2   2   1   3   3
#>
#> $O
#>  O1  O2  O3  O4  O5  O6  O7  O8  O9 O10 O11 O12 O13 O14 O15 O16 O17 O18 O19 O20
#>   3   2   2   2   1   2   1   1   3   2   3   3   3   2   2   2   2   3   2   2
#> O21 O22 O23 O24 O25 O26 O27 O28 O29 O30 O31 O32 O33 O34 O35 O36 O37 O38 O39 O40
#>   2   3   2   3   2   2   1   1   1   2   2   3   2   3   2   2   3   2   2   3
fit$vbound # A vector of the varational bound of the VEM algorithm
#> [1] -2088.490 -2088.447 -2088.447 -2088.447
tau <- fit$membership # The variational parameters of the model
pred <- fit$X_hat # The links predictions for each level
```

2.G. Hard to Infer Levels: Benefits of the Multilevel Modeling

```
library(MLVSBM)
library(aricode)
library(dplyr)
library(ggplot2)
library(tidyr)
```

```
library(pbmcapply)
```

2.G.1. Simulation Scenario

In this vignette we are interested in the behavior of our inference algorithm when both level of our multilevel network are hard to infer.

For a standard Stochastic Block Model, a well known result is the detectability threshold when the mesoscale structure is the one of a planted partition (Decelle et al., 2011), i.e. the connectivity parameter matrix has the following shape:

$$\begin{pmatrix} p & q \\ q & p \end{pmatrix},$$

where p and q are the connection probabilities for respectively an intra-block interaction and an extra-block interaction.

Let define a multilevel network with 120 individuals belonging to one of 2 blocks of individuals and 40 organizations belonging to one of 2 blocks of organizations of equal size on average.

```
Q     <- list(I = 2, O = 2)
n     <- list(I = 60*Q$I, O = 20*Q$O)
pi      <- rep(1/Q$O, Q$O)
gamma   <- .1 * (diag(8, Q$I, Q$O) + 2/Q$I)
```

Then the detectability threshold for each level is given by:

```
detect_threshold <- function(n, q) {
  polyroot(c(n*q*(n*q-2),-2*n*(n*q+1),n**2 ))
}
```

```
Re(detect_threshold(120, .1))
#> [1] 0.0500000 0.1666667
```

for the inter-individual level and

```
Re(detect_threshold(40, .1))
#> [1] 0.02192236 0.22807764
```

for the inter-organizational level.

So we will fix the connectivity parameters a little bit above the detectability threshold in order to get a very challenging inference for a single level SBM and to see how the information in each level of the MLVSBM might help to recover the structure of the other level, and how does it vary with the strength of the inter-level dependence.

```
alpha <- list()
alpha$I <- matrix(c(.1, .21, .21, .1), 2, 2)
alpha$O <- matrix(c(.29, .1, .1, .29), 2, 2)
```

Then simulate 50 networks for each value of γ. Let recall that γ is the mixture parameter for the block membership of the individuals given the one of their

respective organizations, for $g \in [.5, 1]$,

$$\begin{pmatrix} g & 1-g \\ 1-g & g \end{pmatrix}.$$

So the two levels are independent when $g = .5$ and the block membership of an individual is entirely determine by the block membership of his/her organization when $g = 1$.

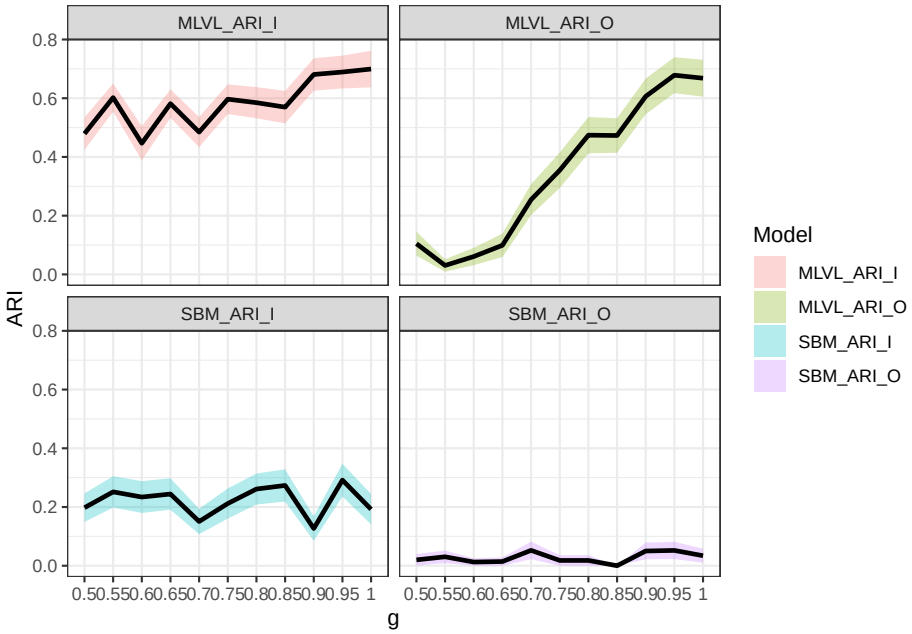
```
set.seed(42)
res_detect_thresh <- tibble()
for (g in seq(.5, 1, .05)) {
  gamma <-   matrix(c(g, 1-g, 1-g, g), 2, 2)
  res <- pbmcapply::pbmclapply(
    X = seq(50L),
    FUN = function(i) {
      mlvl <-
        MLVSBM::mlvsbm_simulate_network(
          n = n, Q = Q, pi = pi, gamma = gamma, alpha = alpha,
          directed = list(I = FALSE, O = FALSE))
      fit <- MLVSBM::mlvsbm_estimate_network(mlvl, nb_cores = 2L)
      return(tibble(
        "SBM_ARI_I" = aricode::ARI(
            c1 = mlvl$memberships$I,
            c2 = mlvl$fittedmodels_sbm$lower[[which.max(mlvl$ICL_sbm$lower)]]$Z),
        "SBM_ARI_O" = aricode::ARI(
            c1 = mlvl$memberships$O,
            c2 = mlvl$fittedmodels_sbm$upper[[which.max(mlvl$ICL_sbm$upper)]]$Z),
        "MLVL_ARI_I" = aricode::ARI(
            c1 = mlvl$memberships$I,
            c2 = fit$Z$I),
        "MLVL_ARI_O" = aricode::ARI(
            c1 = mlvl$memberships$O,
            c2 = fit$Z$O)))
    },
    mc.cores = 6L)
  res <- bind_rows(res)
  res$g <- as.factor(g)
  res_detect_thresh <- bind_rows(res_detect_thresh, res)
}
```

2.G.2. Results

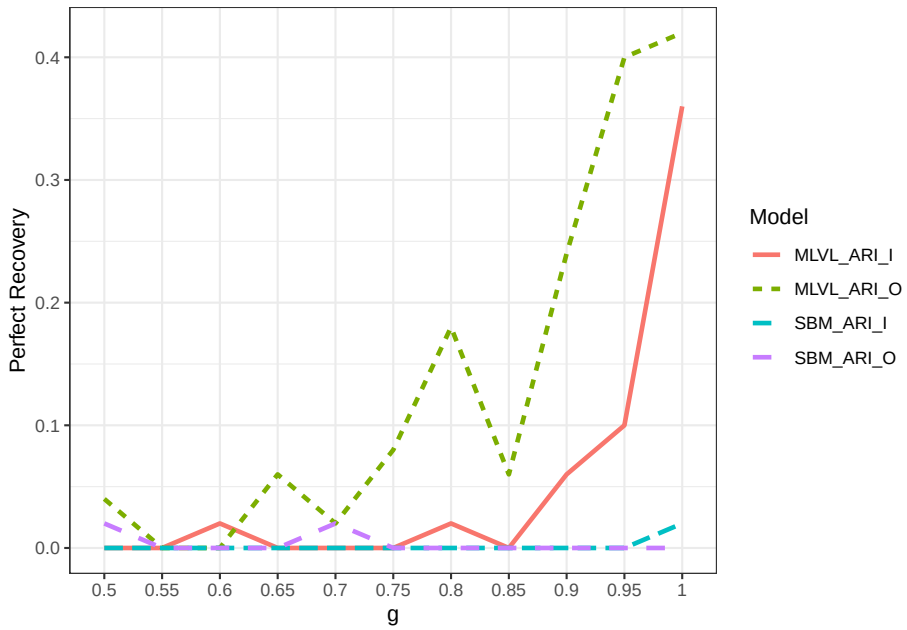
With g, the diagonal entry of the mixture dependency parameter γ, we notice that the greater the interdependence the better the levels are able to help each other in the recovery of the clustering. This is obvious from the beginning for the inter-organizational level which is slightly harder to infer. For the inter-individual we notice a great improvement when $g \geq .9$. In all case, the problem is hard so we do not obtain a perfect recovery of the blocks every time.

```
res_detect_thresh %>%
  pivot_longer(-g, names_to = "Model", values_to = "ARI") %>%
  ggplot(aes(x = g, y = ARI, group = Model, fill = Model)) +
```

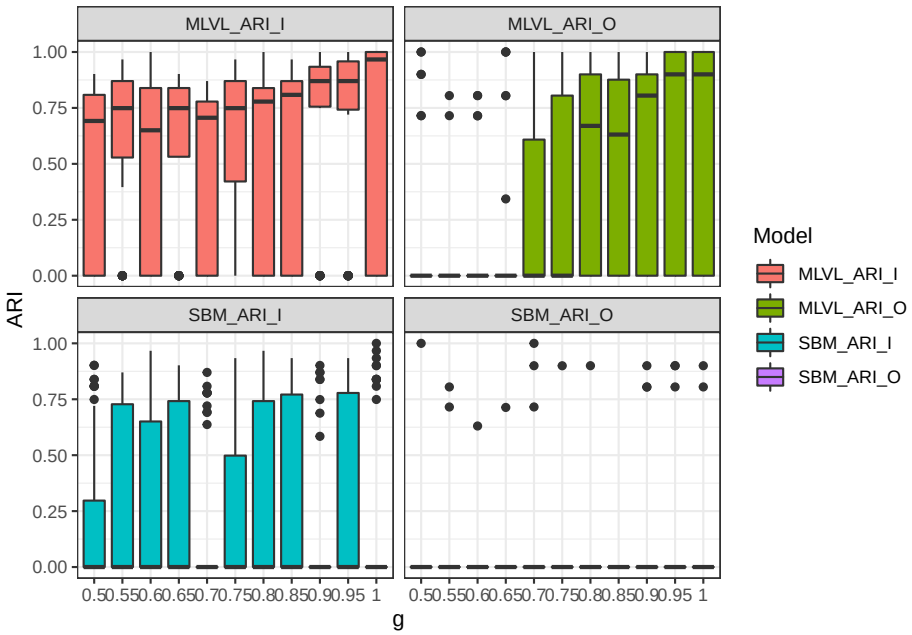
```
  stat_summary(geom = "ribbon", fun.data = "mean_se", alpha = .3) +
  stat_summary(geom = "line", fun = "mean", col = "black", size = 1.1) +
  facet_wrap(~ Model) +
  theme_bw()
```




```
res_detect_thresh %>%
  group_by(g) %>%
  summarise(across(.cols = everything(),  ~mean(.x == 1)), .groups = "keep") %>%
  pivot_longer(-g, names_to = "Model", values_to = "Perfect Recovery") %>%
  ggplot(aes(x = g, y = `Perfect Recovery`,
             group = Model, col = Model, linetype = Model)) +
  geom_line(size = 1.1) +
  theme_bw()
```



```
res_detect_thresh %>%
  pivot_longer(-g, names_to = "Model", values_to = "ARI") %>%
  ggplot(aes(x = g, y = ARI, group = g, fill = Model)) +
  geom_boxplot() +
  facet_wrap(~ Model) +
  theme_bw()
```



Chapter 3

Joint inference of a collection of networks using a stochastic block model framework

Motivation Plusieurs idées d'applications ont motivé ce travail. Dans le cadre du groupe de recherche ResoDiv (https://resodiv.cnrs.fr/, groupe interdisciplinaire étudiant des réseaux de circulation d'objets biologiques tels que des plantes ou des animaux), Étienne Polge a exposé des travaux concernant l'analyse des réseaux socio-économiques de collectifs d'agriculteurs et de leur partenaires (Polge and Torre, 2018; Pachoud et al., 2019). Nous nous sommes demandés, dans le cas où les différents réseaux socio-économiques partagent une partie de leurs structures, comment retrouver les agriculteurs ayant le même rôle structurel aux seins des différents collectifs. Parallèlement, au sein de l'ANR Econet, de nombreuses questions se posent autour de la classification et la comparaison d'écosystèmes. Nous proposons dans ce chapitre une méthode pour traiter ces différentes problématiques.

Résumé Une collection de réseaux consiste en un ensemble de réseaux qui ne partagent pas de nœuds mais qui décrivent le même type d'interactions observées dans différentes situations ou contextes. Une hypothèse est que les réseaux de la collection partagent une structure commune puisque la nature des interactions est la même. Par exemple, dans le cas d'une collection de réseaux trophiques rassemblés dans différents écosystèmes, on s'attend à ce que certains groupes d'espèces (espèces basales ou superprédateurs, par exemple) présentant des profils similaires de relations trophiques puissent être rencontrés dans tous les réseaux.

Nous proposons de nous appuyer sur le populaire modèle à blocs stochastiques (SBM) pour identifier la structure commune dans la collection. Le SBM est un modèle probabiliste qui suppose l'existence de variables latentes représentant les groupes de nœuds (blocs) du réseau et dont les paramètres fournissent une description succincte de la structure du réseau à l'échelle mésoscopique. Nous appelons *colSBM* notre extension du SBM à une modélisation conjointe d'une collection de réseaux. Les réseaux de la collection sont supposés être des réalisations indépendantes de

différents SBM, qui partagent par des paramètres communs la même structure de connectivité, éventuellement à proportions des blocs et/ou un facteur de densité près. Les paramètres du modèle sont estimés et les blocs latents sont récupérés en utilisant un algorithme EM variationnel. L'existence d'un bon compromis entre les structures méso-échelle de ces réseaux n'est pas garantie. Nous utilisons un critère ad-hoc, basé sur la vraisemblance classifiante intégrée (ICL) pour sélectionner le nombre de blocs et évaluer l'adéquation du consensus trouvé entre les structures des différents réseaux. Ce critère peut également être utilisé pour regrouper les réseaux sur la base de leurs structures de connectivité. Il fournit ainsi une partition de la collection en sous-ensembles de réseaux structurellement homogènes.

Une application à une collection de trois réseaux trophiques de cours d'eau révèle l'homogénéité de leurs structures et fournit une structure plus détaillée des plus petits réseaux. Enfin, nous montrons comment 67 réseaux trophiques peuvent être regroupés et ainsi décrits par un petit nombre de structures de connectivité.

Diffusion Le contenu de ce chapitre et des 3 premières appendices sera soumis très prochainement dans une revue de statistiques. Une application détaillée à une collection de 4 réseaux de conseils entre juges, avocats, prêtres ou chercheurs est donnée en appendice 3.D. Elle fournie quelques idées d'extensions dans les applications potentielles des méthodes $colSBM$.

Notation for this chapter

n_m	The number of nodes of network m
$\mathbf{X}$	A collection of M adjacency matrices
X^m	The adjacency matrix of network m of size $n_m \times n_m$
$\boldsymbol{Z}$	The set of latent variables
Z^m	The blocks memberships of the nodes of network m
$\boldsymbol{\theta}$	The set of model parameters
$\boldsymbol{\tau}$	The set of variational parameters
$\boldsymbol{\alpha}$	The set of connectivity parameters
$\boldsymbol{\pi}$	The set of mixture parameters
$\boldsymbol{\delta}$	The set of density parameters
π^m	The mixture parameters for network m
$\mathcal{F}$	The emission distribution
S	The support of the blocks of the $colSBM$, a $Q \times M$ binary matrix
$\mathcal{M}$	A subset of $\{1, \ldots, M\}$, the indices of a subcollection
BIC-L	The model selection criterion
$\mathcal{Q}_m$	The set of allowed blocks for network m
$\mathcal{G}$	A partition of networks
NP	The number of parameters in the model

Abstract Let a collection of networks consists of a set of networks which do not share nodes but which describe the same kind of interactions observed in different situations or contexts. An hypothesis is that the networks in the collection share a common structure since the nature of the interactions is the same. For example in the case of a collection of food webs collected in different ecosystems, one expects that some groups of species (basal species or apex predators e.g.) with similar patterns of trophic relations could be encountered in all the networks.

We propose to build on the popular stochastic block model (SBM) to identify the common structure in the collection. The SBM is a probabilistic model that assumes the existence of latent variables representing the groups of nodes (blocks) of the network and the parameters of which provide a succinct description of the structure of the network at the mesoscopic scale. We call $colSBM$ our extension of the SBM to a joint modeling of a collection of networks. The networks in the collection are assumed to be independent realizations of different SBMs, which share –through common parameters– the same connectivity structure, possibly up to the block proportions and/or a density factor.

The model parameters are estimated and the latent blocks are retrieved using a variational EM algorithm. The existence of a good compromise between the mesoscale structures of these networks is not guaranteed. We use an ad-hoc criterion, based on the integrated classification likelihood to select the number of blocks and to evaluate the adequacy of the consensus found between the structures of the different networks. This criterion can also be used to cluster networks on the basis of their connectivity structures. It thus provides a partition of the collection into subsets of structurally homogeneous networks. An application to a collection of three stream food webs reveals the homogeneity of their structures and provides a more detailed structure of the smaller networks. Finally, we demonstrate how 67 food webs can be clustered and thus described by a small number of connectivity structures.

3.1. Introduction

Context For a long time the statistical analysis of network data has focused on analyzing a single network at a time. This could be done by looking at local or global topological features, or by setting a probabilistic model inferred from the network (Kolaczyk, 2009). When several networks describing the same kind of interactions are available, a natural question is to assess to what extent they are similar or different. As network data are complex by nature, this comparison of different networks is not an easy task and has mainly focused on comparing statistical topological features on the local, global or mesoscale levels (see Donnat and Holmes, 2018, for a review of graph distances).

In this paper, we focus on the mesoscale structure of the networks by assuming that the nodes can be clustered into groups on the basis of their connectivity pattern (White et al., 1976). We assume that the considered networks have no nodes in common and that the nodes of different networks are not linked as it may be the case

in multilayer networks (Kivelä et al., 2014). Furthermore, the networks are assumed to have interactions of the same type (directed or not) and with the same valuation (binary, discrete or continuous). A set of such networks forms what we call in this paper a collection of networks, although some authors may use this terminology in a different meaning. When observing such a collection, we aim to determine if the respective structures of the networks are similar.

A classical tool to infer the mesoscale structure of a single network is the Stochastic Block Model (Holland et al., 1983; Snijders and Nowicki, 1997, SBM). In this model, a latent variable is associated with each node giving its group/block membership. Nodes belonging to the same block share the same connectivity pattern. The SBM has easily interpretable parameters and its framework allows multiple extensions such as modeling the interactions with various distributions (Mariadassou et al., 2010).

Inferring independently the structures for each network and comparing them may be misleading. Indeed, a given network may have several possible clusterings of the nodes (Peel et al., 2017) with similar probabilities (Peixoto, 2014b). Furthermore, the observation of a network may be noisy (Guimerà and Sales-Pardo, 2009), especially for ecological networks, the sampling of which is known to be incomplete (Rivera-Hutinel et al., 2012).

Our contribution Thus, we propose to jointly model a collection of networks by extending the SBM. We assume that the networks are independent realizations of SBMs that share common parameters. The natural and interesting consequence is the correspondence between the blocks of the different networks. The proposed model called *colSBM* comes with a few variants, the simplest one assumes that the parameters of the SBMs are identical leading to a collection of i.i.d. networks. As this assumption might be too restrictive for real networks, we introduce two relaxations on this assumption. The first one is to allow the distribution of the block memberships to vary between networks and even to allow some networks to not populate certain blocks. This enables to model a collection of networks where the structure of certain networks is encompassed in the structure of other networks. The second mechanism is to allow networks to have the same structure up to a density parameter. This is particularly useful to model networks with different sampling efforts as it has a direct impact on the density of ecological networks (Blüthgen et al., 2006). The inference of the block memberships, the model parameters and the model selection are done through an ad-hoc version of classic tools when inferring SBM, namely a Variational EM algorithm for the inference and an adaptation of the integrated classification likelihood (ICL) criterion for the model selection (Daudin et al., 2008).

The interest of the model is then two-folds, the first one is to find a common connectivity pattern which explains the structure of the different networks in the collection and to assess via model selection whether these structures is a reasonable fit for the collection. As a by-product, it allows a fine analysis of the role structure of the different nodes in the networks. In social/ecological networks, individuals/species

with the same block membership play the same social/ecological role in its system (White et al., 1976; Allesina and Pascual, 2009). By sharing the blocks between the networks, $colSBM$ allows to recover sets of nodes which play the same structural roles in different networks. The second one is to provide a partition of the collection of networks into groups of structurally homogeneous networks. will have some practical implications (Michalska-Smith and Allesina, 2019).

As a side effect, by modeling these networks together, provided that the networks have common connectivity patterns, we can use the information of certain networks to recover noisy information from other networks by improving the prediction of missing links (Clauset et al., 2008). Hence $colSBM$ has a stabilizing effects on the block-clustering of the nodes of the networks and might give a block membership that is closer to the one of the full real network than just a single SBM as this will be shown in the numerical studies and application.

Related work As the SBM is a very flexible model, it has been adapted to multilayer networks. To name a few, Matias and Miele (2017) model a collection of networks along a time gradient, the connectivity structure varies from time to time but they integrate a sparsity parameter, which is similar to our density parameter in the binary case.When dealing with networks with no common nodes, Chabert-Liddell et al. (2021) deal with multilevel networks where the networks are linked by a hierarchical relation between the nodes of the different levels. Within the SBM framework the closest work to ours is the strata multilayer SBM (Stanley et al., 2016), in that it looks for both common connectivity patterns and network clustering. However, it does not consider a collection of networks but a multiplex network where all the networks share the same nodes.

Most contributions about collections of networks rely on some node correspondence between the networks. Recently, motivated by the analysis of fMRI data a few works extend the SBM to model population of networks (Paul and Chen, 2018; Pavlović et al., 2020). Le et al. (2018) make the assumption that the networks of the collection are noisy realizations of the true network, while Reyes and Rodriguez (2016) use in a Bayesian framework a hierarchical SBM to model the collection. Signorelli and Wit (2020) propose a mixture of network models which is not restricted to the SBM. The contributions dealing with networks with no node correspondence include a hierarchical mixed membership SBM, using a common bayesian prior on the connectivity parameter of the different networks (Sweet et al., 2014). Finally on network clustering, Mukherjee et al. (2017) propose to fit a graphon when the networks in the collection have the same number of nodes while they use graph moments when this is not the case.

Outline Section 3.2 recalls the definition of the Stochastic Block Model. We motivate our new approach by inferring it independently on a collection of food webs. Then in Section 3.3, we present the different $colSBM$s. The likelihood expression is provided in Section 3.4, together with some identifiability conditions. We develop the methodology for the parameter estimation in Section 3.5 and for model selection

in Section 3.6. Section 3.7 deals with network clustering. After some numerical studies to demonstrate the efficiency of our inference procedure and the pertinence of our model selection criterion in Section 3.8, we deal with two applications on food webs in Section 3.9. First, we compare the structures of 3 networks and show the information transfer between these networks. Second, we seek a partition of a collection of 67 networks. The technical details are provided in appendix sections.

3.2. Data Motivation and the Stochastic Block Model

Consider a collection of M independent networks where each network indexed by m involves its own n_m nodes. The networks are encoded into their adjacency matrices $(X^m)_{m\in\{1,\dots,M\}}$ such that: $\forall m \in \{1,\dots,M\}$, $\forall (i \neq j) \in \{1,\dots,n_m\}^2$,

$$\begin{cases} X^m_{ij} & = 0 \quad \text{if no interaction is observed between species } i \text{ and } j \text{ of network } m \\ X^m_{ij} & \neq 0 \quad \text{otherwise.} \end{cases}$$

If the networks represent binary interactions then $X^m_{ij} \in \mathcal{K} = \{0,1\}$, $\forall (m,i,j)$; if the interactions are weighted such as counts, then $X^m_{ij} \in \mathcal{K} = \mathbb{N}$. Moreover, all the networks encompass the same type of interactions (binary, count...) and no self-interaction is considered. Besides, for the sake of simplicity, we assume that all the networks are directed. The extension to undirected networks i.e. such that $X^m_{ij} = X^m_{ji}$ for any $i \neq j$ is straightforward. $\mathbf{X} = (X^1,\dots,X^m)$ denotes the collection of adjacency matrices.

A first ecological example: three stream food webs As a first example, we consider the collection of three stream food webs from Thompson and Townsend (2003). The three networks collected respectively in Martins (Maine USA), Cooper and Herlzier (North-Carolina, USA) involve respectively 105, 58 and 71 species resulting in 343, 126 and 148 binary edges respectively. Classically, the food web edges represent directed trophic links showing the energy flow ie. $X^m_{ij} = 1$ if species j preys on species i, with no reciprocal interactions. When aiming at unraveling the structure of these networks the Stochastic Block Model (SBM) is an interesting tool which has proven its high flexibility by encompassing a large variety of structures (see Allesina and Pascual, 2009, for the particular case of food webs). When dealing with three networks, the standard strategy that we describe below, is to fit separately one SBM per network.

Separate SBM (sepSBM) The SBM introduces clusters of nodes and assumes that the interaction between two nodes is driven by the clusters the nodes belong to. More precisely, for network m, let the n_m nodes be divided into Q_m clusters. Let

$Z^m = (Z_i^m, \ldots, Z_{n_m}^m)$ be independent latent random variables such that $Z_i^m = q$ if node i of network m belongs to cluster q with $q \in \{1, \ldots, Q_m\}$ and

$$P(Z_i^m = q) = \pi_q^m \tag{3.1}$$

where $\pi_q^m > 0$ and $\sum_{q=1}^{Q_m} \pi_q^m = 1$. Given the latent variables Z^m, the X_{ij}^m's are assumed to be independent and distributed as

$$X_{ij}^m | Z_i^m = q, Z_j^m = r \sim \mathcal{F}(\cdot; \alpha_{qr}^m), \tag{3.2}$$

where $\mathcal{F}$ is referred to as the emission distribution. $\mathcal{F}$ is the Bernoulli distribution for binary interactions, and the Poisson distribution for weighted interactions such as counts. Let f be the density of the emission distribution, then:

$$\log f(X_{ij}^m; \alpha_{qr}^m) = \begin{cases} X_{ij}^m \log\left(\alpha_{qr}^m\right) + (1 - X_{ij}^m) \log\left(1 - \alpha_{qr}^m\right) & \text{for Bernoulli emission} \\ -\alpha_{qr}^m + X_{ij}^m \log\left(\alpha_{qr}^m\right) - \log(X_{ij}!) & \text{for Poisson emission} \end{cases}. \tag{3.3}$$

Equations (3.1), (3.2) and (3.3) define the SBM model and we will now use the following short notation:

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q_m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m). \tag{sep-SBM}$$

where $\mathcal{F}$ encodes the emission distribution, n_m is the number of nodes, Q_m is the number of blocks and $\boldsymbol{\pi}^m = (\pi_q^m)_{q=1,\ldots,Q_m}$ is the vector of their proportions. The $Q \times Q$ matrix $\boldsymbol{\alpha}^m = \left(\alpha_{qr}^m\right)_{q,r=1,\ldots,Q_m}$ denotes the connection parameters i.e. the parameters of the emission distribution. Moreover, $\alpha_{qr}^m \in \mathcal{A}_{\mathcal{F}}$ where $\mathcal{A}_{\mathcal{F}} = (0, 1)$ (resp. $\mathcal{A}_{\mathcal{F}} = \mathbb{R}^{*+}$) for the Bernoulli (resp. Poisson) emission distribution. In the *sep-SBM* model, each network m is modeled independently with its own parameters $(\boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$.

Application to the three stream food webs We fit the *sepSBM* on the 3 stream food webs, respectively referred to as Martins, Cooper and Herlzier. To do so, we use the `sbm` R-package (Chiquet et al., 2021; Leger et al., 2020) on each network, which implements a variational version of the EM algorithm to estimate the parameters and selects the number of clusters Q_m using a penalized likelihood criterion ICL. These inference tools will be recalled hereafter.

We obtain respectively $\widehat{Q}_1 = 5$ blocks for Martins, $\widehat{Q}_2 = 3$ blocks for Cooper and $\widehat{Q}_3 = 4$ blocks for Herlzier. The adjacency matrices of the food webs reordered by block membership are plotted in Figure 3.1. Each food web is composed of 2 blocks of basal species (the 2 bottom blocks). For Cooper, the higher trophic levels are grouped together in the same block, as the lack of statistical power does not allow refinement of the species clustering. For Herlzier the higher trophic level is separated into 2 blocks mainly determined on how much they prey on the less preyed basal block. Martins has a separation into 3 blocks, the third one is a medium trophic level, which preys on basal species and is highly preyed on by species of the first

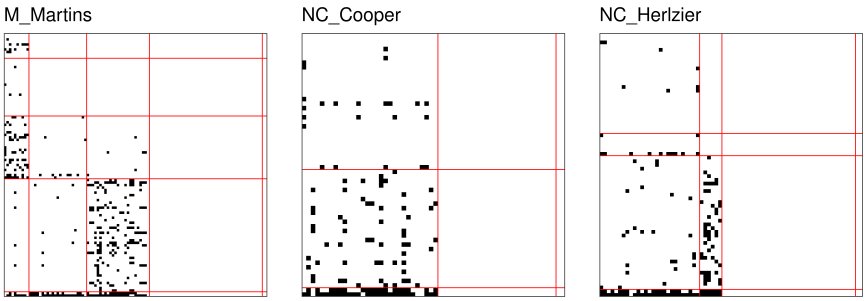


Figure 3.1 – Matricial view of 3 stream food webs. The species are reordered by blocks and blocks are ordered by expected out-degrees to emulate the trophic levels (bottom to top and right to left). The blocks have been obtained by fitting a SBM on each network separately.

block. The first two blocks are made up of higher trophic level species, with the last two blocks being much less connected than the first.

As can be seen in Figure 3.1, the connectivity structures of these three networks seem to have a lot of similarities. To explore further this aspect, the following section is dedicated to the presentation of several *colSBM* models assuming common structures among the networks of a given collection.

3.3. Joint Modeling of a Collection of Networks

We now present a set of probabilistic models designed to introduce structure consensus into a collection of networks of interest. For ease of notation, we develop the models for directed networks; extensions to the undirected cases are straightforward. Note that the networks $(X^m)_{m=1,\dots,M}$ are always assumed to be independent random objects. A summary of the various models is provided in Table 3.1, from the most to the less constrained model

3.3.1. A collection of i.i.d. SBM

The first model we propose is the most constrained one and assumes that the networks are independent realizations of the same Q-blocks SBM model with identical parameters. The so-called *iid-colSBM* states that:

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q, \boldsymbol{\pi}, \boldsymbol{\alpha}), \qquad \forall m = 1, \dots M, \qquad (\textit{iid-colSBM})$$

where $\forall (q,r) \in \{1,\dots,Q\}^2$, $\alpha_{qr} \in \mathcal{A}_{\mathcal{F}}$, $\pi_q \in]0,1]$ and $\sum_{q=1}^{Q} \pi_q = 1$. The model involves $(Q-1)+Q^2$ parameters, the first term corresponding to the block proportions and the second term to the connection parameters.

However, assuming that the blocks are represented in the same proportions in each network is a strong assumption that may lead to the model being of little practical use. The following model relaxes this assumption.

3.3.2. A collection of networks with varying block sizes

π-*colSBM* assumes that the networks share a common connectivity structure encoded in $\boldsymbol{\alpha}$, but that the proportions of the blocks are specific to each network. Moreover, by allowing some block proportions to be null, the model encompasses situations where some blocks may not be represented in all the realized networks. More precisely, for $m \in \{1, \ldots, M\}$, the X^m are independent and

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \boldsymbol{\alpha}) \qquad \forall m = 1, \ldots M. \qquad (\pi\text{-}colSBM)$$

where $\alpha_{qr} \in \mathcal{A}_{\mathcal{F}}$. Moreover, we assume that $\sum_{q=1}^{Q} \pi_q^m = 1, \forall m \in \{1, \ldots, M\}$ where $\pi_q^m \in [0, 1]$ and if $\pi_q^m = 0$ then the block q is not represented in network m. In addition, we assume that for any $q \in \{1, \ldots, Q\}$, $\exists m \in \{1, \ldots, M\}$ such that $\pi_q^m > 0$, meaning that any block q is represented in at least one network. Let S be the $M \times Q$ support matrix such that $\forall (m, q)$

$$S_{mq} = \mathbf{1}_{\pi_q^m > 0}. \tag{3.4}$$

Then, the set of admissible supports is

$$\mathcal{S}_Q := \left\{ S \in \mathcal{M}_{M,Q}(\{0, 1\}), \sum_{m=1}^{M} S_{mq} \geq 1 \quad \forall q = 1, \ldots, Q \right\} \tag{3.5}$$

For a given matrix S, the number of parameters of the π-*colSBM* model is deduced as follows:

$$\text{NP}(\pi\text{-}colSBM) \quad = \quad \sum_{m=1}^{M} \left(\sum_{q=1}^{Q} S_{qm} - 1 \right) + \sum_{q,r=1}^{Q} \mathbf{1}_{(S'S)_{qr} > 0} \tag{3.6}$$

The first term corresponds to the non-null block proportions in each network. The second quantity accounts for the fact that some blocks may never be represented simultaneously in any network, so the corresponding connection parameters are not useful for defining the model (see the illustration below). The number of parameters is bounded by $M(Q-1) + Q^2$, this upper bound corresponding to the case where all the blocks are represented in all the networks, but with varying proportions.

Illustration We illustrate the flexibility of this model with three examples, all with $Q = 3$ and $M = 2$.

1. First consider the situation where the 3 blocks are represented in the two networks but with different block proportions:

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \qquad \begin{matrix} \pi^1 = [.25, .25, .50] \\ \pi^2 = [.20, .50, .30] \end{matrix}.$$

In that case, $S = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ and the number of parameters is $2(3-1)+3\times 3 = 13$.

2. Now imagine two networks with imbricated structures. Blocks 1 and 3 are represented in the two networks while block 2 only exists in network 1.

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{pmatrix} \qquad \begin{array}{l} \pi^1 = [.25, .25, .50] \\ \pi^2 = [.40,\ 0\ , .60] \end{array}.$$

In that case, $S = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ and the number of parameters is $(3-1)+(2-1)+3\times 3 = 12$.

3. Finally, let us consider two networks with partially imbricated structures. The two networks share block 1 (for instance super predators) but the remaining nodes of each network cannot be considered as equivalent in terms of connectivity:

$$\alpha = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \cdot \\ \alpha_{31} & \cdot & \alpha_{33} \end{pmatrix} \qquad \begin{array}{l} \pi^1 = [.25, .75,\ 0\ ] \\ \pi^2 = [.40,\ 0\ , .60] \end{array}.$$

In that case, $S = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$. Moreover, blocks 2 and 3 never interact because their elements do not belong to the same network and so α_{23} and α_{32} are not required to define the model. As a consequence, the number of parameters is equal to $(2-1)+(2-1)+7 = 9$.

3.3.3. A collection of networks with varying density (δ-*colSBM*)

The *iid-colSBM* can be relaxed in another direction, assuming that the M networks exhibit similar intra- and inter- blocks connectivity patterns but with different densities. More precisely, let $\delta_m \in \mathbb{R}$ be a density parameter for network m. The δ-*colSBM* is defined as follows:

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q, \boldsymbol{\pi}, \delta_m \boldsymbol{\alpha}). \qquad (\delta\text{-}colSBM)$$

with $\pi_q > 0, \forall q = 1, \ldots, Q$, $\sum_{q=1}^{Q} \pi_q = 1$. Moreover $\forall (m, q, r)$, $\delta_m \alpha_{qr} \in \mathcal{A}_{\mathcal{F}}$ and one of the density parameter equal to one ($\delta_1 = 1$) for identifiability purpose. This model mimics different intensity of connections between networks. δ-*colSBM* involves $\text{NP}(\delta\text{-}colSBM) = (Q-1) + Q^2 + (M-1)$ parameters.

Model name	Block prop.	Connection param.	Nb of param.
iid-colSBM	$\pi_q^m = \pi_q$, $\pi_q > 0$	$\alpha_{qr}^m = \alpha_{qr}$	$(Q-1)+Q^2$
π-*colSBM*	π_q^m, $\pi_q^m \geq 0$	$\alpha_{qr}^m = \alpha_{qr}$	$\leq M(Q-1)+Q^2$
δ-*colSBM*	$\pi_q^m = \pi_q$, $\pi_q > 0$	$\alpha_{qr}^m = \delta_m \alpha_{qr}$	$(Q-1)+Q^2+(M-1)$
$\delta\pi$-*colSBM*	π_q^m, $\pi_q^m \geq 0$	$\alpha_{qr}^m = \delta_m \alpha_{qr}$	$\leq MQ+Q^2-1$
sepSBM	π_q^m, $\pi_q^m > 0$	α_{qr}^m	$\sum_{m=1}^M (Q_m-1)+Q_m^2$

Table 3.1 – Summary of the various models defined in Section 3.3. The last line corresponds to modeling separately each network as presented in Section 3.2.

3.3.4. Collection of networks with varying block sizes and density ($\delta\pi$-*colSBM*)

Finally, we propose to mix the models π-*colSBM* and δ-*colSBM* to obtain a more complex one which allows each network to have its own block proportions $\boldsymbol{\pi}^m$ as well as a specific scale density parameter δ_m. Then, the $(X^m)_{m\in\{1,\dots,M\}}$ are independent and

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \delta_m \boldsymbol{\alpha}), \qquad (\delta\pi\text{-}colSBM)$$

where $\forall (m,q,r)$, $\delta_m \alpha_{qr} \in \mathcal{A}_{\mathcal{F}}$, $\delta_1 = 1$, $\pi_q^m \geq 0$ and $\sum_{q=1}^Q \pi_q^m = 1$. The number of parameters is given by

$$\text{NP}(\delta\pi\text{-}colSBM) = \text{NP}(\pi\text{-}colSBM) + M - 1, \tag{3.7}$$

the last term corresponding to the aditional proper density of each network. Note that $\text{NP}(\delta\pi\text{-}colSBM) \leq M(Q-1)+Q^2+M-1 = MQ+Q^2-1$.

3.4. Likelihood and identifiability of the models

In this section, we derive the expression of the likelihood of the most complex model $\delta\pi$-*colSBM* and provide conditions to ensure the identifiability of the parameters for each of the four models.

3.4.1. Log-likelihood expression

For a given matrix S, let $\boldsymbol{\theta}_S$ be:

$$\boldsymbol{\theta}_S = (\boldsymbol{\pi}^1, \dots \boldsymbol{\pi}^M, \delta_1, \dots, \delta_M, \boldsymbol{\alpha}) = (\boldsymbol{\pi}, \boldsymbol{\delta}, \boldsymbol{\alpha}),$$

where $\pi_q^m = 0$ for any q such that $S_{mq} = 0$. Let $Z_{iq}^m = \mathbf{1}_{Z_i^m = q}$ be the latent variable such that $Z_{iq}^m = 1$ if node i of network m belongs to block q, $Z_{iq}^m = 0$ otherwise. We define $Z^m = (Z_{iq}^m)_{i=1\dots,n_m, q=1\dots,Q}$. Then the log likelihood is:

$$\ell(\mathbf{X}; \boldsymbol{\theta}_S) = \sum_{m=1}^M \log \int_{Z^m} \exp\left\{\ell(X^m | Z^m; \boldsymbol{\alpha}, \boldsymbol{\delta}) + \ell(Z^m; \boldsymbol{\pi})\right\} dZ^m, \tag{3.8}$$

where

$$\begin{array}{rcl} \ell(X^m|Z^m;\boldsymbol{\alpha},\boldsymbol{\delta}) & = & \sum_{\substack{i,j=1\\ i\neq j}}^{n_m}\sum_{(q,r)\in\mathcal{Q}_m} Z^m_{iq}Z^m_{jr}\log f\left(X^m_{ij};\delta_m\alpha_{qr}\right), \\ \ell(Z^m;\boldsymbol{\pi}) & = & \sum_{i=1}^{n_m}\sum_{q\in\mathcal{Q}_m} Z^m_{iq}\log\pi^m_q \end{array}.$$

with $\mathcal{Q}_m = \{q \in \{1,\dots,Q\}|\pi^m_q > 0\}$ and f defined as in Equation 3.3. The log-likelihood functions of the other models can be deduced from this one, setting $\delta_m = 1$ for *iid-colSBM* and *π-colSBM* and $\boldsymbol{\pi}^m = \boldsymbol{\pi}$ for *iid-colSBM* and *δ-colSBM* with S being a matrix of ones (all blocks are represented in each network).

Identifiability

We aim at giving conditions ensuring the identifiability of the models we propose, in the sense that $\forall\mathbf{X},\ \ell(\mathbf{X};\boldsymbol{\theta}) = \ell(\mathbf{X};\boldsymbol{\theta}')$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}'$. The proof relies on the identifiability for the standard -SBM model demonstrated by Celisse et al. (2012). Note that, like any mixture models, all the models are identifiable upto a label switching of the blocks.

Properties 3.1.

iid-colSBM *The parameters $(\boldsymbol{\pi},\boldsymbol{\alpha})$ are identifiable upto a label switching of the blocks provided*

1. $\exists m^* \in \{1,\dots,M\} : n_{m^*} \geq 2Q$
2. $(\boldsymbol{\alpha}\cdot\boldsymbol{\pi})_q \neq (\boldsymbol{\alpha}\cdot\boldsymbol{\pi})_r\ \forall(q,r)\in\{1,\dots,Q\}^2, q\neq r$

δ-colSBM *The parameters $(\boldsymbol{\pi},\delta_1,\dots,\delta_M,\boldsymbol{\alpha})$ are identifiable upto a label switching of the blocks provided*

1. $\exists m^* \in \{1,\dots,M\} : n_{m^*} \geq 2Q$
2. $\delta_{m^*} = 1$
3. $(\alpha\cdot\pi)_q \neq (\alpha\cdot\pi)_r\ \forall(q,r)\in\{1,\dots,Q\}^2, q\neq r$

π-colSBM *Assume that $\forall m = 1,\dots,M, X^m \sim \mathcal{F}\text{-}SBM_{n_m}(Q,\boldsymbol{\pi}^m,\boldsymbol{\alpha})$. Let $Q_m = |\mathcal{Q}_m| = |\{q = 1\dots,Q,\pi^m_q > 0\}|$ be the number of non empty blocks in network m. Then the parameters $(\boldsymbol{\pi}^1,\dots,\boldsymbol{\pi}^M,\boldsymbol{\alpha})$ are identifiable upto a label switching of the blocks under the following conditions:*

1. $\forall m \in \{1,\dots,M\} : n_m \geq 2Q_m$
2. $(\alpha\cdot\pi^m)_q \neq (\alpha\cdot\pi^m)_r$ *for all* $(q\neq r)\in\mathcal{Q}^2_m$
3. *Each diagonal entry of $\boldsymbol{\alpha}$ is unique*

$\delta\pi$-colSBM *Assume that $\forall m = 1,\dots,M, X^m \sim \mathcal{F}\text{-}SBM_{n_m}(Q,\boldsymbol{\pi}^m,\delta_m\boldsymbol{\alpha})$. Let $Q_m = |\mathcal{Q}_m| = |\{q = 1\dots,Q,\pi^m_q > 0\}|$ be the number of non empty blocks in network m. Then the parameters $(\boldsymbol{\pi}^1,\dots,\boldsymbol{\pi}^M,\boldsymbol{\alpha},\delta_1,\dots,\delta_M)$ are identifiable upto a label switching of the blocks under the following conditions:*

1. $\forall m : n_m \geq 2|\mathcal{Q}_m|$
2. $\delta_1 = 1$

   For $Q \geq 2$:
3. $(\boldsymbol{\alpha} \cdot \pi^m)_q \neq (\boldsymbol{\alpha} \cdot \pi^m)_r$ *for all* $(q \neq r) \in \mathcal{Q}_m^2$
4. $\forall m \in \{1, \ldots, M\}, Q_m \geq 2$
5. *Each diagonal entry of* $\boldsymbol{\alpha}$ *is unique*

   For $Q \geq 3$:
6. *There is no configuration of four indices* $(q, r, s, t) \in \{1, \ldots, Q\}$ *such that* $q \neq s$, $r \neq t$ *and* $\alpha_{qq}/\alpha_{rr} = \alpha_{ss}/\alpha_{tt}$.
7. $\forall m \geq 2, |\mathcal{Q}_m \cap \cup_{l:l<m} \mathcal{Q}_l| \geq 2$.

3.5. Variational estimation of the parameters

We now tackle the estimation of the parameters $\boldsymbol{\theta}_S \in \Theta_S$ for a given support matrix S. For ease of reading, the index S is dropped in this section. The likelihood given in Equation (3.8) is not tractable in practice, even for a small collection of networks as it relies on summing over $\sum_{m=1}^{M} |\mathcal{Q}_m|^{n_m}$ terms. A well-proven approach to handle this problem for the inference of the SBM is to rely on a variational version of the EM algorithm. This is done by maximizing a lower (variational) bound of the log-likelihood of the observed data (Daudin et al., 2008). The approach is similar for both Bernoulli and Poisson models. More precisely,

$$\begin{aligned} \ell(\mathbf{X}; \boldsymbol{\theta}) &= \sum_{m=1}^{M} \ell(X^m; \boldsymbol{\theta}) \\ &\geq \sum_{m=1}^{M} \Big(\ell(X^m; \boldsymbol{\theta}) - D_{\mathrm{KL}}(\mathcal{R}_m(Z^m) \| p(Z^m | X^m; \boldsymbol{\theta})) \Big) \end{aligned}$$

where D_{KL} is the Kullback-Leibler divergence and $\mathcal{R}_m$ stands for any distribution on Z^m and $\mathcal{R}$ denotes the product distribution: $\mathcal{R} = \otimes_{m=1}^{M} \mathcal{R}_m$. The last equation can be reformulated as:

$$\ell(\mathbf{X}; \boldsymbol{\theta}) \geq \sum_{m=1}^{M} \Big(\mathbb{E}_{\mathcal{R}_m}[\ell(X^m, Z^m; \boldsymbol{\theta})] + \mathcal{H}(\mathcal{R}_m(Z^m)) \Big) =: \mathcal{J}(\mathcal{R}, \boldsymbol{\theta}). \tag{3.9}$$

where $\mathcal{H}$ denotes the entropy of a distribution. Now, if $\mathcal{R}_m$ for all $m \in \{1, \ldots, M\}$ is chosen in the set of fully factorizable distributions and if one sets $\tau_{iq}^m = \mathbb{P}_{\mathcal{R}_m}(Z_{iq}^m = 1)$ then $\mathcal{H}(\mathcal{R}_m(Z^m))$ is equal to:

$$\mathcal{H}(\mathcal{R}_m(Z^m)) = -\sum_{i=1}^{n_m} \sum_{q \in \mathcal{Q}_m} \tau_{iq}^m \log \tau_{iq}^m. \tag{3.10}$$

Besides, the complete likelihood of network m for the $\delta\pi$-$colSBM$ marginalized over $\mathcal{R}_m$ is given by:

$$\mathbb{E}_{\mathcal{R}}[\ell(X^m, Z^m; \boldsymbol{\theta})] = \sum_{\substack{i,j=1\\ i\neq j}}^{n_m} \sum_{(q,r)\in\mathcal{Q}_m} \tau_{iq}^m \tau_{jr}^m \log f(X_{ij}^m; \delta_m \alpha_{qr}) + \sum_{i=1}^{n_m} \sum_{q\in\mathcal{Q}_m} \tau_{iq}^m \log \pi_q^m. \tag{3.11}$$

Finally, the variational lower bound $\mathcal{J}(\mathcal{R}, \boldsymbol{\theta}) := \mathcal{J}(\boldsymbol{\tau}, \boldsymbol{\theta})$ is obtained by plugging Equations (3.10) and (3.11) into the right member of Equation (3.9). Note that the lower bound $\mathcal{J}(\boldsymbol{\tau}, \boldsymbol{\theta})$ is equal to the log-likelihood if $\mathcal{R}_m(Z^m) = p(Z^m|X^m)$ for all $m \in \{1, \ldots, M\}$.

The variational EM (VEM) algorithm consists in optimizing the lower bound $\mathcal{J}(\boldsymbol{\tau}, \boldsymbol{\theta})$ with respect to $(\boldsymbol{\tau}, \boldsymbol{\theta})$, by iterating two optimization steps with respect to $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ respectively, also referred to as VE-step and M-step. The details of each step are specific to the model at stake and are detailed hereafter.

VE-step At iteration (t) of the VEM algorithm, the VE-step consists in maximizing the lower bound with respect to $\boldsymbol{\tau}$:

$$\widehat{\boldsymbol{\tau}}^{(t+1)} = \arg\max_{\boldsymbol{\tau}} \mathcal{J}(\boldsymbol{\tau}, \widehat{\boldsymbol{\theta}}^{(t)}).$$

Note that by doing so, one minimizes the Kullback-Leibler divergences between $\mathcal{R}_m(Z^m)$ and $p(Z^m|X^m)$, and so approximates the true conditional distribution $p(Z^m|X^m)$ in the space of fully factorizable probability distributions. The $\boldsymbol{\tau}^m$'s can be optimized separately by iterating the following fixed point systems for all $m \in \{1, \ldots, M\}$:

$$\widehat{\tau}_{iq}^{m(t+1)} \propto \widehat{\pi}_q^{m(t)} \prod_{\substack{j=1\\ j\neq i}}^{n_m} \prod_{r\in\mathcal{Q}_m} f(X_{ij}^m; \widehat{\delta}_m^{(t)} \widehat{\alpha}_{qr}^{(t)})^{\widehat{\tau}_{jr}^{m(t+1)}} \quad \forall i = 1, \ldots, n_m, q \in \mathcal{Q}_m. \tag{3.12}$$

M-Step At iteration (t) of the VEM algorithm, the M-step maximizes the variational bound with respect to the model parameters $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{J}(\widehat{\boldsymbol{\tau}}^{(t+1)}, \boldsymbol{\theta}).$$

The update depends on the chosen model and the estimations are derived by canceling the gradient of the lower bound. For the sake of simplicity, the iteration index (t) is dropped in the following formulae. The obtained formulae involve the following quantities:

$$e_{qr}^m = \sum_{\substack{i,j=1\\ i\neq j}}^{n_m} \tau_{iq}^m \tau_{jr}^m X_{ij}^m, \qquad n_{qr}^m = \sum_{\substack{i,j=1\\ i\neq j}}^{n_m} \tau_{iq}^m \tau_{jr}^m, \quad n_q^m = \sum_{i=1}^{n_m} \tau_{iq}^m.$$

On the one hand, the $(\pi_q^{(m)})_{q\in\mathcal{Q}_m}$ are estimated as

$$\widehat{\pi}_q^m = \frac{n_q^m}{n_m} \qquad \text{for } \pi\text{-}colSBM \text{ and } \delta\pi\text{-}colSBM,$$

which is the expected proportion of the nodes in each allowed block for network m. On the other hand,

$$\widehat{\pi}_q = \frac{\sum_{m=1}^{M} n_q^m}{\sum_{m=1}^{M} n_m} \qquad \text{for } iid\text{-}colSBM \text{ and } \delta\text{-}colSBM,$$

taking into account all the networks at the same time. The connection parameters α_{qr} of *iid-colSBM* and π-*colSBM* are estimated as the ratio of the number of interactions between blocks q and r among all networks over the number of possible interactions:

$$\widehat{\alpha}_{qr} = \frac{\sum_{m=1}^{M} e_{qr}^m}{\sum_{m=1}^{M} n_{qr}^m} \quad \text{for } iid\text{-}colSBM \text{ and } \pi\text{-}colSBM\ .$$

For the δ-*colSBM* and $\delta\pi$-*colSBM*, there is no closed form for $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\alpha}}$ for a given value of $\boldsymbol{\tau}$. If $\mathcal{F} = \mathcal{P}oisson$, then $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\alpha}}$ can be iteratively updated using the following formulae:

$$\hat{\alpha}_{qr} = \frac{\sum_{m=1}^{M} e_{qr}^m}{\sum_{m=1}^{M} n_{qr}^m \hat{\delta}^m} \qquad \text{and} \qquad \hat{\delta}^m = \frac{\sum_{q,r\in\mathcal{Q}_m} e_{qr}^m}{\sum_{q,r\in\mathcal{Q}_m} n_{qr}^m \hat{\alpha}_{qr}}$$

If $\mathcal{F} = \mathcal{B}ernoulli$, no explicit expression can be derived and one has to rely on a gradient ascent algorithm to update the parameters at each M-Step.

Remark. *In the VE-Step, each network can be treated independently, so the computation can be parallelized with ease. Also, it can be more efficient to update only a subset of networks at each step to avoid being stuck in local maxima. So we use a slightly modified VEM algorithm where we just compute the VE-step on one network at a time (the order of which is taken uniformly at random) before updating the corresponding parameters in the M-Step.*

3.6. Model selection

There are two model selection issues. First, under a fixed *colSBM*, we aim to choose Q and determine the support matrix S for π-*colSBM* and $\delta\pi$-*colSBM*. This task is tackled in Subsection 3.6.1 by introducing a penalized likelihood criterion. Second, the comparison of the the *colSBM* models –each one introducing various degrees of consensus between the networks– with the *sep-SBM* – which assumes that each network has its own structure– is dealt with in Subsection 3.6.2.

3.6.1. Selecting the number of blocks Q

A classical tool to choose the number of blocks in the SBM context is the Integrated Classified Likelihood (ICL) proposed by Biernacki et al. (2000); Daudin et al. (2008). ICL derives from an asymptotic approximation of the marginal complete likelihood $m(\mathbf{X}, \mathbf{Z}) = \int_{\boldsymbol{\theta}} \exp\{\ell(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\} p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ where the parameters are integrated

out against a prior distribution, resulting in a penalized criterion of the form $\max_{\boldsymbol{\theta}} \ell(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}) - \frac{1}{2}\text{pen}$. In the ICL, the latent variables $\mathbf{Z}$ are integrated out against an approximation of $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ obtained via the variational approximation. This leads to the following expression

$$\text{ICL} = \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{R}_{\widehat{\boldsymbol{\tau}}}} \left[\ell(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})\right] - \frac{1}{2}\text{pen}$$

Using the fact that $\mathbb{E}_{\mathcal{R}_{\boldsymbol{\tau}}} \left[\ell(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})\right] \approx \ell(\mathbf{X}; \boldsymbol{\theta}) - \mathcal{H}(\mathcal{R}_{\boldsymbol{\tau}})$, one understands that, as emphasized in the literature, ICL favors well separated blocks by penalizing for the entropy of the clustering. However, in this work, our goal is not only to cluster the nodes into coherent blocks but also to evaluate the similarity of the connectivity patterns between the different networks. As such we would like to authorize models providing clustering that may be more fuzzy by not penalizing for the entropy. This leads to a BIC-like criterion of the form:

$$\text{BIC-L} \quad = \quad \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{R}_{\widehat{\boldsymbol{\tau}}}} \left[\ell(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta})\right] + \mathcal{H}(\mathcal{R}_{\widehat{\boldsymbol{\tau}}}) - \frac{1}{2}\text{pen} = \max_{\boldsymbol{\theta}} \mathcal{J}(\widehat{\boldsymbol{\tau}}, \boldsymbol{\theta}) - \frac{1}{2}\text{pen}$$

We now supply the expression of the penalty term for the four models we proposed and discuss possible variations of the criterion.

Selection of Q for *iid-colSBM* and *δ-colSBM*

For *iid-colSBM* and *δ-colSBM*, the derivation of the penalty is a straightforward extension of the classical SBM model, leading to:

$$\text{BIC-L}(\mathbf{X}, Q) = \max_{\boldsymbol{\theta}} \mathcal{J}(\widehat{\boldsymbol{\tau}}, \boldsymbol{\theta}) - \frac{1}{2}\left[\text{pen}_{\pi}(Q) + \text{pen}_{\alpha}(Q) + \text{pen}_{\delta}(Q)\right], \tag{3.13}$$

where

$$\begin{aligned}
\text{pen}_{\pi}(Q) &= (Q-1)\log\left(\sum_{m=1}^{M} n_m\right), \\
\text{pen}_{\alpha}(Q) &= Q^2 \log(N_M), \\
\text{pen}_{\delta}(Q) &= \begin{cases} 0 & \text{for } \textit{iid-colSBM} \\ (M-1)\log(N_M) & \text{for } \delta\textit{-colSBM} \end{cases}.
\end{aligned}$$

where $N_M = \sum_{m=1}^{M} n_m(n_m - 1)$ is the number of possible interactions. The first term $\text{pen}_{\pi}(Q)$ corresponds to the clustering part where the $Q-1$ block proportions have to be estimated from the $\sum_{m=1}^{M} n_m$ nodes. The terms $\text{pen}_{\alpha}(Q)$ and $\text{pen}_{\delta}(Q)$ are linked to the connection parameters. Finally, Q is chosen as:

$$\widehat{Q} = \text{argmax}_{Q \in \{1,\ldots,Q_{\max}\}} \text{BIC-L}(\mathbf{X}, Q).$$

Selection of Q for *π-colSBM* and *δπ-colSBM*

Here, in addition to the choice of Q, the collection of support matrices S is considered. In order to penalize the complexity of the model space, we introduce a prior

distribution on S defined as follows. Let us introduce $Q_m = \sum_{q=1}^{Q} S_{mq}$ the number of blocks represented in network m. Assuming independent uniform prior distributions on the (Q_m)'s and a uniform prior distribution on S for fixed numbers of blocks $Q_1, \ldots, Q_M$ represented in each network, we obtain the following prior distribution on S:

$$\log p_Q(S) \quad = \quad -M \log(Q) - \sum_{m=1}^{M} \log \binom{Q}{Q_m}$$

where $\binom{Q}{Q_m}$ is the number of choices of Q_m non-empty blocks among the Q possible blocks in network m. Now, combining the Laplace asymptotic approximation of the marginal complete likelihood (where the parameters have been integrated out) and introducing the prior distribution on S, we obtain the following penalized criterion:

$$\text{BIC-L}(\mathbf{X}, Q) = \max_{S} \left[\max_{\boldsymbol{\theta}_S \in \Theta_S} \mathcal{J}(\hat{\boldsymbol{\tau}}, \boldsymbol{\theta}_S) - \frac{1}{2} \left[\text{pen}_\pi(Q,S) + \text{pen}_\alpha(Q,S) + \text{pen}_\delta(Q,S) + \text{pen}_S(Q) \right] \right], \tag{3.14}$$

where

$$\begin{aligned}
\text{pen}_\pi(Q,S) &= \sum_{m=1}^{M} (Q_m - 1) \log(n_m), \\
\text{pen}_\alpha(Q,S) &= \left(\sum_{q,r=1}^{Q} \mathbf{1}_{(S'S)_{qr}>0} \right) \log\left(N_M\right), \\
\text{pen}_\delta(Q,S) &= \begin{cases} 0 & \text{for } \pi\text{-}colSBM \\ (M-1)\log\left(N_M\right) & \text{for } \delta\pi\text{-}colSBM \end{cases}, \\
\text{pen}_S(Q) &= -2 \log p_Q(S).
\end{aligned}$$

Finally, Q is chosen such that:

$$\widehat{Q} = \text{argmax}_{Q \in \{1,\ldots,Q_{\max}\}} \text{BIC-L}(\mathbf{X}, Q).$$

The details about the derivation of this criterion are provided in Appendix 3.B.

Practical model selection

The practical choice of Q and the estimation of its parameters are computationally intensive tasks. Indeed, we should compare all the possible models through the chosen model selection criterion. Furthermore, for each model, the variational EM algorithm should be initialized at a large number of initialization points (due to its sensitivity to the starting point), resulting in an unreasonable computational cost. Instead, we propose to adopt a stepwise strategy, resulting in a faster exploration of the model space, combined with efficient initializations of the variational EM algorithm. The procedure we suggest is given in Algorithm 4 and is implemented in an R-package colSBM available on GitHub and that will be on CRAN soon. For initializing a *colSBM* with Q blocks, we start from fitting a *sep-SBM*, then, the Q blocks of the M networks have to be associated. This association can be done in

many ways due to label switching within each network which provides us with a lot of possible initializations. Then, the stepwise procedure explores the possible number of blocks by building on the previously fitted models. Note that when fitting the π-*colSBM* or the $\delta\pi$-*colSBM*, the support S has to be determined which is done through an extra-step that consists in thresholding the parameters $\boldsymbol{\pi}^m$ related with the block proportions leading to an exploration over the set $\mathcal{S}_Q$.

Algorithm 4: Model selection algorithm

Data: $\mathbf{X}$ a collection of networks.

begin initialization
- -Infer *sep-SBM* on $\mathbf{X}$, with $Q \in [Q_{min}, Q_{max}]$
- -Get $\hat{Z}^m_{sep\text{-}SBM}(Q)$
- -Fit *colSBM*s with VEM starting from merged $\hat{Z}^m_{sep\text{-}SBM}(Q)$ (many initializations as a result of permutations within each $\hat{Z}^m_{sep\text{-}SBM}(Q)$)
- -Keep the b fitted models with the best BIC-L for each Q

while *BIC-L is increasing* **do**
- - Forward loop
- **for** $Q = Q_{min} + 1, \ldots, Q_{max}$ **do**
  - - Fit *colSBM* with Q blocks from initializations obtained by splitting a block in models with $Q - 1$ blocks
  - **if** π-$\delta\pi$*colSBMs* **then**
    - - Fit *colSBM* with $\widehat{S}_{qm} = \mathbf{1}_{\hat{\pi}^m_q > t}$ for different value of threshold t
- -Backward loop
- **for** $Q = Q_{max} - 1, \ldots, Q_{min}$ **do**
  - - Fit *colSBM* with Q blocks from initializations obtained by merging two blocks in models with $Q + 1$ blocks
  - **if** π-$\delta\pi$*colSBMs* **then**
    - - Fit *colSBM* with $\widehat{S}_{qm} = \mathbf{1}_{\hat{\pi}^m_q > t}$ for different value of threshold t
- - Among all fitted models, keep the b fitted models with the highest BIC-L for each Q

return $\widehat{Q} = \arg\max$ *BIC-L*$(\mathbf{X}, Q)$, *with the corresponding* $\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{Z}}$ *and* $\widehat{S}$ *for* π-*colSBM and* $\delta\pi$-*colSBM*.

3.6.2. Testing common connectivity structure

We can also use a model selection approach to choose which model from the 4 *colSBM*s and the *sep-SBM* is the most adapted to the collection. The most interesting comparison is to decide whether a collection of networks share the same connectivity structure by comparing the model selection criteria obtained for a given *colSBM* model with the one of *sep-SBM*. We decide that a collection of networks

share the same connectivity structure if:

$$\max_{Q} \text{BIC-L}_{colSBM}(\mathbf{X}, Q) > \sum_{m=1}^{M} \max_{Q_m} \text{BIC-L}_{SBM}(X^m, Q_m).$$

3.7. Partition of networks according to their mesoscale structure

If the networks in a collection do not have the same connectivity structure, we aim to partition them accordingly. We present hereafter a strategy to perform a clustering of the networks.

Clustering a collection of networks consists in finding a partition $\mathcal{G} = (\mathcal{M}_g)_{g=1,\dots,G}$ of $\{1, \dots, M\}$. Given $\mathcal{G}$, we set the following model on $\mathbf{X}$:

$$\forall g \in \{1, \dots, G\}, \quad \forall m \in \mathcal{M}_g, \quad X^m \sim \mathcal{F}\text{-SBM}(Q^g, \boldsymbol{\pi}^m, \delta_m \boldsymbol{\alpha}^g) \tag{3.15}$$

with $\delta_1 = 1$. Moreover, $\delta_m = 1$ for all m for *iid-colSBM* and π-*colSBM*s and $\boldsymbol{\pi}^m = \boldsymbol{\pi}^g$ for *iid-colSBM* and δ-*colSBM*. In other words, the networks belonging to the subcollection $\mathcal{M}_g$ share the same mesoscale structure given by a particular *colSBM*. To any partition $\mathcal{G}$ we associate the following score:

$$\text{Sc}(\mathcal{G}) = \sum_{g=1}^{G} \max_{Q^g=1,\dots,Q_{\max}} \text{BIC-L}((X^m)_{m\in\mathcal{M}_g}, Q^g). \tag{3.16}$$

where $\text{BIC-L}((X^m)_{m\in\mathcal{M}_g}, Q^g)$ is the BIC-L computed on the $\mathcal{M}_g$ subcollection of networks. The best partition $\mathcal{G}$ is chosen as follows:

$$\mathcal{G}^* = \text{argmax}_{\mathcal{G}} \text{Sc}(\mathcal{G}). \tag{3.17}$$

Computing the BIC-L for all the partitions $\mathcal{G}$ requires to consider the $2^M - 1$ non-empty subcollections of the networks $\mathcal{M}$, fit the *colSBM*s on these subcollections and then combine the associated BIC-L in order to be able to compute the scores (3.17). This can be done exhaustively provided that M is not too large but the computational cost becomes prohibitive as M grows.

To circumvent this point, we propose a less computationally intensive forward strategy, starting from $\mathcal{G} = (\{1, \dots, M\})$, and then progressively splitting the collection of networks. In order to explore the space of partitions of $\{1, \dots, M\}$, we define a dissimilarity measure between any pair of networks (m, m').

Definition of a dissimilarity measure between networks of a collection $(X^m)_{m\in\mathcal{M}}$. This relies on the following steps.

1. Infer *colSBM* on $(X^m)_{m\in\mathcal{M}}$ to get coherent clusterings of the nodes encoded in $\widehat{\boldsymbol{\tau}}^m = (\widehat{\tau}^m_{iq})_{i=1,\dots,n_m,q=1,\dots,\widehat{Q}}$, for any $m \in \mathcal{M}$. This step supplies the clusterings of the nodes of each network in terms of mesoscale structure. Note that the inference also supplies the $(\widehat{\delta}^m)_{m\in\mathcal{M}}$, these quantities being set to 1 if we work with the π-*colSBM* and *iid-colSBM*.

2. For each network m, compute:

$$\widetilde{n}^m_{qr} = \sum_{\substack{i,j=1\\ i\neq j}}^{n_m} \widehat{\tau}^m_{iq}\widehat{\tau}^m_{jr}, \quad \widetilde{\alpha}^m_{qr} = \frac{\sum_{\substack{i,j=1\\ i\neq j}}^{n_m} \widehat{\tau}^m_{iq}\widehat{\tau}^m_{jr}X^m_{ij}}{\widetilde{n}^m_{qr}}, \quad \widetilde{\pi}^m_q = \frac{\sum_{i=1}^{n_m}\widehat{\tau}^m_{iq}}{n_m}, \quad \widetilde{\delta}^m = \sum_{\substack{i,j=1\\ i\neq j}}^{n_m} X^m_{ij}$$

   with the convention that $\widetilde{\pi}^m_q = 0$ if $q \notin \mathcal{Q}_m$ and $\alpha^m_{qr} = 0$ if $\{q,r\} \not\subset \mathcal{Q}_m$. These quantities are the separated estimates of the parameters encoding the mesoscale structure for each network, computed from node clusterings obtained by considering all the networks jointly.

3. Then, for any pair of networks $(m,m') \in \mathcal{M}$ compute the dissimilarity:

$$D_{\mathcal{M}}(m,m') = \sum_{(q,r)=1}^{Q} \max\left(\widetilde{\pi}^m_q, \widetilde{\pi}^{m'}_q\right) \max\left(\widetilde{\pi}^m_r, \widetilde{\pi}^{m'}_r\right) \left(\frac{\widetilde{\alpha}^m_{qr}}{\widehat{\delta}_m} - \frac{\widetilde{\alpha}^{m'}_{qr}}{\widehat{\delta}_{m'}}\right)^2. \tag{3.18}$$

This dissimilarity measure quantifies to what extent the connectivity parameters inferred separately on each network of the pair are different. This is weighted by the size of the blocks and corrected in the case of δ-*colSBM* and $\delta\pi$-*colSBM* by the density parameter. If this dissimilarity measure is large, it means that enforcing the same connectivity patterns by estimating common connectivity parameters for these two networks is not relevant and the networks cannot be considered to be part of the same group.

An algorithm to cluster of the collection of networks Now, we use this dissimilarity to guide the search for the best partition of the collection of networks by using Algorithm 5 which consists in a recursive partitioning of the collection.

Algorithm 5: Clustering a collection of networks into two groups

Call: `Clust2Coll(`$\mathbf{X} = (X^m)_{m\in\mathcal{M}}$`)`

Data: $\mathbf{X} = (X^m)_{m\in\mathcal{M}}$ a collection of networks and $\mathcal{G} = \{\mathcal{M}\}$ the trivial partition in a unique group

begin
- -Fit *colSBM* on $\mathbf{X}$
- -Compute the score $Sc_0 = Sc(\mathcal{G})$
- -Compute the dissimilarity for all the networks in the collection $\left(D_{\mathcal{M}}(m,m')\right)_{m,m'\in\mathcal{M}}$
- -Apply a 2-medoïds algorithm to obtain G_1 and G_2 giving a partition of $\mathcal{M}$.
- -Compute $Sc^* = Sc(\mathcal{G}^*)$ where $\mathcal{G}^* = \{G_1, G_2\}$.

if $Sc_0 > Sc^*$ **then**
  return $\mathcal{G}$
else
  return $\left\{\texttt{Clust2Coll}\left((X^m)_{m\in G_1}\right), \texttt{Clust2Coll}\left((X^m)_{m\in G_2}\right)\right\}$

3.8. Simulation studies

In this section, we perform a large simulation study. The first study aims at testing the ability of the inference method to recover the number of blocks and the parameters for the π-*colSBM* model. The second study highlights the performances in terms of clustering of networks based on their mesoscale structure.

3.8.1. Efficiency of the inference procedure

Simulation paradigm Let us simulate data under the π-*colSBM* model with $M = 2$, $n_m = 100$ and $Q = 4$. α and $\boldsymbol{\pi}$ are chosen as:

$$\boldsymbol{\alpha} = .25 + \begin{pmatrix} 3\epsilon_\alpha & 2\epsilon_\alpha & \epsilon_\alpha & -\epsilon_\alpha \\ 2\epsilon_\alpha & 2\epsilon_\alpha & -\epsilon_\alpha & \epsilon_\alpha \\ \epsilon_\alpha & -\epsilon_\alpha & \epsilon_\alpha & 2\epsilon_\alpha \\ -\epsilon_\alpha & \epsilon_\alpha & 2\epsilon_\alpha & 0 \end{pmatrix}, \quad \boldsymbol{\pi}^1 = \sigma_1(.2, .4, .4, 0), \quad \boldsymbol{\pi}^2 = \sigma_2(0, \tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}). \tag{3.19}$$

with ϵ_α taking eight equally spaced values ranging from 0 to 0.24. For each value of ϵ_α, 30 datasets (X^1, X^2) are simulated, resulting in $8 \times 30 = 240$ datasets. More precisely, for each dataset, we pick uniformly at random two permutations of $\{1, \ldots, 4\}$ (σ_1, σ_2) with the constraint that $\sigma_1(4) \neq \sigma_2(1)$. This ensures that each of the two networks have a non-empty block that is empty in the other one. Then the networks are simulated with $\mathcal{B}$ern-SBM$_{100}(4, \boldsymbol{\alpha}, \boldsymbol{\pi}^m)$ with the previous parameters.

Each network has 2 blocks in common and their connectivity structures encompass a mix of core-periphery, assortative community and disassortative community

structures, depending on which 3 of the 4 blocks are selected for each network. ϵ_α represents the strength of these structures, the larger, the easier it is to tell apart one block from another.

Inference On each simulated dataset, we fit the *iid-colSBM*, *π-colSBM* and *sep-SBM* models. The inference is performed with the VEM algorithm and the BIC-L criterions presented in Sections 3.5 and 3.6.1.

Quality indicators The assess the quality of the inference, we compute the following set of indicators for each simulated dataset.

- First, for each dataset, we put in competition *π-colSBM* with *sep-SBM* and *iid-colSBM* respectively. To do so, for each dataset, we compute the BIC-L of each model *π-colSBM* is preferred to *sep-SBM* (resp. *iid-colSBM*) if its BIC-L is greater.
- Secondly, when considering the *π-colSBM*, we compare $\widehat{Q}$ to its true value ($Q = 4$).
- For *π-colSBM* and Q fixed to its true value ($Q = 4$), we evaluate the quality of recovery of the support matrix S by calculating:

$$\mathrm{Rec}(\widehat{S}, S) = \max_{\sigma \in \mathfrak{S}_4} \mathbf{1}_{\{\forall q, m S_{mq} = \widehat{S}_{m\sigma(q)}\}} \tag{3.20}$$

  the greater the better.
- In order to evaluate the ability to recover the true connectivity parameter in the *π-colSBM* model, we compare $\widehat{\boldsymbol{\alpha}}$ to its true value for the true number of blocks $Q = 4$ through:

$$\mathrm{RMSE}(\widehat{\alpha}, \alpha) = \min_{\sigma \in \mathfrak{S}_4} \sqrt{\frac{1}{16} \sum_{1 \leq q, r \leq 4} (\widehat{\alpha}_{\sigma(q)\sigma(r)} - \alpha_{qr})^2}, \tag{3.21}$$

  the σ being there to correct the possible label switching of the blocks.
- Finally, we judge the quality of our clustering with the Adjusted Rand Index (Hubert and Arabie, 1985, ARI = 0 for a random clustering and 1 for a perfect recovery). For each network, for the *π-colSBM*, using $\hat{Q}$, we compare the block memberships to the real ones by taking the average over the two networks

$$\overline{\mathrm{ARI}} = \tfrac{1}{2}\left(\mathrm{ARI}(\widehat{\boldsymbol{Z}}^1, \boldsymbol{Z}^1) + \mathrm{ARI}(\widehat{\boldsymbol{Z}}^2, \boldsymbol{Z}^2)\right)$$

  and by computing it on the whole collection of nodes

$$\mathrm{ARI}_{1,2} = \mathrm{ARI}\left((\widehat{\boldsymbol{Z}}^1, \widehat{\boldsymbol{Z}}^2), (\boldsymbol{Z}^1, \boldsymbol{Z}^2)\right).$$

	Model comparison (π-*colSBM* vs $\cdot$)		Estimation of Q and S under π-*colSBM*				Parameter & Clustering accuracy under π-*colSBM* (mean $\pm$ sd)		
ϵ_α	*sep-SBM*	*iid-colSBM*	$\mathbf{1}_{\widehat{Q}<4}$	$\mathbf{1}_{\widehat{Q}=4^\star}$	$\mathbf{1}_{\widehat{Q}>4}$	$\mathrm{Rec}(\hat{S}, S)$	$\mathrm{RMSE}(\widehat{\alpha},\alpha)$	$\overline{\mathrm{ARI}}$	$\mathrm{ARI}_{1,2}$
0	1	0	1	0	0	0	$.1 \pm .002$	0	0
.04	.9	0	1	0	0	0	$.13 \pm .003$	0	0
.08	.47	.33	.97	.03	0	.03	$.14 \pm .035$	$.24 \pm .27$	$.15 \pm .2$
.12	.47	.8	.3	.7	0	.7	$.1 \pm .069$	$.91 \pm .06$	$.6 \pm .27$
.16	.8	1	0	.93	.07	.93	$.04 \pm .06$	$.98 \pm .01$	$.89 \pm .2$
.2	.93	1	0	.97	.03	.93	$.02 \pm .04$	1	$.98 \pm .08$
.24	1	1	0	1	0	1	$.01 \pm .003$	1	1

Table 3.2 – **Accuracy of the inference for varying $\boldsymbol{\alpha}$.** All the quality indicators are averaged over the 30 simulated datasets.

.

All these quality indicators are averaged among the 30 simulated datasets. The results are provided in Table 3.2. Each line corresponds to the 30 datasets simulated with a given value of ϵ_α. The first columns concatenate the results of the model comparison task. The following set of columns is about the selection of Q and the estimation of S. The last columns supply the RMSE on $\boldsymbol{\alpha}$ and the ARI.

Results For the model comparison, when ϵ_α is small ($\epsilon_\alpha \in [0, .04]$), the simulation model is close to the Erdős-Rényi network and it is very hard to find any structure beyond the one of a single block. As such, the *iid-colSBM* and π-*colSBM* models are equivalent and *iid-colSBM* is preferred to *sep-SBM*.

We observe a transition when $\epsilon_\alpha = .08$ where we become able to recover the true number of blocks $\widehat{Q} = 4$ and the support of the blocks given the true number of blocks. During this transition, the model selection criterion is about half of the time in favor of *sep-SBM* i.e. the model with no common connectivity structure between the networks.

From $\epsilon_\alpha = .16$, we recover the true number of blocks and their support most of the time and the common structure obtained by the π-*colSBM* is found to be relevant. Note that when we are able to recover the true number of blocks, we are also able to recover their support almost every time.

For both the estimation of the parameters and the ARIs, the results mainly follow our ability to recover the true number of blocks, with the error of estimation of the parameters slowly decreasing from $\epsilon_\alpha = 0.12$. $\overline{\mathrm{ARI}}$ goes to 1 a bit faster than $\mathrm{ARI}_{1,2}$, denoting our ability to recover faster the real clustering of each network than to match the blocks between the networks. This is directly linked with the detection of the true number of blocks and their support. Indeed, to get $\mathrm{ARI}_{1,2} = 1$, we need $\mathrm{Rec}(\widehat{S}, S) = 1$ while the effective block number for each network is of only $Q = 3$, meaning that even with the wrong selected model we can still reach $\overline{\mathrm{ARI}} = 1$.

	Model comparison			Estimation of Q and S under π-*colSBM*	
ϵ_π	*iid-colSBM*	π-*colSBM*	*sep-SBM*	$\mathbf{1}_{\widehat{Q}=4}$	$\mathrm{Rec}(\hat{S}, S)$
0	**1**	0	0	1	1
.04	1	**0**	0	1	1
.08	.9	**.1**	0	1	1
.12	.5	**.5**	0	1	1
.16	.37	**.63**	0	1	.97
.2	.03	**.97**	0	1	1

Table 3.3 – **Model selection for varying mixture parameters**. The number of blocks $\widehat{Q}$ is given for the π-*colSBM*. The similarity of the block support to the true one $\mathrm{Rec}(\hat{S}, S)$ is given for π-*colSBM* with $Q = 4$.

3.8.2. Capacity to distinguish π-*colSBM* from *iid-colSBM*

We aim to understand how well we are able to differentiate *iid-colSBM* from π-*colSBM* depending on the block proportions. To do so, we fix $\boldsymbol{\alpha}$ as in equation (3.19) with $\epsilon_\alpha = 0.16$ and set $\boldsymbol{\pi}$ as follows:

$$\boldsymbol{\pi}^1 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \quad \text{and} \quad \boldsymbol{\pi}^2 = \sigma\left(\frac{1}{4} - \epsilon_\pi, \frac{1}{4} - \frac{\epsilon_\pi}{2}, \frac{1}{4} + \epsilon_\pi, \frac{1}{4} + \frac{\epsilon_\pi}{2}\right),$$

with ϵ_π taking 6 values equally spaced in $[0, .20]$. σ is a random permutation of the blocks. We simulate 30 different collections for each value of ϵ_π.

Here again, we put in competition π-*colSBM* with *iid-colSBM* and *sep-SBM* and select a model if its BIC-L the greater than the two other ones. Then, for π-*colSBM* we compare $\widehat{Q}$ to 4 and evaluate our ability to recover S. The results are provided in Table 3.3.

First notice that, since we chose $\epsilon_\pi \ll 0.25$, we do not simulate any empty block. As a consequence, the inference of the model is quite easy and we are able to recover the true number of blocks and the right support for the π-*colSBM* model almost always. When $\epsilon_\pi = 0$, $\boldsymbol{\pi}^1 = \boldsymbol{\pi}^2$ and the model reduces to *iid-colSBM*. This remark explains why *iid-colSBM* is preferred to π-*colSBM* when $\epsilon_\pi < .08$. As ϵ_π increases, π-*colSBM* gets more and more selected, highlighting our capacity to recover the simulated structure.

3.8.3. Partition of networks

The third simulation experiment aims at illustration our capacity to perform a partition of a collection of networks based on their structure, as presented in Section 3.8.3.

Simulation scenario For *iid-colSBM*, π-*colSBM* and δ-*colSBM* and $\delta\pi$-*colSBM*, we simulate $M = 9$ undirected networks with 60 nodes and $Q = 3$

blocks. The block proportions are chosen as follows:

$$\boldsymbol{\pi}^1 = (.2, .3, .5)$$

and for all $m = 2, \ldots, 9$

$$\boldsymbol{\pi}^m = \begin{cases} \boldsymbol{\pi}^1 & \text{for} \quad \textit{iid-colSBM} \quad \text{and} \quad \delta\textit{-colSBM} \\ \sigma_m(\boldsymbol{\pi}^1) & \text{for} \quad \pi\textit{-colSBM} \quad \text{and} \quad \delta\pi\textit{-colSBM} \end{cases}$$

where σ_m is a permutation of $\{1, 2, 3\}$ proper to network m and $\sigma(\boldsymbol{\pi}) = (\pi_{\sigma(i)})_{i=1,\ldots,3}$. The networks are divided into 3 groups of 3 networks with connectivity parameters as follows:

$$\boldsymbol{\alpha}^{\text{as}} = .3 + \begin{pmatrix} \epsilon & -\frac{\epsilon}{2} & -\frac{\epsilon}{2} \\ -\frac{\epsilon}{2} & \epsilon & -\frac{\epsilon}{2} \\ -\frac{\epsilon}{2} & -\frac{\epsilon}{2} & \epsilon \end{pmatrix}, \qquad \boldsymbol{\alpha}^{\text{cp}} = .3 + \begin{pmatrix} \frac{3\epsilon}{2} & \epsilon & \frac{\epsilon}{2} \\ \epsilon & \frac{\epsilon}{2} & 0 \\ \frac{\epsilon}{2} & 0 & -\frac{\epsilon}{2} \end{pmatrix}, \qquad \boldsymbol{\alpha}^{\text{dis}} = .3 + \begin{pmatrix} -\frac{\epsilon}{2} & \epsilon & \epsilon \\ \epsilon & -\frac{\epsilon}{2} & \epsilon \\ \epsilon & \epsilon & -\frac{\epsilon}{2} \end{pmatrix},$$

with $\epsilon \in [.1, .4]$. $\boldsymbol{\alpha}^{\text{as}}$ represents a classical assortative community structure, while $\boldsymbol{\alpha}^{\text{cp}}$ is a layered core-periphery structure with block 2 acting as a semi-core. Finally, $\boldsymbol{\alpha}^{\text{dis}}$ is a disassortative community structure with stronger connections between blocks than within blocks. If $\epsilon = 0$, the three matrices are equal and the 9 networks have the same connection structure. Increasing ϵ differentiates the 3 clusters of networks. For δ-*colSBM* an $\delta\pi$-*colSBM*, we add density parameters $\delta^1 = \delta^4 = \delta^7 = 1$, $\delta^2 = \delta^5 = \delta^8 = 0.75$ and $\delta^3 = \delta^6 = \delta^9 = 0.5$.

We simulate each of these configurations 30 times. We apply the strategy exposed in Section 3.7 and evaluate the recovery of the simulated network partition.

Results We assess the quality of our procedure by comparing the obtained classifications of the network collection with the simulated one through the ARI index. As ϵ grows we are able to better differentiate the networks and do so almost perfectly on all *colSBM* setup. Note that Adding complexity slightly deteriorates the results as we recover the partition better for *iid-colSBM* and π-*colSBM*s than for δ-*colSBM* and $\delta\pi$-*colSBM* .

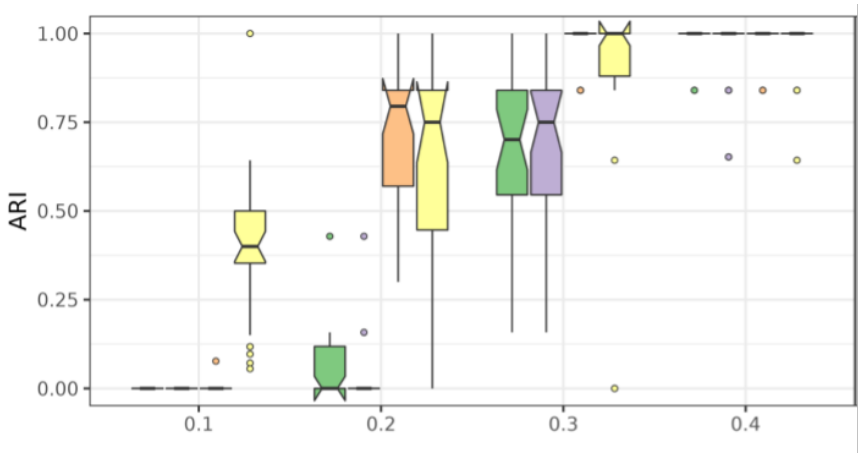


Figure 3.2 – **Partition of networks**. ARI of the recovered partition of networks. Orange is for the *iid-colSBM*, green for the δ-*colSBM*, purple for the $\delta\pi$-*colSBM* and yellow for the π-*colSBM*.

3.8.4. Finding finer block structures

Finally, we perform a last simulation study in order to illustrate the fact that, for particular configurations, using a *colSBM* model on a collection of networks favors the transfer of information between networks and allows to find finer block structures on the networks. We consider the core-periphery structure configuration described in Equation (3.22) with $\epsilon = .4$. In that case $Q = 3$.

We simulate a collection of 5 networks. 4 networks are of respective size $(90, 90, 120, 120)$. The last network is either smaller with only 60 nodes in the case of *iid-colSBM* or has a less marked structure ($\delta = .5$) for the δ-*colSBM* and $\delta\pi$-*colSBM* models.

Our goal is to recover the true connection structure of this last network X^5. To do so, we compare the results obtained using either a standard single *SBM* on X^5, or using the corresponding *colSBM* inferred with $M = 2, 3$ and 5 networks. We study $\widehat{Q}$ in the various scenarii. In the simulation experiment, we obtained only $\widehat{Q} = 2$ or 3. The experiment is repeated 30 times. The results are depicted in Figure 3.3.

For the 4 models of simulation, the simple *SBM* recovers 2 blocks most of the time. For *iidcolSBM*, we always recover the 3 blocks while for the other case, we improve the ability to recover the true number of blocks when the quantity of information available from the other networks grows, either by augmenting the number of networks or by augmenting the number of nodes.

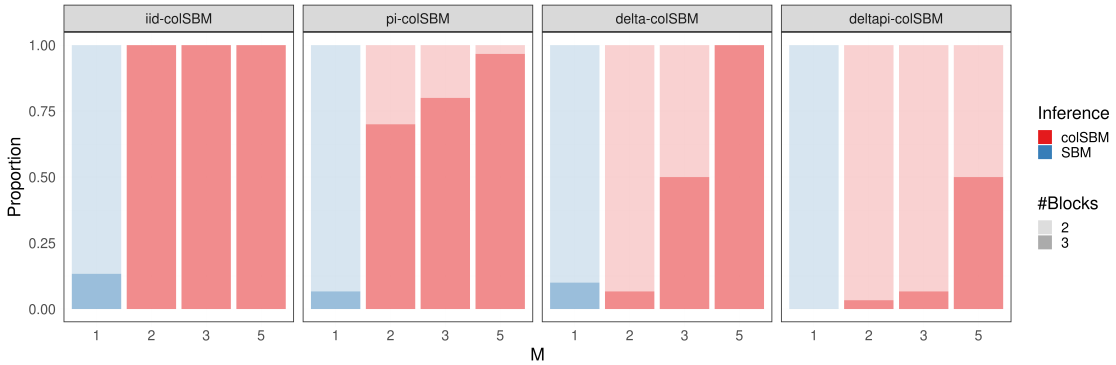


Figure 3.3 – **Finding finer block structures.** Cumulative barplot of $\widehat{Q}$ by the SBMs (blue) and the adequate $colSBM$ (red) under the different simulation scenario. The number of blocks to be recovered is 3 and the darkest shade corresponds to $\widehat{Q} = 3$.

3.9. Application to Food Webs

In this section, we demonstrate the interest of our models on 2 collections of ecological networks . The first one consists of a collection of 3 stream food webs issued from the dataset of Thompson and Townsend (2003) described in Section 3.2 at page 100. We analyze in detail the different structures given by the different models and we show how using networks with some common structure helps the prediction of missing information in the networks.

The second dataset is a collection of 67 networks issued from the Mangal database (Vissault et al., 2020). We will use our model to propose a partition of the networks into groups of networks with common mesoscale structures.

3.9.1. Joint analysis of 3 stream food webs

In Section 3.2 at page 101, we fitted *sep-SBM* obtained 5 blocks for Martins, 3 blocks for Cooper and 4 blocks for Herlzier. For reminder, a matricial representation of the block reordered food webs was shown in Figure 3.1. Each food web comprises of 2 blocks of basal species (the 2 bottom blocks).

Finding a common structure between the networks We now fit the four $colSBM$ models in order to find a common structure among the 3 networks. First, notice that our model selection criterion greatly favors common network structure above separated one: BIC-L $= -2080$ for *sep-SBM* versus respectively -1966, -1982, -1969 and -1989 for *iid-colSBM*, π-*colSBM*, δ-*colSBM*, $\delta\pi$-*colSBM*. The structures of the collection under the different models are represented in Figure 3.4. In this figure, the square matrix represents the estimated connection matrix $\widehat{\boldsymbol{\alpha}}$, while the cumulative bar plot on the right represents the $\widetilde{\boldsymbol{\pi}}^m$ (see definition page 114).

For each model, the basal species are separated into 2 blocks (bottom blue blocks in Figure 3.4), similar to the one obtained with the SBMs. The obtained structure slightly differs as described hereafter.

- *iid-colSBM* highlights 5 blocks in total. Block 3 (light green) is a small block of intermediate trophic level species (ones that prey on basal species and are being preyed on by higher trophic levels) with some within block predation. The higher trophic level is divided into 2 more blocks, block 2 (dark green) only preys on the 2 basal blocks, while block 1 (pink) preys on the intermediate block 3 level but only on the most connected basal species block.

- π-*colSBM* leads to only 4 blocks, with blocks 1 and 2 corresponding to the top and intermediate trophic levels. There are no empty blocks and the block proportions are still quite homogeneous, with block 2 grossly corresponding to block 1 and 2 of *iid-colSBM*, rendering the flexibility of the π-*colSBM* of little use compared to the *iid-colSBM* on this collection.

- With δ-*colSBM*, the species are clustered into 6 blocks and the networks have different estimated density parameters: $\hat{\delta} \approx (1, 1.4, 1.6)$. Block 1 (red) corresponds to the top trophic levels, with blocks 2, 3 and 4 being intermediate levels. Block 4 (light green) is a well connected group with both block 1 and the basal species blocks. Block 2 (pink) is huge and only preys on block 6 while block 3 (dark green) is a small group of species that preys on both basal species blocks.

- Finally, $\delta\pi$-*colSBM* clusters the species into 5 blocks with more heterogeneous block proportions and the networks have different estimated density parameters: $\hat{\delta} \approx (1, 1.1, 1.5)$. Block 1 (pink) is empty on network 2 (Herlzier) and very small on network 1 (Cooper), it corresponds to a block of high trophic level species with within blocks predation, it is well connected to block 3 (light green, intermediate trophic level species) and 5 (dark blue, well connected basal species). Block 2 is of much larger size (especially for network Herlzier) and preys on the same blocks than block 1 but with lower probability for block 3.

Remark. *On this collection, the entropy of the clustering is much lower on π-colSBM and $\delta\pi$-colSBM than on iid-colSBM and δ-colSBM. In these models, to ensure homogeneous block proportions between networks, some nodes tend to get a fuzzy clustering and sit between several blocks (the variational parameters do not concentrate on one block). This phenomenon is taken into account by our model selection criterion which tends to favor models with higher entropy than models with well separated clusters.*

Prediction of missing links and dyads Since we have been able to find some common structures between the 3 networks, we now examine if these structures could be used to help recover some information on networks with incomplete information. We proceed as follows: we choose a network m and remove a proportion $K \in [.1, .8]$ of

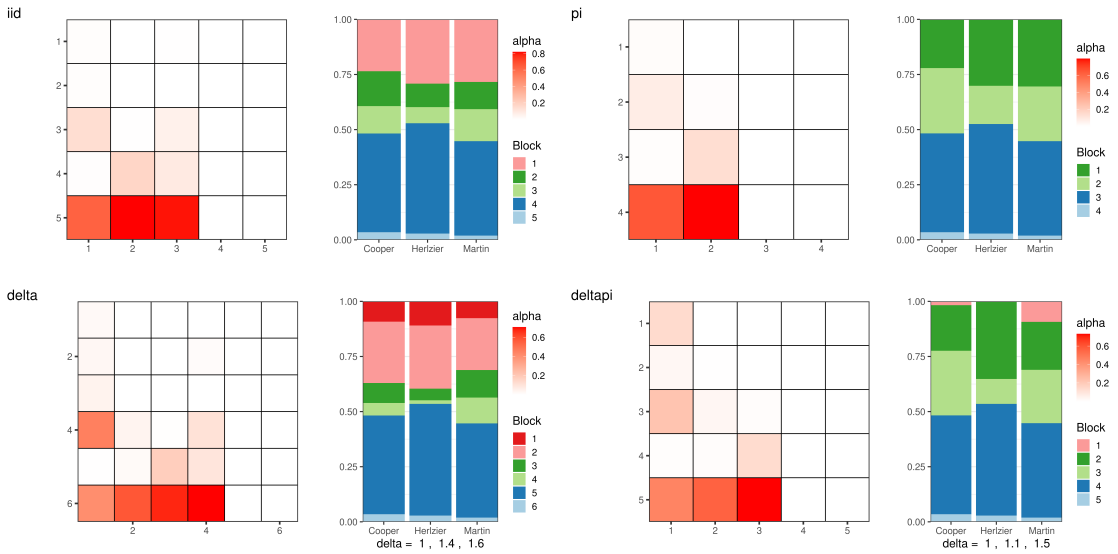


Figure 3.4 – **Estimated structure of the collection of 3 stream food webs with the four *colSBM* models**. For each network, the matrix on the left is the estimated parameter connectivity $\widehat{\boldsymbol{\alpha}}$, the barplot on the right depicts $\widetilde{\boldsymbol{\pi}}^{(m)}$. The ordering is done by trophic level from bottom to top and right to left. For (δ-$\delta\pi$)*colSBM*s we give $\hat{\delta}$ below the barplot.

- the existing links uniformly at random *for the missing link experiment*
- or of the existing dyads (both 0 and 1) by encoding them as `NA` *for the missing dyads experiment.*

Then, for the missing link experiment, we try to recover where the missing links are among all non existing ones. For the missing dyad experiment, we predict the probability of existence of missing dyads (`NA` entries). Under the *colSBM*, the probability of a link between species i and j for network m is predicted by:

$$\widehat{p}_{ij}^{m} = \sum_{q,r \in \hat{\mathcal{Q}}_m} \widehat{\tau}_{iq}^{m} \widehat{\tau}_{jr}^{m} \widehat{\delta}^{m} \widehat{\alpha}_{qr}.$$

We resort to the area under the ROC curve to evaluate the capacity of the different models to recover this information. For each value of K, each experiment is repeated 30 times and the results are shown in Figure 3.5.

First, let us notice that these stream food webs networks have a structure that is well explained by an SBM. When there is little information missing ($K < .3$) the ROC AUC is over 0.9. Besides, with 70% of missing links or dyads, we still predict better than a random guess (ROC AUC $\approx .75$). As there is a common structure between the 3 networks, there is a lot of information to be taken from the ones with no missing information. Starting from $K \geq .3$, the *colSBM*s outperform the *sep-SBM* on both experiments. Even for $K = .8$, the prediction is still high. About the difference between the *colSBM*s, for the missing links experiment, as we remove links from one of the network, its density decreases and the models with a density parameters, δ-*colSBM* and $\delta\pi$-*colSBM* have a built-in mechanism that

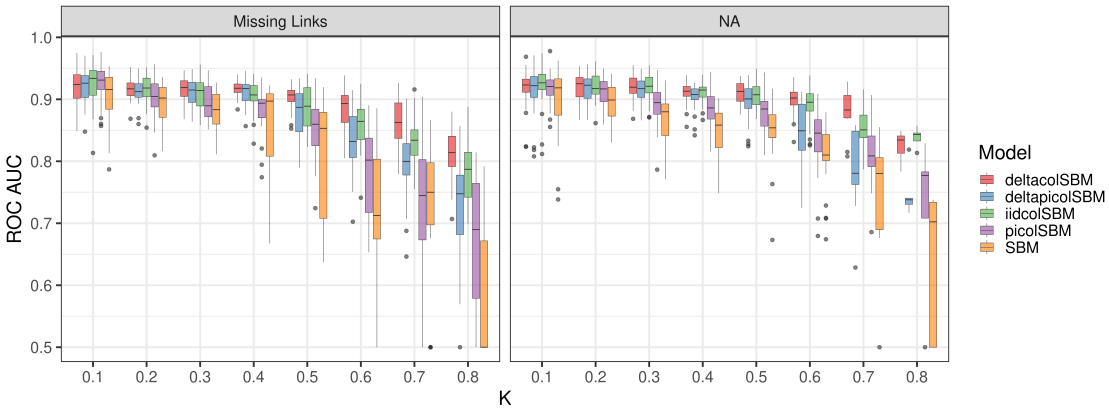


Figure 3.5 – Prediction of missing links and `NA` entries on stream food webs.

compensates this fact. As a consequence, these models yield to better predictions than *iid-colSBM* or π-*colSBM* for large values of K.

Another noteworthy comment on both the `NA` and missing links experiments is that as K grows, the amount of information on the modified network gets lower. Hence, $(\pi-\delta\pi)colSBM$s lacks the statistical power to separate blocks and will empty blocks on this network. This affects our capability to predict the trophic links. On the other hand, $(iid\text{-}\delta)colSBM$s will force some separation of the species into blocks, and as the information from the other networks is relevant it still has good prediction performance for large K.

3.9.2. Partition of a collection of 67 predation networks

Now, we consider a collection of 67 predation networks which are all directed networks with more than 30 species, from the Mangal database (Vissault et al., 2020). They are issued from 33 datasets each containing between 1 and 10 networks. The number of species ranges from 31 to 106 (3395 in total) by networks; the networks have density ranging from .01 to .32 (14934 total predation links). This dataset is too heterogeneous to find a common structure that will fit well on all the networks. Therefore we propose to use a π-*colSBM* to look for a partition of the networks into groups sharing common connectivity structures.

We present the obtained network clustering and the connectivity structure of each group in Figure 3.6 as well as a contingency table of the obtained groups with the different datasets of the Mangal database. Our comments on each group follow.

A This group consists of 7 networks and 12 blocks are required to describe this cluster. 5 networks of which are issued from the same dataset (id: 80) and These 5 networks populate the 12 blocks, while the other 2 networks only populate parts of them. The average density is about .18. From the ecological point of view, the blocks can be divided into 3 heterogeneous sets: block 1 to 3 represent the higher trophic levels, block 4 to 8 the intermediate ones and block 9 to 12 the lower ones.

B This group of 50 networks is the most heterogeneous, it consists mostly of sparse networks, issued from various datasets. Most networks populate only parts of the blocks. The first 3 blocks represent higher trophic levels, with block 1 feeding on almost all the blocks. From block 9 to 11, we observe some symmetry in the connectivity matrix rendering it difficult to order the blocks by trophic order.

C This is a small group of only 3 networks of heterogeneous size ($n = (79, 38, 33)$) but homogeneous density (average .17) issued from two different datasets. The structure of the 2nd and 3rd networks is encompassed into the 1st one, occupying just blocks 1, 5, 6 and 7 of the 7 blocks. The ecological connectivity structure is clear, a top trophic level on block 1 feeds on all the other blocks. Blocks 2 to 4 are the intermediate level species that only feed on the three last blocks. Block 5 is also an intermediate trophic levels block, while the basal species are divided into the last two blocks depending on there propension to be preyed.

D A small group of 2 homogeneous networks from the same dataset with 50 species each. The 6 block structure follows a kind of stairs shape structure: blocks 1 and 2 are the top trophic species that only feed on block 3 and 4. Block 3 to 5 consist of intermediate trophic level species, with block 3 preying on block 5, and blocks 4 and 5 preying on both blocks 5 and 6.

E The last group consists of 5 networks containing from 38 to 45 species, four of them are issued from the same dataset (id = 144) while the last one is an outlier of a dataset of 10 networks (id = 157), the others networks from this dataset belonging to group **B**. The structure consists of 4 blocks, and is a simpler structure of the stair shape structure of group **D**.

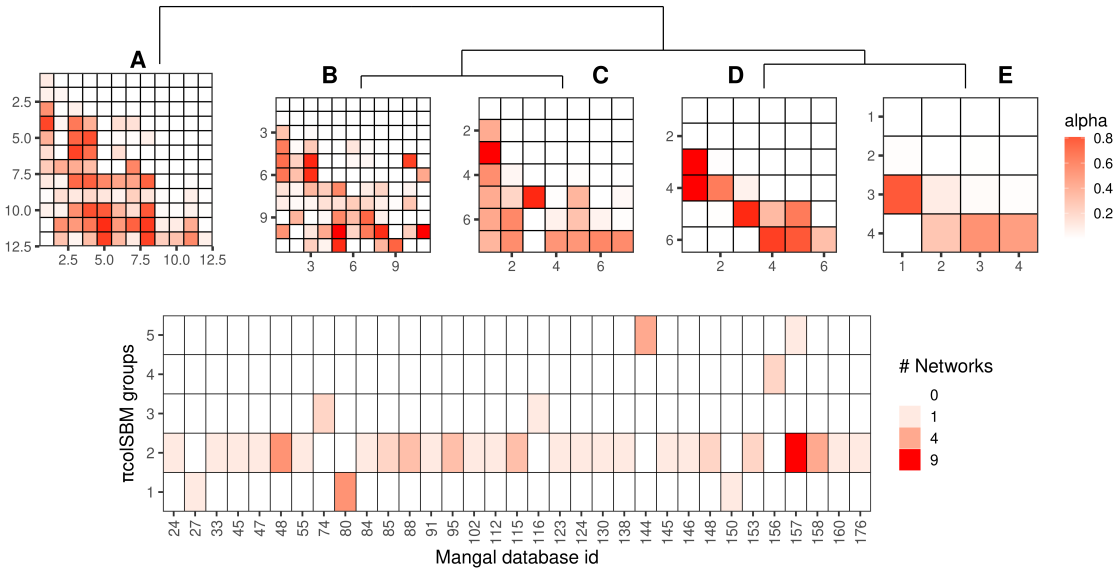


Figure 3.6 – Above: Classification and connectivity structures of a collection of 67 predation networks from the Mangal database into 5 groups. The length of the dendrogram is given by the difference in BIC-L to the best model. Below: Contingency table of the classification found by $\pi colSBM$ and the different datasets from the Mangal database.

3.10. Discussion

In this paper, we proposed a new method to find a common structure and compare different structures of a collection of networks which we do not assume to share common nodes. This method is very general and could be applied to networks sharing common nodes as well, such as temporal or multiplex networks. Starting from our most basic model, – namely the *iid-colSBM* – we refined it by proposing models allowing for different mixture distributions and even empty blocks (π-$colSBM$), models allowing to find common structure for networks of different density (δ-$colSBM$) or even models allowing both ($\delta\pi$-$colSBM$). The model selection criterion we derived can be used to select the number of blocks but also to choose which $colSBM$ fits better the data. We also presented a strategy providing a partition of a collection of networks into groups of networks sharing common connectivity patterns.

The idea behind these models is very general and could be extended to other types of networks. In ecology, bipartite and multipartite networks are common and the model extension is straightforward (although some additional modeling choices arise when considering π-$colSBM$, δ-$colSBM$ or $\delta\pi$-$colSBM$), the main difficulty would then lie in the algorithmic part. The main idea of this article could also be extended to the Degree Corrected SBM (Karrer and Newman, 2011) which is quite used in practice. Finally, the nested version of the SBM proposed by Peixoto (2014b), by allowing a hierarchy on the blocks would be particularly adapted to π-$colSBM$ and $\delta\pi$-$colSBM$s.

Finally, we notice during our simulations and applications that the *colSBM*s allow to find a larger number of blocks compared to the *sep-SBM* and so lead to a finer resolution of the mesoscale structure of the networks. This resolution limit problem was one of the motivations of Peixoto (2014b) and we believe that this direction should be explored further for collections of networks.

Acknowledgments

The authors would like to thank Stéphane Robin for his helpful advice. This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. This work was partially supported by the grant ANR-18-CE02-0010-01 of the French National Research Agency ANR (project EcoNet).

3.A. Proof of identifiability

Proof. Celisse et al. (2012) states that the parameters $(\boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$ of a $SBM_{n_m}(Q_m, \boldsymbol{\pi}^m, \boldsymbol{\alpha}^m)$ are identifiable upto label switching of the blocks from a single observed network X^m provided

1. $n_m > 2Q_m$
2. $(\boldsymbol{\alpha}^m \cdot \boldsymbol{\pi}^m)_q \neq (\boldsymbol{\alpha}^m \cdot \boldsymbol{\pi}^m)_r$ for all $(q \neq r) \in \{1, \dots, Q_m\}^2$

From this result, we prove the identifiability of our *colSBM*s.

iid-colSBM Under this model, for all $m = 1, \dots M$, $X^m \sim SBM_{n_m}(Q, \boldsymbol{\pi}, \boldsymbol{\alpha})$. As a consequence, assuming that $\exists m^* : n_{m^*} \geq 2Q$, we obtain the identifibaility of $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ (Celisse et al., 2012), provided that $(\boldsymbol{\alpha} \cdot \boldsymbol{\pi})_q \neq (\boldsymbol{\alpha} \cdot \boldsymbol{\pi})_r$ for all $(q \neq r) \in \{1, \dots, Q\}^2$.

δ-colSBM Under this model, for all $m = 1 \dots, M$, $X^m \sim SBM_{n_m}(Q, \boldsymbol{\pi}, \delta_m \boldsymbol{\alpha})$. Assuming that $\exists m^* : n_{m^*} \geq 2Q$ and $\delta_{m^*} = 1$, we apply the theorem of Celisse et al. (2012) and obtain the identifiability of $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ under the condition $(\boldsymbol{\alpha} \cdot \boldsymbol{\pi})_q \neq (\boldsymbol{\alpha} \cdot \boldsymbol{\pi})_r$ for all $(q \neq r) \in \{1, \dots, Q\}^2$. Now, for any $m \neq m^*$, by definition of the SBM model,

$$\mathbb{E}[X_{ij}^m] = \delta_m \boldsymbol{\pi}' \boldsymbol{\alpha} \boldsymbol{\pi},$$

which proves that δ_m is identifiable.

π-colSBM Note that under *π-colSBM*, we have $X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q_m, \widetilde{\boldsymbol{\pi}}^m, \widetilde{\boldsymbol{\alpha}}_m)$ where $\widetilde{\boldsymbol{\pi}}^m$ is a vector of non-null proportions of length Q_m and $\widetilde{\boldsymbol{\alpha}}^m$ is the restriction of $\boldsymbol{\alpha}$ to $\mathcal{Q}_m$. Under assumptions 1 and 2, applying the theorem of Celisse et al. (2012) on the distribution of X^m, we obtain the identifiability of the parameters $\widetilde{\boldsymbol{\pi}}^m$ and $\widetilde{\boldsymbol{\alpha}}^m$. However, the identifibaility of each $\widetilde{\boldsymbol{\pi}}^m$ and $\widetilde{\boldsymbol{\alpha}}^m$ is established up to a label switching of the blocks in each network. We now have to find a coherent reordering between the networks and the null block proportions in each network.

We are know able to build the complete matrix $\boldsymbol{\alpha}$ using the $\widetilde{\boldsymbol{\alpha}}^m$. First we fill the diagonal of $\boldsymbol{\alpha}$ which is composed of $(\text{diag}(\widetilde{\boldsymbol{\alpha}}^m))_{m=1,\dots,M}$, taking the unique values and sorting them by increasing order, such that $\alpha_{11} < \alpha_{22} < \dots < \alpha_{QQ}$. This task is possible because of Assumption 3.

Now we get back to $\widetilde{\boldsymbol{\pi}}^m$ and reorganize them to match with $\boldsymbol{\alpha}$. For any m we define $\phi_m : \{1, \dots, Q_m\} \to \{1, \dots, Q\}$ such that $\alpha_{\phi_m(q), \phi_m(q)} = \widetilde{\alpha}_{qq}^m$. With the (ϕ_m) we are able to fullffil the rest of $\boldsymbol{\alpha}$ as: $\alpha_{\phi_m(q)\phi_m(r)} = \widetilde{\alpha}_{qr}^m$ for all $(q, r) \in \{1, \dots, Q_m\}^2$. Finally, we define π^m a vector of size Q such that:

$$\pi_q^m = \begin{cases} 0 & \forall q \in \{1, \dots, Q\} \backslash \phi^m(\{1, \dots, Q_m\}) \\ \widetilde{\pi}_{\phi_m^{-1}(q)}^m & \forall q \in \phi^m(\{1, \dots, Q_m\}) \end{cases}$$

$\delta\pi$-colSBM We now consider the model where

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q, \boldsymbol{\pi}^m, \delta_m \boldsymbol{\alpha}) \tag{3.22}$$

We set the following conditions.

1. $\forall m = 1, \dots M, : n_m \geq 2Q_m$
2. $\delta_1 = 1$

   For $Q \geq 2$:
3. $(\boldsymbol{\alpha} \cdot \boldsymbol{\pi}^m)_q \neq (\boldsymbol{\alpha} \cdot \boldsymbol{\pi}^m)_r$ for all $(q \neq r) \in \mathcal{Q}_m^2$
4. $\forall m \in \{1, \dots, M\}, Q_m \geq 2$
5. Each diagonal entry of $\boldsymbol{\alpha}$ is unique

   For $Q \geq 3$:
6. There is no configuration of four indices $(q, r, s, t) \in \{1, \dots, Q\}$ such that $q \neq s$, $r \neq t$ and $\alpha_{qq}/\alpha_{rr} = \alpha_{ss}/\alpha_{tt}$.
7. $\forall m \geq 2, |\mathcal{Q}_m \cap \cup_{l:l<m} \mathcal{Q}_l| \geq 2$.

Like in model *π-colSBM*, Equation (3.22) implies that marginally,

$$X^m \sim \mathcal{F}\text{-SBM}_{n_m}(Q_m, \tilde{\boldsymbol{\pi}}^m, \tilde{\boldsymbol{\alpha}}^m) \quad \text{where} \quad \tilde{\boldsymbol{\alpha}}^m = \delta_m (\boldsymbol{\alpha}_{qr})_{q,r \in \mathcal{Q}_m}.$$

Applying Celisse et al. (2012) on the distribution of X^m, we obtain the identifiability of the parameters $\tilde{\boldsymbol{\pi}}^m$ and $\tilde{\boldsymbol{\alpha}}^m$ (Assumptions 1 and 3). We know have to use the label switching to make the sructures of the networks match and to take into account the empty blocks. We separate the cases $Q = 2$ and $Q > 3$.

• For $Q = 2$, we do not allow empty clusters (Assumption 4, $Q_m \geq 2$) so $\tilde{\boldsymbol{\alpha}}^m = \delta_m \boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\pi}}^m = \boldsymbol{\pi}^m$. Using the fact that $\delta_1 = 1$ (Assumption 2), we identify $\boldsymbol{\pi}^1$ and $\boldsymbol{\alpha}$. Since we know that the diagonal entries of $\boldsymbol{\alpha}$ are unique (Assumption 5), $\boldsymbol{\alpha}$ can be chosen such that $\alpha_{11} > \alpha_{22}$, thus provided we order well the blocks in each network (with strictly increasing intro block density) we identify the $\boldsymbol{\pi}^m$ uniquely and not up to label switching.

• For $Q \geq 3$, for each $m \in \{1, \dots, M\}$, by Assumptions 1 and 3 and using the marginal distributions, we are able to identify $\tilde{\boldsymbol{\alpha}}^m$ and $\tilde{\boldsymbol{\pi}}^m$.

Using the fact that $\delta_1 = 1$ (Assumption 2) and the fact that the entries of the diagonal of $\boldsymbol{\alpha}$ are unique, we can do as *π-colSBM* and identify $\boldsymbol{\pi}^1$ and $\boldsymbol{\alpha}_{qr}$, for any $q, r \in \mathcal{Q}_1$

Then for $m = 2$, up to a relabelling of the blocks in $\tilde{\boldsymbol{\alpha}}^2$, we can define $\delta_2 = \tilde{\alpha}_{11}^2/\alpha_{11} = \tilde{\alpha}_{22}^2/\alpha_{22}$ since there are at least two blocks in network $m = 2$ that correspond to two blocks already identified in the first network by Assumption 7. We need to prove this parameter δ_2 is uniquely defined. Assume that there exists a permutation ϕ on $\{1, \dots, M\}$ such that a similar identification occurs. Then we will observe $\delta_2' = \tilde{\alpha}_{\phi(1)\phi(1)}^2/\alpha_{11} = \tilde{\alpha}_{\phi(2)\phi(2)}^2/\alpha_{22}$ with ϕ such that $\phi(1) \neq 1$ or $\phi(2) \neq 2$, which

implies that $\widetilde{\alpha}_{11}/\widetilde{\alpha}_{\phi(1)\phi(1)} = \widetilde{\alpha}_{22}/\widetilde{\alpha}_{\phi(2)\phi(2)}$. This is in contradiction with Assumption 6. Therefore, $\delta_2 = \delta'_2$. We can then identify the blocks in network $m = 2$ by matching $\widetilde{\alpha}_{qq}/\delta$ with the α_{qq} already identified. The ones that does not match complete the matrix $\boldsymbol{\alpha}$. The process is iterated with networks $m = 3, \ldots, M$. Once the matrix $\boldsymbol{\alpha}$ and the parameters in $\boldsymbol{\delta}$ are identified, injections from $\{1, \ldots, Q_m\} \to \{1, \ldots, Q\}$ corresponding to the matching of the blocks provide the $\boldsymbol{\pi}^m$.

□

3.B. Details of the Model Selection when Allowing for Empty Blocks

For π-*colSBM* and $\delta\pi$-*colSBM*, the model is decribed by its support S. We can compute the likelihood for a given support. We recall that $\boldsymbol{\theta}_S = \{\boldsymbol{\alpha}_S, \boldsymbol{\delta}, \boldsymbol{\pi}_S\}$ are the parameters restricted to their support. Then for the model represented by S, the complete likelihood is given by:

$$\begin{aligned} p(\mathbf{X}, \mathbf{Z}|S) &= \int_{\boldsymbol{\theta}_S} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_S, S) p(\boldsymbol{\theta}_S) \mathrm{d}(\boldsymbol{\theta}_S) \\ &= \int_{(\boldsymbol{\alpha}_S, \boldsymbol{\delta})} \int_{\boldsymbol{\pi}_S} p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}_S, \boldsymbol{\delta}, S) p(\mathbf{Z}|\boldsymbol{\pi}_S, S) p(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) p(\boldsymbol{\pi}_S) \mathrm{d}(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) \mathrm{d}(\boldsymbol{\pi}_S) \\ &= \underbrace{\int_{(\boldsymbol{\alpha}_S, \boldsymbol{\delta})} p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\alpha}_S, \boldsymbol{\delta}, S) p(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) \mathrm{d}(\boldsymbol{\alpha}_S, \boldsymbol{\delta})}_{B1} \underbrace{\int_{\boldsymbol{\pi}_S} p(\mathbf{Z}|\boldsymbol{\pi}_S, S) p(\boldsymbol{\pi}_S) \mathrm{d}(\boldsymbol{\pi}_S)}_{B2}, \end{aligned} \tag{3.23}$$

where we assumed the prior on the emission parameters and the mixture parameters to be independent.

The restriction of the parameter space to the one associated with the support S is needed. Otherwise, some parameters would not be defined or would lie on the boundary of the parameters space, and the following asymptotic derivation would not be properly defined. We use a BIC approximation on $B1$ where we rewrite:

$$\begin{aligned} p(\mathbf{X}|\mathbf{Z}, S) &= \int_{(\boldsymbol{\alpha}_S, \boldsymbol{\delta})} \left(\prod_{m=1}^{M} p(X^m|Z^m, \boldsymbol{\alpha}_S, \boldsymbol{\delta}, S) \right) p(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) \mathrm{d}(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) \\ &= \max_{(\boldsymbol{\alpha}_S, \boldsymbol{\delta})} \exp \left(\sum_{m=1}^{M} \ell(X^m|Z^m; \boldsymbol{\alpha}_S, \boldsymbol{\delta}, S) - \frac{1}{2} (\nu(\boldsymbol{\alpha}_S, \boldsymbol{\delta})) \log(\sum_m n_m(n_m - 1)) + \mathcal{O}(1) \right), \end{aligned}$$

where $\nu(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) = \nu(\boldsymbol{\alpha}_S) = \sum_{q,r=1}^{Q} \mathbf{1}_{(S'S)_{qr}>0}$ (number of free parameters in $\boldsymbol{\alpha}_S$) in a π-*colSBM*, $\nu(\boldsymbol{\alpha}_S, \boldsymbol{\delta}) = \nu(\boldsymbol{\alpha}_S) + M - 1$ in a $\delta\pi$-*colSBM*.

For $B2$, we use a Q_m-dimensional Dirichlet prior for each π_S^m:

$$\begin{aligned} p(\mathbf{Z}|S) &= \prod_{m\in\mathcal{M}} \int_{\pi_S^m} p(Z^m|\pi_S^m)p(\pi_S^m)\mathrm{d}(\pi_S^m) \\ &= \max_{\boldsymbol{\pi}_S} \exp\left(\sum_{m=1}^{M} \ell(Z^m;\pi_S^m) - \frac{Q_m-1}{2}\log(n_m) + \mathcal{O}(1)\right). \end{aligned}$$

Then, we input $B1$ and $B2$ into Equation (3.23) to obtain:

$$\log p(\mathbf{X},\mathbf{Z}|S) \approx \max_{\boldsymbol{\theta}_S} \ell(\mathbf{X},\mathbf{Z};\boldsymbol{\theta}|S) - \frac{1}{2}\Big(\mathrm{pen}_\pi(Q,S) + \mathrm{pen}_\alpha(Q,S) + \mathrm{pen}_\delta(Q,S)\Big), \tag{3.24}$$

with the penalty terms as given in the main text.

As $\mathbf{Z}$ is unknown we replace each Z_{iq}^m by the variational parameters $\hat{\tau}_{iq}^m$ which maximizes the variational bound for a given support S. Then, we add the entropy of the variational distribution $\mathcal{H}(\widehat{\mathcal{R}}(\mathbf{Z}))$ to Equation (3.24). This leads to the variational bound of Equation (3.9), as

$$\max_{\boldsymbol{\theta}_S} \mathcal{J}(\hat{\boldsymbol{\tau}},\boldsymbol{\theta}_S) = \max_{\boldsymbol{\theta}_S} \ell(\mathbf{X},\mathbb{E}_{\widehat{\mathcal{R}}}[\mathbf{Z}];\boldsymbol{\theta}_S) + \mathcal{H}(\widehat{\mathcal{R}}(\mathbf{Z})),$$

which we recall is a surrogate of the log-likelihood of the observed data. We define

$$\text{BIC-L}(\mathbf{X},Q,S) = \max_{\boldsymbol{\theta}_S} \mathcal{J}(\hat{\boldsymbol{\tau}},\boldsymbol{\theta}|S) - \frac{1}{2}\Big(\mathrm{pen}_\pi(Q,S) + \mathrm{pen}_\alpha(Q,S) + \mathrm{pen}_\delta(Q,S)\Big)$$

which is a penalized likelihood criterion when the support S is known. Finally to obtain the criterion BIC-L$(\mathbf{X},Q)$ for π-*colSBM* and $\delta\pi$-*colSBM*, we need to penalize for the size of the space of possible models that depends on the support S. For a given Q corresponding to the number of different blocks in the collection of network $\mathbf{X}$, we set the prior on S decomposed as the product of uniform priors on the numbers of blocks (between 1 and Q) actually represented in each network and uniform priors for the choice of these Q_m blocks among the Q possible blocks (Q_m is the number of blocks that are represented in network m):

$$p_Q(S) = p_Q(Q_1,\ldots,Q_M)\cdot p_Q(S|Q_1,\ldots,Q_M) = \frac{1}{Q^M}\cdot\prod_{m=1}^{M} 1\Big/\binom{Q}{Q_m}.$$

The prior is given on the space $\mathcal{S}_Q$ of admissible support.

Using a BIC approximation under a concentration assumption on the correct support, we derive

$$\begin{aligned} \log p(\mathbf{X},\mathbf{Z}|\mathbf{Q}) &= \log\int_S p(\mathbf{X},\mathbf{Z}|S,\mathbf{Q})p_Q(S)\mathrm{d}S \\ &\approx \log\int_S \exp\Big(\text{BIC-L}(\mathbf{X},Q,S)p_Q(S)\Big)\mathrm{d}S \\ &\approx \max_{S\in\mathcal{S}_Q}\Big(\text{BIC-L}(\mathbf{X},Q,S) - \log p_Q(S)\Big). \end{aligned}$$

Thus, by denoting $\mathrm{pen}_S(Q) = -2\log p_Q(S)$ in the equation above, we obtain the expression of BIC-L$(\mathbf{X},Q)$ given in Equation (3.14).

3.C. Partition of Food Webs with $\delta colSBM$

We present the partition of networks and the connectivity structure of each group obtained by fitting a $\delta colSBM$ with Poisson distribution and using the binary tree partition method in Figure 3.7 as well as a contingency table of the obtained group with the different dataset of the Mangal database. There are 3 groups, with group A, B being closer in terms of connectivity structure than group C. The blocks are ordered to minimize the upper diagonal interactions. Comments on each groups follow:

A A group of 8 networks, 5 being networks with 78 species issued from the same database (id 58, stream food webs) and the 3 others being smaller networks. The structure is composed of 11 blocks, the top trophic level does not prey on the bottom 5 ones. Most of the blocks have within blocks interactions.

B A group of 28 networks of various sizes issued from a lot of different databases. The connectivity matrix has a kind of L-shape. 2 blocks, containing about a quarter of the species, are preyed on by the 4 other blocks. The top two blocks feed on all the other blocks, differing only in intensity.

C A group of 31 networks with a more "stair"-shape structure. The top trophic blocks feed on medium's ones with high intensity but not so much on the lower ones. 10 of these networks are New Zealand stream food webs issued from the same database (id 157).

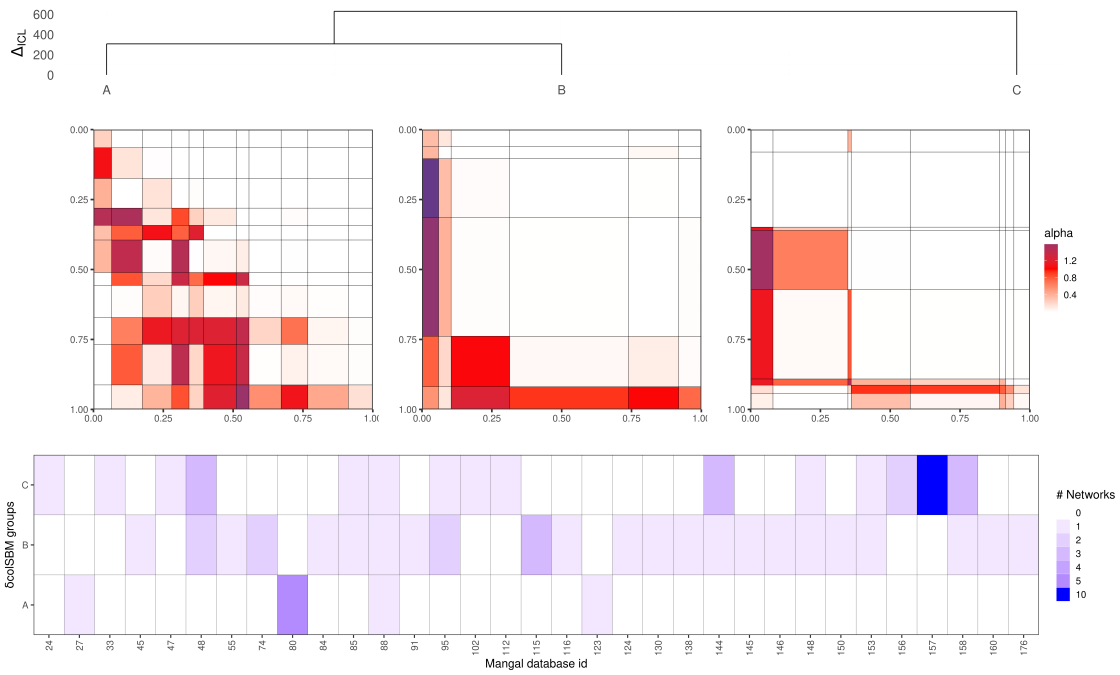


Figure 3.7 – Above: Classification and connectivity structures of a collection of 67 predation networks from the Mangal database into 3 groups. The length of the dendrogram is given by the difference in BIC-L to the best model. Below: Contingency table of the classification found by $\delta colSBM$ and the different datasets from the Mangal database.

3.D. Analyze of advice networks

3.D.1. Presentation of the advice networks

In this section, we consider 4 advice networks with binary and directed interactions. The individuals involved are judges (Lazega et al., 2011), lawyers (Lazega, 2001), priests (Lazega and Wattebled, 2011) and researchers(Lazega et al., 2008). Using this collection of networks, Lazega and Brailly (2021) analyze and compare the homophily of the different types of actors. The 4 networks are heterogeneous in size and density, with the lawyers advice network having between 2 to 3.5 times the density of the other networks:

$$n = (153, 71, 104, 126) \quad \text{and} \quad \hat{d} = (.053, .179, .049, .061).$$

In order to make comparison with the different structures, we introduce the following indices which can be computed directly from the SBM parameters for $\mathcal{F}$ a Bernoulli distribution:

- Modularity: The propensity of the nodes to interact with nodes from the same

block. The higher the more modular:

$$\mathrm{Mod}(\mathcal{F}\text{-SBM}_n(Q,\alpha,\pi)) = \sum_{q=1}^{Q}(a_{qq} - (\sum_{r=1}^{Q} a_{qr})(\sum_{r=1}^{Q} a_{rq})),$$

where $a_{qr} = (\frac{\pi_q \alpha_{qr} \pi_r}{\sum_{q,r} \pi\alpha\pi^\intercal})$ is the the proportion of total interactions being between blocks q and r.

- Reciprocity: The symmetry in the block structure. How close are α_{qr} from α_{rq}. The higher the more reciprocal the network:

$$\mathrm{Rec}(\mathcal{F}\text{-SBM}_n(Q,\alpha,\pi)) = 1 - \frac{1}{2}\sum_{q,r=1}^{Q} |a_{qr} - a_{rq}| \quad \in [0,1].$$

- Nestedness: High expected degree nodes connect to high expected degree nodes. A smooth version of core-periphery. The higher the most nested. 1 corresponds to an Expected Degree Distribution (EDD) model.

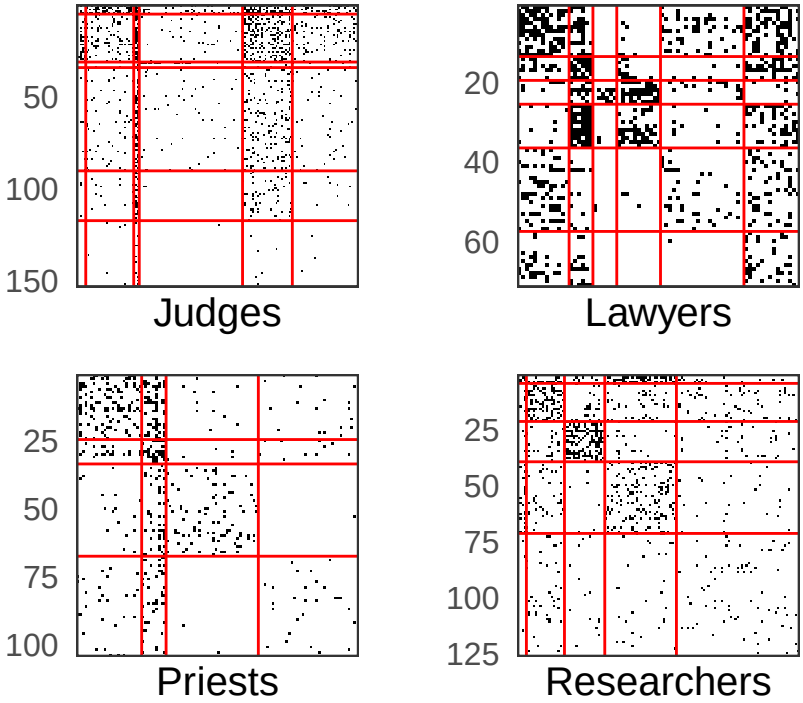
$$\mathrm{Nes}(\mathcal{F}\text{-SBM}_n(Q,\alpha,\pi)) = 1 - \frac{1}{2}\sum_{q,r=1}^{Q} |a_{qr} - b_{qr}| \quad \in [0,1],$$

where $b_{qr} = \frac{(\sum_{q=1}^{Q} \pi_q \alpha_{qr})_r (\sum_{r=1}^{Q} \alpha_{qr} \pi_r)_q}{\sum_{q,r=1}^{Q} (\sum_{q'=1}^{Q} \pi_{q'} \alpha_{q'r})_r (\sum_{r'=1}^{Q} \alpha_{qr'} \pi_{r'})_q}$ is the the proportion of total interactions being between blocks q and r for an EDD model.

We fit a SBM on each of the 4 networks. The adjacency matrices reordered by blocks are depicted in the next pictures as well as a table with the 3 indices described above in Table 3.4. A short analysis follows:

- Judges: A structure with 6 blocks of heterogeneous size, with a clear asymmetry in the block connections (low reciprocity index of .371). In particular, block 5 has much higher average indegree than outdegree. The within blocks connections are not very strong, implying a low modularity. It is the most nested network structure of the 4 advice networks.

- Lawyers: A structure with 5 blocks, which the first 3 have very strong within block connections, however this is not the case for the last 2 blocks, leading to a medium modularity. Block 5 is a highly asymmetric block.
- Priests: A structure with 4 blocks with a core of 2 blocks and a periphery with assortative community structures on the 2 other blocks. The structure has high modularity but average nestedness and reciprocity compared to the other networks.
- Researchers: A structure with 5 blocks, with a kind of assortative community structure for the first 4 blocks. Block 5 consists in the residual individuals with homogeneous connection pattern with the rest of the researchers. The larger the block, the sparser the connection. The structure is the most reciprocal, modular and the least nested of the 4 networks.



	Judges	Lawyers	Priests	Researchers
Modularity	0.014	0.132	0.256	0.277
Nestedness	0.877	0.749	0.725	0.671
Reciprocity	0.371	0.624	0.757	0.877

Table 3.4 – Indices for the structures obtained by SBM.

3.D.2. Pairwise analysis of the advice networks

In this section, we analyze for each pair of networks if they share common structures. We fit each colSBM on all pairs of networks and compared the obtained BIC-L criterion to the one of two independent SBMs. The results are shown in Table 3.5

Model	Jud_Law	Jud_Pri	Jud_Res	Law_Pri	Law_Res	Pri_Res
sep-SBM	-6084	-5913	-7278	-3932	-5297	-5126
iid-colSBM	-6144	-5945	-7334	-3956	-5287	-5091
π-*colSBM*	-6130	-5914	-7298	-3950	-5276	-5099
δ-*colSBM*	-6172	-5950	-7331	-3940	-5289	-5095
$\delta\pi$-*colSBM*	-6124	-5920	-7308	-3930	-5299	-5106

Table 3.5 – BIC-L criterion for pairs of advice networks.

The model selection criterion states that when considering networks two by two, the judges advice network does not share common structure with the other networks. For lawyers and priests, the common structure is relevant only for $\delta\pi$-*colSBM*, this

is caused by the huge difference in density between the two networks forcing some non-common clusters. Priests and researchers are found to have some common structures on all the models, the best fit being for *iid-colSBM*, while *π-colSBM* provides the best model for lawyers and researchers.

On collections where we improved the model selection criterion from separated SBM, we plot the structure obtained by the colSBM with the highest BIC-L. We comment on the change in the clustering corresponding to the old separated SBMs structure and the new colSBM structure.

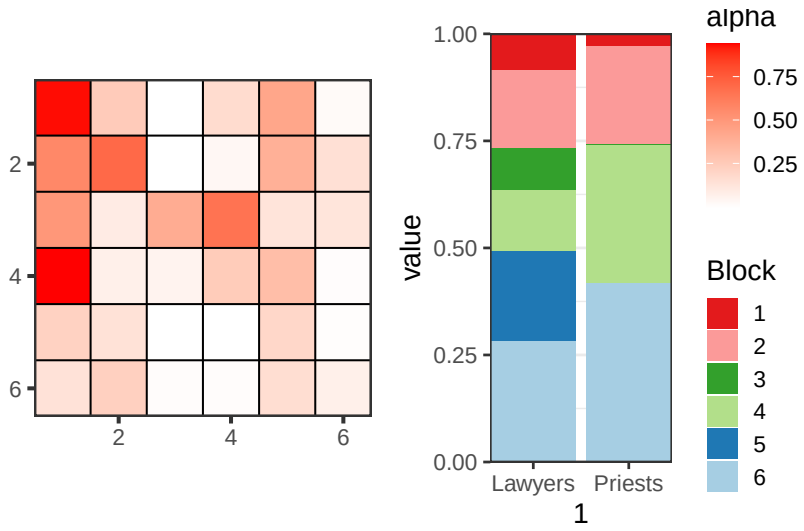
Lawyers and priests

For those two networks with highly different density, only the *δπ-colSBM* model improved on the separated SBMs. The lawyers advice network has a structure with 6 blocks which embeds the structure of the priests advice network. The obtained ARI between the separated SBMs and the *δπ-colSBM* is given by .93 for the lawyers and .72 for the priests. Saying it otherwise, the mesoscale structure of the priests network was slightly modified to agree with a partial structure of the lawyers network. The new structure for the priests network is slightly less modular and more reciprocal than the one obtained by the SBM. The changes for the lawyers network are not significant.

We show in Table 3.6, the indices obtained from the *δπ-colSBM* model parameters and the one given by computing $\tilde{\alpha}$ and $\tilde{\pi}$ from the obtained variational parameters of this network.

	Lawyers	Lawyers (colSBM)	Priests	Priests (colSBM)
Modularity	0.127	0.121	0.216	0.224
Nestedness	0.747	0.760	0.725	0.693
Reciprocity	0.618	0.596	0.788	0.851

Table 3.6 – Indices for the structures obtained by *δπ-colSBM* for the lawyers and the priests networks.



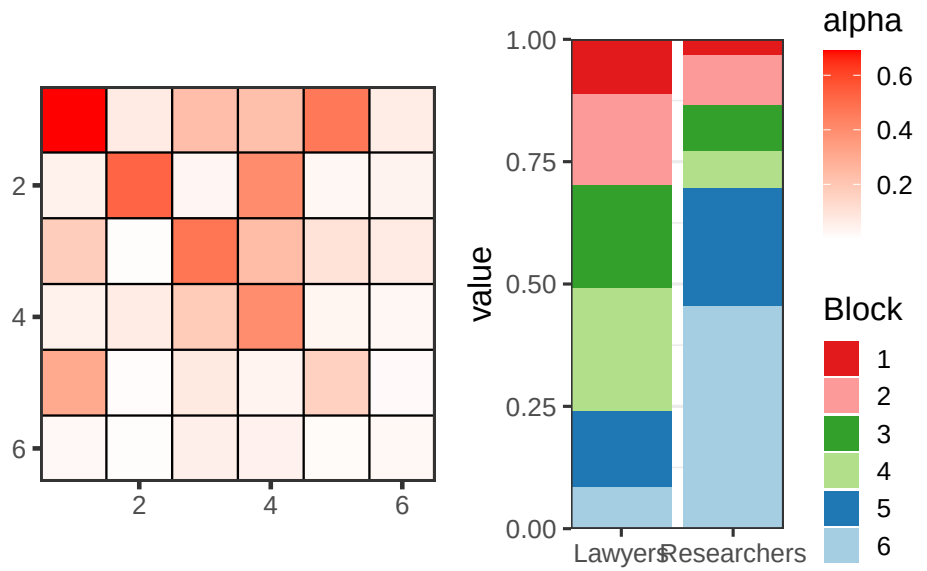
Lawyers and researchers

On these two networks all of the 4 colSBMs improved over separated SBMs. The model selection criterion for π-$colSBM$ is the highest by a large margin, we examine the results further. Individuals of both networks populate the 6 blocks but with large differences in terms of block proportions. For the lawyers, the proportions are homogeneous, while the researchers are mostly in blocks 5 and 6 (blue blocks), which are the least connected ones. The blocks membership for the researchers is only slightly modified compared to the one obtained by the SBM (ARI = .88), but the memberships for the lawyers has a weak agreement with the one obtained from the SBM (ARI = .27).

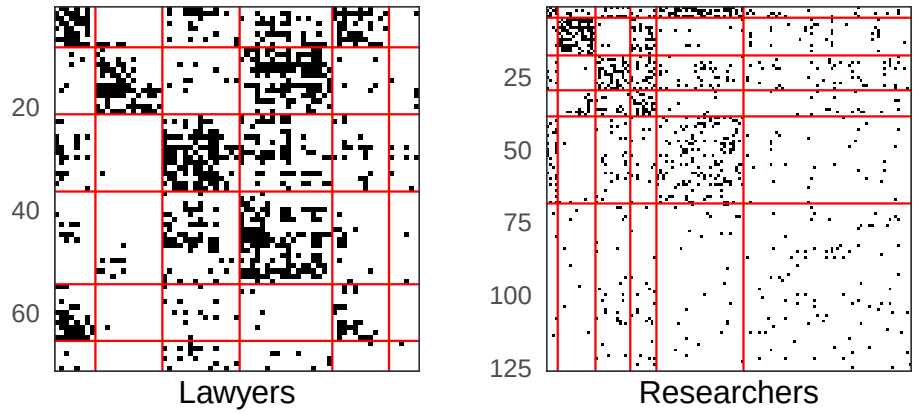
The new structure of the lawyers network is more modular, more reciprocal and slightly less nested. As for the structure of the researchers network, it became slightly less reciprocal and modular. This exhibits the compromise in the structure found by the π-$colSBM$. The indices are shown in Table 3.7.

	Lawyers	Researchers	Lawyers (colSBM)	Researchers (colSBM)
Modularity	0.222	0.257	0.233	0.245
Nestedness	0.681	0.673	0.694	0.670
Reciprocity	0.796	0.828	0.815	0.848

Table 3.7 – Indices for the structures obtained by π-$colSBM$ for the lawyers and the researchers networks.



In order to understand the compromise in the structure done for the two networks, we plot the adjacency matrices reordered by blocks. The mesoscale structure of the researchers network did not change much, being mainly an assortative community structure with decreasing within block connectivity and a very small block more connected to the individuals from other blocks. The memberships of the lawyers were reordered in order for its structure to be more assortative and hence to match the one of the researchers network.

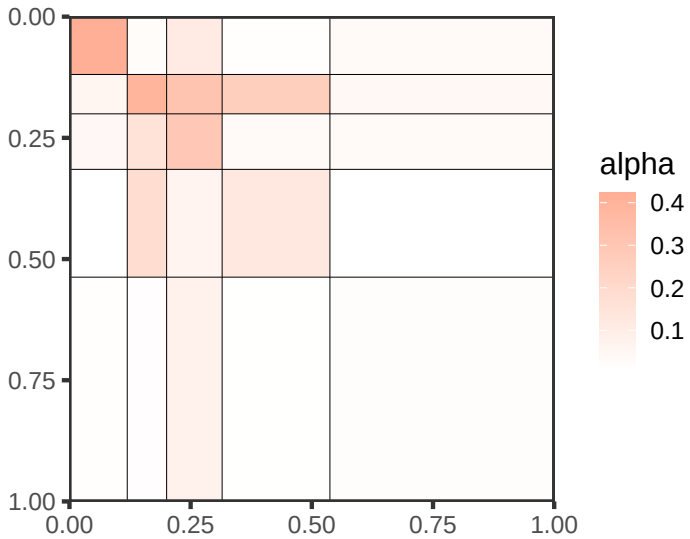


Priests and researchers

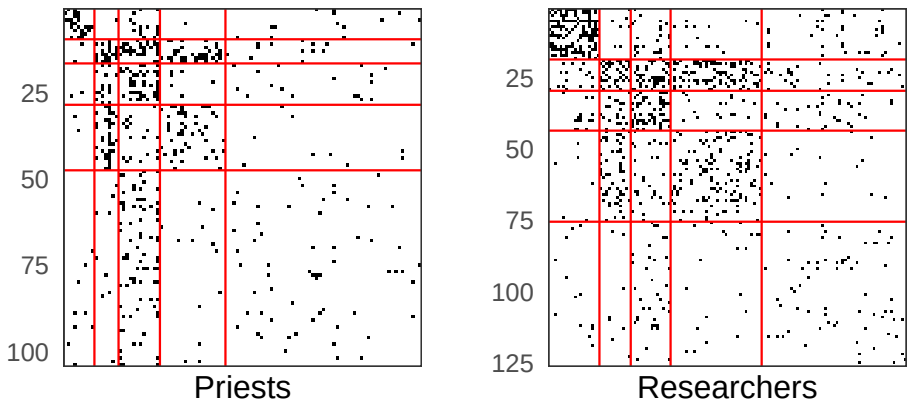
On these two networks all of the 4 colSBMs improved over separated SBMs. The structures of this pair of networks are the most similar of the collection and the *iid-colSBM* model has the highest model selection criterion. The new structure is shown in the following matricial view of the colSBM. For the researchers, the clustering agree with the one obtained for the separated SBM ($ARI = .82$), while this is not the case for the priests ($ARI = .31$. However the number of blocks differs in this case. The new structure finds a compromise in the reciprocity and nestedness of both networks. But, by doing so, the structures are less modular than the ones obtained with the SBM. The indices are shown in Table 3.8.

	Priests	Researchers	Priests (colSBM)	Researchers (colSBM)
Modularity	0.187	0.249	0.225	0.225
Nestedness	0.710	0.683	0.693	0.693
Reciprocity	0.789	0.845	0.824	0.824

Table 3.8 – Indices for the structures obtained by *iid-colSBM* for the researchers and the priests networks.



We plot the adjacency matrices reordered by blocks for both of the networks. The two structures are very similar the main differences being the residual blocks being much larger for the priest network and the first block being denser for the researcher network.



3.D.3. Looking for larger collections

When considering 3 networks at a time, all of the colSBMs are able to find common structures for the lawyers–priests–researchers collection. This is further evidence that the judges advice networks has a different structure than the 3 other networks, as no collections involving the judges has a higher BIC-L than separated SBMs. The model selection criteria for collections of 3 and 4 networks are shown in Table 3.9.

Model	Jud_Law_Pri	Jud_Law_Res	Jud_Pri_Res	Law_Pri_Res	Jud_Law_Pri_Res
sep-SBM	-7964	-9329	-9158	-7177	-11210
iid-colSBM	-8032	-9397	-9187	-7128	-11254
π-colSBM	-7988	-9377	-9176	-7130	-11213
δ-colSBM	-8034	-9406	-9169	-7138	-11256
$\delta\pi$-colSBM	-7983	-9377	-9168	-7137	-11231

Table 3.9 – BIC-L criterion for collections of 3 and 4 advice networks.

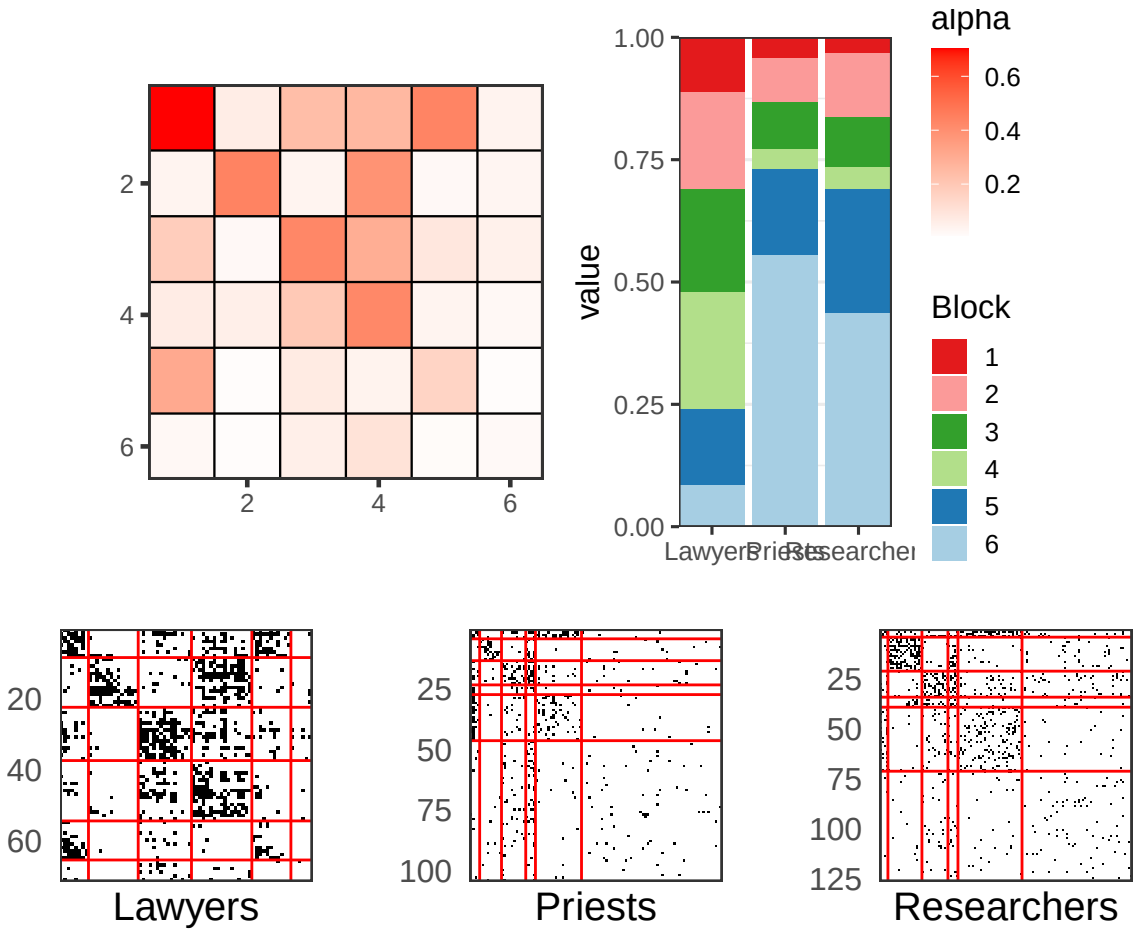
Lawyers, priests and researchers

We plot the *π-colSBM* for the lawyers–priests-researchers collection of networks, as well as individual adjacency matrices with the newly found clustering. The indices of the obtained structures are given in Table 3.11. We obtain a structure of 6 blocks with block proportions being different for the lawyers compared to the one of the priests and judges. Compared to the one obtained from separated SBMs, the structure for the lawyers is less nested, but much more reciprocal and modular; the structure for the priests is less modular, and slightly more reciprocal, while the structure for the researchers did not change much.

	Lawyers	Priests	Researchers
Modularity	0.228	0.186	0.261
Nestedness	0.679	0.705	0.667
Reciprocity	0.797	0.786	0.838

	Lawyers (colSBM)	Priests (colSBM)	Researchers (colSBM)
Modularity	0.212	0.211	0.254
Nestedness	0.714	0.703	0.656
Reciprocity	0.805	0.855	0.860

Table 3.11 – Indices of the structures obtained by *π-colSBM* for the lawyers, priests and researchers advice networks.



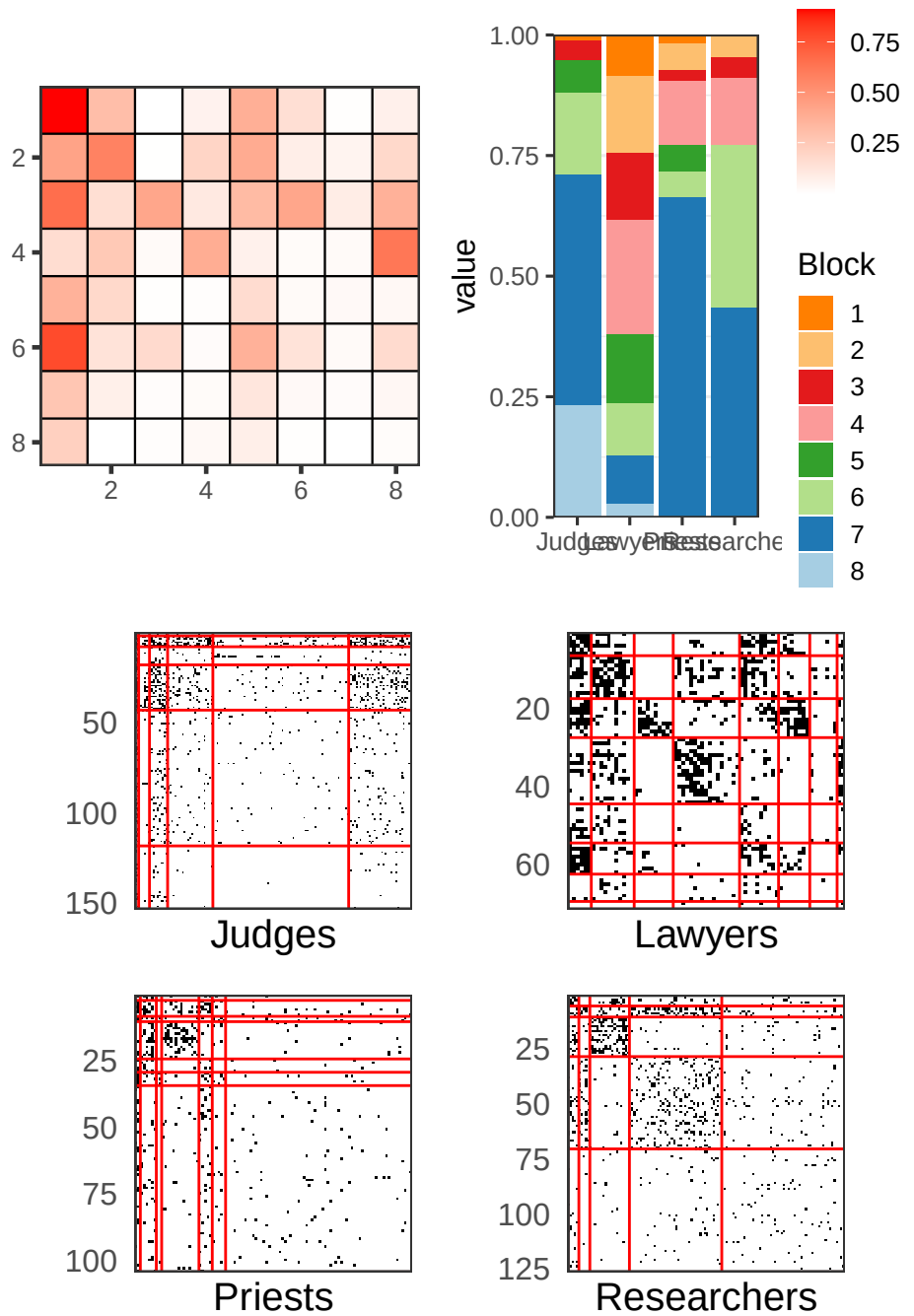
Judges, lawyers, priests and researchers

For the whole collection of advice networks, we explain the structure found by π-$colSBM$, the model selection criterion being very close to the one of separated SBMs. We find a 8 blocks structure, with empty blocks for each network. The judges are present on 6 blocks. When looking at the indices for the structure in Table 3.13, the indices for the judges did not change much, but the structure for the researchers, and the priests to a lesser extent, is less modular and more nested than before. The structure of the researchers network had to change in order to find a good compromise with the other networks, which was not the case for collections with 2 or 3 networks without the judges advice network. Structure of the π-$colSBM$ and the matricial view of the 4 networks reordered by blocks are plotted below.

	Judges	Lawyers	Priests	Researchers
Modularity	0.010	0.178	0.133	0.231
Nestedness	0.882	0.695	0.773	0.715
Reciprocity	0.392	0.672	0.725	0.865

	Judges (colSBM)	Lawyers (colSBM)	Priests (colSBM)	Researchers (colSBM)
Modularity	0.039	0.159	0.141	0.196
Nestedness	0.870	0.714	0.775	0.737
Reciprocity	0.447	0.662	0.728	0.868

Table 3.13 – Indices of the structure obtained by π-*colSBM* for the judges, lawyers, priests and researchers advice networks.



Clustering networks and choosing the best model to explain the collection

We look for each colSBM for the partition of networks which leads to the best BIC-L. The 4 models agree on making a group with lawyers, priests and researchers advice netwroks, which the structure has already been analyzed for π-$colSBM$. The judges advice network is on a group of its own with the 6 blocks structure obtained by the SBM.

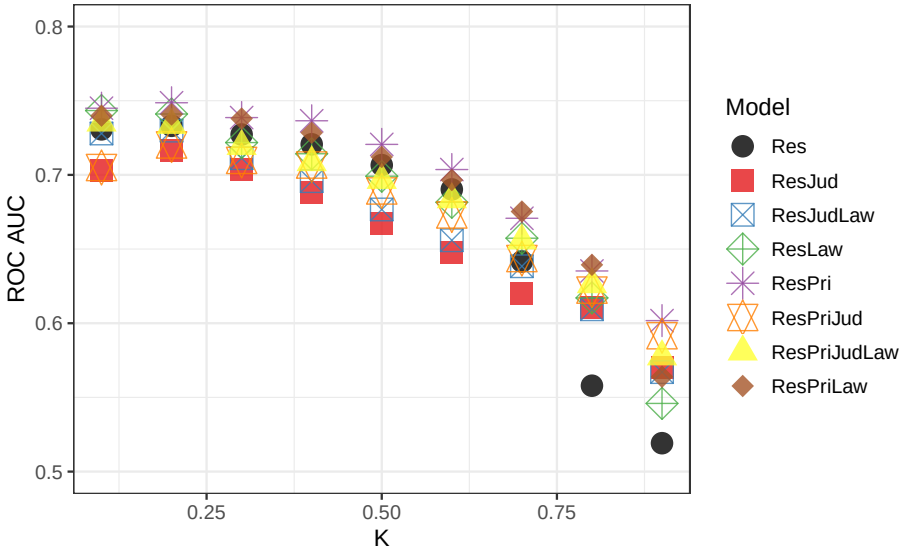
3.D.4. Using dyad prediction to quantify the link between networks

As we have seen before, some advice networks share common connectivity structures with other advice networks, but not all. To further illustrate this fact, we will try to predict the value of missing dyads between the individuals of a given network, using the information from the other networks. Recall that these networks do not share any individuals and that any improvements in the prediction are due to similar connection patterns between the networks in the collection.

We will plot the ROC AUC of the prediction obtained from an SBM on the researchers and the judges network and the one with δ-$colSBM$ on various collections of networks. We make the number of missing dyads vary between 10% and 80% (K) of the total number of dyads and simulate 30 times the missing dyads for each value of K. A value of .5 means that the prediction is not better than a uniform one on all missing dyads. The baseline will be the black dots on the figures, which is the prediction obtained by an SBM on a single network (judges or researchers). Information obtained from other networks of the collection improve the prediction of interactions between researchers or judges if it has a higher ROC AUC than the one obtained from the SBM.

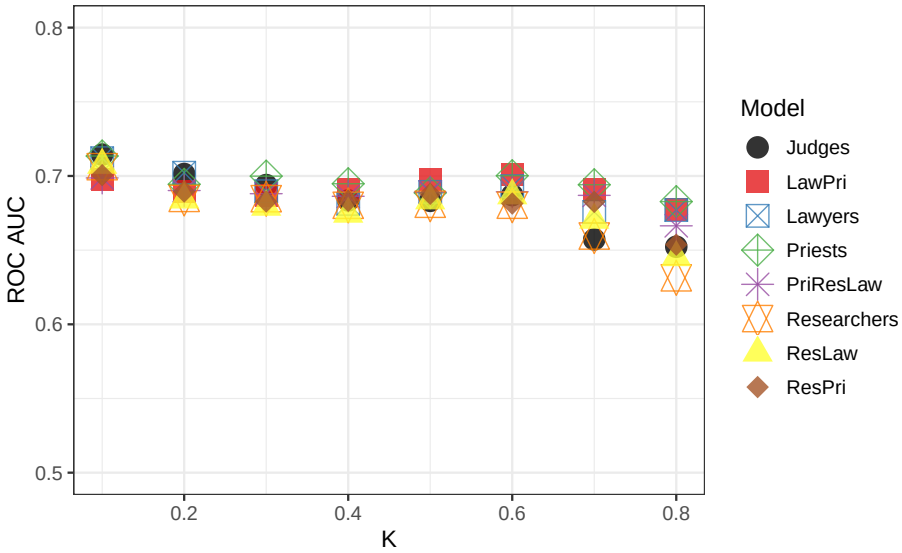
Predicting interactions between Researchers

colSBM found common structure between the researchers network and both the priests and the lawyers networks. For small and medium K, collections involving the judges provide poor prediction while the ones involving priests improve the prediction of advice between researchers. When K is higher than .8, the information obtained from the judge network significantly improves significantly the prediction of missing dyads, as almost no structure remains in the researchers network.



Predicting interactions between judges

Improving the prediction of advice between judges with the information coming from other networks is harder than for the researcher when K is small. But the capacity to predict stays good when K becomes larger. Information from priests and lawyers–priest advice networks improve the predictions, while the information coming from the researchers are just noise, providing worst prediction than considering the judges on their own.



3.D.5. Conclusion

We show on a collection of advice networks how considering a joint model on the networks leads to changes on the structure explaining the connection in the networks and how information contained in other networks might help the prediction of the

interactions on a given network.

Among all our experiments the advice networks from the judges exhibit different connection patterns than the 3 other networks, but common connectivity patterns still exist as shown in the prediction section. Priests and researchers advice networks are the most similar. About the lawyers advice network, which is a lot denser than the other ones, it is possible to find another mesoscale structure to explain the connectivity patterns of the network. The new connectivity structure is a good compromise with the one of the researchers and priests advice networks.

Advice networks share common connectivity patterns, which could be grossly described as an assortative community structure with various within block interaction probability, and an asymmetry in the connections of the most connected blocks. For the judges advice network, the assortativity is weaker and the reciprocity in the connection between blocks is much lower.

Chapter 4

Estimating the robustness of a bipartite ecological networks through a probabilistic modeling

Motivation L'idée de ce travail est née des discussions avec les écologues Sonia Kéfi et François Massol dans le cadre de l'ANR Econet. L'objectif est de mieux comprendre le lien entre la structure du réseau et la robustesse et de proposer une méthode permettant de comprendre comment la robustesse des réseaux écologiques varie selon le type d'interaction concerné.

Résumé La robustesse d'un réseau écologique quantifie la résilience de l'écosystème qu'il représente à la perte d'espèces. Elle correspond à la proportion d'espèces qui sont déconnectées du reste du réseau lorsque des extinctions se produisent de manière séquentielle. Classiquement, la robustesse est calculée pour un réseau donné, à partir de la simulation d'un grand nombre de séquences d'extinction. Le lien entre la structure du réseau et la robustesse reste une question ouverte. La définition d'un modèle probabiliste conjointement sur le réseau et les séquences d'extinction permet d'analyser cette relation. Les modèles de blocs stochastiques bipartites ont prouvé leur capacité à modéliser les réseaux bipartites, par exemple dans les réseaux plantes-pollinisateurs : les espèces sont divisées en blocs et les probabilités d'interaction sont déterminées par les blocs d'appartenance. Des expressions analytiques de l'espérance et de la variance de la robustesse sont obtenues sous ce modèle, pour différentes distributions de séquences d'extinction primaire. L'impact de la structure du réseau sur la robustesse est analysé à travers un ensemble de propriétés et d'illustrations numériques. L'analyse d'une collection de réseaux écologiques bipartites nous permet de comparer l'approche empirique à notre approche probabiliste, et illustre la pertinence de cette dernière lorsqu'il s'agit de calculer la robustesse d'un réseau partiellement observé ou incomplètement échantillonné.

Diffusion Le contenu de ce chapitre ainsi que la première annexe ont fait l'objet d'un article (Chabert-Liddell et al., 2022) publié dans le journal Environmetrics. Un package `R` nommé `robber`, disponible sur le CRAN (Chabert-Liddell, 2021b), dont la documentation est disponible à l'adresse suivante https://Chabert-Liddell.github.io/robber/, permet d'appliquer les méthodes développées ci-dessous. Un exemple d'application du package à la comparaison de la robustesse d'une collection de réseaux d'interactions écologiques bipartites est donné en appendice 4.B.

Notations for this chapter

n_r The number of row species

n_c The number of column species

A An incident matrix of size $n_r \times n_c$

Z The block membership of row species

W The block membership of column species

Q_r The number of row blocks

Q_c The number of column blocks

δ The connectivity parameter, a matrix of size $Q_r \times Q_c$

π The mixture parameter of row species

ρ The mixture parameter of column species

$\boldsymbol{\theta}$ The set of parameters (δ, π, ρ)

R The robustness function of the network

$\bar{R}$ The robustness statistic of the network

$\mathbb{A}$ A probability distribution of networks

$\mathbb{S}$ A probability distribution of primary extinction sequences

$\mathbb{U}$ A uniform probability distribution of primary extinction sequences

$\mathbb{B}$ A probability distribution of primary extinction sequences by block memberships ($\mathbb{B}^{\uparrow}$ increasing, $\mathbb{B}^{\uparrow}$ decreasing)

$\mathbb{D}$ A probability distribution of primary extinction sequences which depends on the sequence of the degrees ($\mathbb{D}^{\uparrow}$ increasing, $\mathbb{D}^{\downarrow}$ decreasing)

$\mathbb{D}^{\uparrow}_{ord}$ A probability distribution of primary extinction sequences determined by the increasing ($\mathbb{D}^{\downarrow}_{ord}$, decreasing) sequence of the degrees

$\mathbb{D}^{\uparrow}_{lin}$ A probability distribution of primary extinction sequences which depends linearly on the increasing ($\mathbb{D}^{\downarrow}_{lin}$, decreasing) sequence of the degrees

Abstract The robustness of an ecological network quantifies the resilience of the ecosystem it represents to species loss. It corresponds to the proportion of species that are disconnected from the rest of the network when extinctions occur sequentially. Classically, the robustness is calculated for a given network, from the simulation of a large number of extinction sequences. The link between network structure and robustness remains an open question. Setting a joint probabilistic model on the network and the extinction sequences allows analysis of this relation. Bipartite stochastic block models have proven their ability to model bipartite networks e.g. plant-pollinator networks: species are divided into blocks and interaction probabilities are determined by the blocks of membership. Analytical expressions of the expectation and variance of robustness are obtained under this model, for different distributions of primary extinction sequences. The impact of the network structure on the robustness is analyzed through a set of properties and numerical illustrations. The analysis of a collection of bipartite ecological networks allows us to compare the empirical approach to our probabilistic approach, and illustrates the relevance of the latter when it comes to computing the robustness of a partially observed or incompletely sampled network.

4.1. Introduction

In response to the rapid evolution of ecosystems due to climate change (habitat losses and species extinctions) on the one hand, and the increasing number of available data sets of ecological interaction networks on the other hand, the study of the robustness of ecological networks to species loss has become an active area of research in the ecological scientific community (see Landi et al., 2018, for a review). Given a primary extinction sequence, the influence of these extinctions on the other species is studied by monitoring the proportion of species that get disconnected from the rest of the network (Dunne et al., 2002).

Bipartite networks are used to represent interactions between two separated sets of species that do not interact within their own set. These interactions may be either mutualistic when the species of both groups benefit from the interactions (such as pollination for plant-pollinator networks or seed-dispersal for plant-frugivore bird networks) or antagonistic when one group of species benefits from the interactions at the expense of the other group of species (e.g. plant-herbivore or host-parasite networks). Robustness is a numeric indicator quantifying the impact of the disappearance of one set of (primary) species on the other (secondary species) under given extinction sequences (see Memmott et al., 2004; Curtsdotter et al., 2011, for examples of primary extinction sequences). In a few words, the object of interest is the proportion of secondary species that remain connected to other species after a given number of primary species have disappeared. Since this proportion depends on the order in which the species disappear, robustness is defined as the average of these proportions over a large number of primary extinction sequences. [1]

[1]Note that the robustness of ecological networks bears a different meaning than the robustness

The definition of ecological robustness used in this paper is not the only one, in particular the robustness of ecological networks is also studied through dynamic approaches. Among recent developments in the dynamic approach, Song et al. (2018) are interested in the feasibility domain of ecological communities, i.e. the set of environmental conditions under which all species have positive abundances. Barabás et al. (2014) estimate extinction risk to small environmental variation through the use of sensitivity analysis while others study the capacity of coexistence of species through the so-called invasion criterion (see Grainger et al., 2019, for a review).

Obviously, the robustness defined above is strongly related to the size of the network considered (i.e. the number of species involved or *species richness*) and to its density (i.e. the number of interactions observed compared to the total number of possible interactions or *connectance*). To go further, the question arises to what extent the topological properties of a network also influence its robustness. Indeed, observed ecological networks present different topological structures depending on the type of interactions: for example, mutualistic networks are known to have a strong nested structure, while this is not necessarily the case for antagonistic networks (Fortuna et al., 2010; Bascompte et al., 2003). A key question is to identify the relationship between the ability of a network to withstand species extinctions and its characteristics such as its size and its mesoscale structure.

To that purpose, we propose to assume a parametric probabilistic model for the bipartite network. This model will embed the topological properties of the network in a few number of parameters. Then we propose to study the behavior of the robustness in this framework and in particular the variations of the robustness with respect to the model parameters. More precisely, we focus on the expectation and the variance of the robustness under a network probabilistic model. For some particular probabilistic models and some particular extinction sequences, the relation between the parameters and the expected robustness can be provided in a closed-form, making a fine study of the relation possible.

Furthermore, relying on a probabilistic model which can be adjusted to account for an observation process may enable for correction of the sampling effect on the robustness. Indeed, even though ecological networks are often considered to describe all the possible interactions between species, the sampling may be incomplete (Blüthgen et al., 2008) which may bias the computed network statistics (Rivera-Hutinel et al., 2012; de Manincor et al., 2020) such as the robustness.

The stochastic block model (SBM Nowicki and Snijders, 2001) and its extensions for bipartite networks, such as the biSBM (Govaert and Nadif, 2003, also referred to as Latent Block Model) or its degree corrected counterpart (DCbiSBM Larremore et al., 2014) have gained a lot of attention in the statistical and network science research

traditionally used in the information or epidemic networks literature. In these fields, the goal is to see how a network stays connected when some nodes are disabled (or on the opposite how epidemics are still able to spread when nodes become immune), the task is tackled by analyzing the size of the different connected components and in particular the existence of a giant connected component (see Newman, 2018, chapter 15 and reference therein). In ecology, the key issue is to observe how species get isolated from the rest of the ecosystem. This approach is related to the concept of isolated nodes in a graph (Erdős and Rényi, 1960).

fields during the last decade. While the Erdős-Rényi model (Erdős and Rényi, 1960) assumes that any pair of species has the same probability of interaction, SBMs introduce heterogeneity in the connection behavior. Specifically, in bipartite SBMs, the two sets of nodes are divided into clusters/blocks/groups and the probability that two nodes are connected depends on which blocks the nodes belong to. Depending on the number of blocks, their size, and interaction probabilities (inter- and intra-block), SBMs encompass a wide variety of topologies (such as assortative community or core-periphery) and encode them in a small number of parameters.

Since Allesina and Pascual (2009) has advocated for the use of groups in ecological networks, SBMs (sometimes referred to as group models) have gained in popularity. Some variants adapted to multilayer ecological networks have been proposed for: multiplex networks (Kéfi et al., 2016), multipartite networks (Bar-Hen et al., 2020) or temporal networks (Matias and Miele, 2017). Besides, they have been used to answer specific ecological questions. To name a few, Michalska-Smith et al. (2018) explore the structural role of parasite species in food webs, while Miele et al. (2020) use biSBM to assess the core-periphery structure of plant-pollinator networks. Furthermore, SBMs provide an ecological interpretation in terms of functional groups in the ecosystem: species in the same block interact in a similar way, which means that the exchangeability of species in a block model is related to the concept of ecological equivalence (Sander et al., 2015).

To our knowledge, the behavior of the robustness has never been studied in the SBM framework. Some grouping algorithms are used to derive extinction sequences. Cai and Liu (2016) optimize an objective function to determine communities in the network and then generates primary extinction based on these communities. More generally, some models are used to generate primary extinction sequences, or to model how species of the other functional group react to those extinctions (such as rewiring or cascading, see Bane et al., 2018; Vizentin-Bugoni et al., 2020, for examples). In an approach more similar to ours, Burgos et al. (2007) studied the relationship between nestedness and robustness of mutualistic networks by using the self-organizing network model (Medan et al., 2007) to generate nested networks and derived analytical expression of the robustness under this model.

The robustness given as a decreasing function and a robustness statistic classically used in ecology are described in Section 4.2. Section 4.3 is dedicated to the introduction of the biSBM, the DCbiSBM and related models for sequential species extinctions. Section 4.4 supplies the expression of the expectation and the variance of the robustness under a biSBM for different distributions of primary extinction sequences. The analytical properties of the expected robustness are given in Section 4.5 together with more general studies to illustrate the impact of the network structure on the robustness. In Section 4.6, we apply our approach on a dataset composed of both mutualistic and antagonistic bipartite networks and compare our results to the classical approach. Finally, on the same dataset, we show how our approach allows us to calculate the expected robustness of partially observed ecological networks.

4.2. Robustness of bipartite ecological networks

Robustness aims at measuring the tolerance of a network to species extinctions by quantifying the proportion of remaining species along a species extinction sequence. Formally, let $A \in \{0,1\}^{n_r \times n_c}$ be the $n_r \times n_c$ incidence matrix of a bipartite network representing ecological interactions between two groups of species of respective sizes n_r and n_c (such as plant-pollinators or hosts-parasites). Then:

$$A_{ij} = \begin{cases} 1 & \text{if row species } i \text{ interacts with column species } j, \\ 0 & \text{otherwise.} \end{cases}$$

Let s be an extinction sequence on the row species: $s \in \mathfrak{S}_{n_r}$ where $\mathfrak{S}_{n_r}$ is the symmetric group. The extinctions in row (whose order is given by s) are the *primary extinctions*. These row extinctions lead to secondary extinctions among the column species if these column species remain isolated after the disappearance of row species. More precisely, a column species j is said to be be extinct after m row primary extinctions if these m primary extinctions caused species j to lose all its connections, or equivalently if after these m primary row extinctions, j has no connections left with the remaining rows, which is equivalent to $\sum_{i=m+1}^{n_r} A_{s(i)j} = 0$.

For a given sequence s and a given number of primary extinctions m, $R(A,s,m)$ is the proportion of remaining column species:

$$R(A,s,m) = 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \mathbf{1}_{\{\sum_{i=m+1}^{n_r} A_{s(i)j} = 0\}}. \tag{4.1}$$

Note that for all $m \in \{0,\ldots,n_r\}$, $0 \leq R(A,s,m) \leq 1$ with $R(A,s,n_r) = 0$. The *robustness function* is defined as the expectation of $R(A,S,m)$ against the primary extinction sequences:

$$m \mapsto R_{\mathbb{S}}(A,m) = \mathbb{E}_S\left[R(A,S,m)\right] \quad \text{with} \quad S \sim \mathbb{S}. \tag{4.2}$$

where $\mathbb{S}$ is a probability distribution on $\mathfrak{S}_{n_r}$. Equation (4.2) is a weighted summation over the $n_r!$ possible permutations of $\mathfrak{S}_{n_r}$, which may render the computation intractable. In practice, it is generally approximated by a Monte Carlo integration:

$$\widehat{R}_{\mathbb{S}}(A,m) = \frac{1}{B} \sum_{b=1}^{B} R(A, S^{(b)}, m) \quad \text{where} \quad S^{(b)} \sim_{i.i.d} \mathbb{S}, \quad b = 1,\ldots,B. \tag{4.3}$$

The *robustness statistic* of a network A and a primary extinction sequences distribution $\mathbb{S}$ is defined as:

$$\overline{R}_{\mathbb{S}}(A) = \frac{1}{n_r} \sum_{m=0}^{n_r} R_{\mathbb{S}}(A,m), \tag{4.4}$$

which corresponds to the area under the curve (AUC) of the robustness function where the x-axis has been re-normalized to match with the proportion of removed row species. $\widehat{\overline{R}}_{\mathbb{S}}(A)$ is the Monte Carlo version of $\overline{R}_{\mathbb{S}}(A)$.

Other statistics exist in the literature, such as the median of the robustness function (Dunne et al., 2002), i.e. the proportion of primary extinctions needed to provoke a secondary extinction on half of the species. However, the statistic (4.4) is widely used in ecology and is mathematically convenient for our purpose.

About the extinction sequence distribution $\mathbb{S}$ Several choices for $\mathbb{S}$ corresponding to various ecological scenarios are commonly considered in the literature. The first choice is the uniform distribution over $\mathfrak{S}_{n_r}$ ($\mathbb{S} = \mathbb{U}_{\mathfrak{S}_{n_r}}$), assuming that the species disappear without specific order. For the sake of simplicity, we use $\mathbb{U}$ instead of $\mathbb{U}_{\mathfrak{S}_{n_r}}$.

Distributions $\mathbb{S}$ depending on A are suggested in the literature (see Curtsdotter et al., 2011, for examples). Among these approaches, we will focus on sequences depending on the row degree sequences. For any row species $i = 1, \ldots, n_r$, let $D_i = \sum_{j=1}^{n_c} A_{ij}$ be its degree, i.e. the number of edges involving i. On the one hand, the worst-case scenario in terms of ecological extinctions assumes that the row species with the highest degrees (most connected row species) disappear first. In this case, the species in rows are ordered in decreasing degrees and the primary extinction sequences follow this order; row species of equal degrees disappear in a uniformly distributed order. On the other hand, the generation of primary extinction sequences that first eliminate species of lower degree mimics a more favorable ecological scenario. In these two cases, the support of $\mathbb{S}$ is of cardinal $\prod_k \#\{i : D_i = k\}!$. Depending on the sequence $(D_i)_{i=1,\ldots,n_r}$, the corresponding $R_{\mathbb{S}}(A, m)$ may become tractable or not. If not, a Monte Carlo approximation is used.

Instead of considering a strict monotone ordering of the row degrees, one might relax the constraint and set the probability for a row species to disappear proportional to a function of its degree, as seen in Liu et al. (2019). This would correspond to sampling without replacement a sequence s where the weights of each species $i \in \{1, \ldots, n_r\}$ are, for instance given by:

$$w_i \propto D_i^{\alpha} \,. \tag{4.5}$$

$\alpha = 0$ coincides to the uniform distribution $\mathbb{U}$ while, if $\alpha = 1$, the primary extinction sequence depends linearly on the degrees. The increasing order is obtained by reversing this sequence of primary extinction.

Remark. *The $R_{\mathbb{S}}(A, m)$'s computed for the first two $\mathbb{S}$'s described above are the most widely used. They are implemented in the* `R` *package* `bipartite` *(Dormann et al., 2008) available on* `cran`*.*

The robustness function (4.2) and corresponding statistic (4.4) are then computed from a unique observed network A and a probability distribution $\mathbb{S}$ conditional to A (through the row degrees):

$$R_{\mathbb{S}}(A, m) = \mathbb{E}_{S|A}\left[R(A, S, m)\right] \quad \text{with} \quad S|A \sim \mathbb{S} \,. \tag{4.6}$$

This robustness computed for a given A will be referred to as the *empirical robustness*.

In order to understand the variability of the robustness with respect to the network structure we set a probabilistic model on the network $A \sim \mathbb{A}_\theta$ where the parameters θ embeds the network structure in a small number of parameters and then study the random variable $R_{\mathbb{S}}(A, m)$ for various distributions on (S, A). The following section is dedicated to the description of flexible probabilistic models on A and adapted joint distributions on A and S.

4.3. Bipartite Stochastic Block Model and related sequential extinctions

4.3.1. Probabilistic model on bipartite ecological networks

Bipartite stochastic block models The bipartite SBM, a.k.a. the Latent Block Model (Govaert and Nadif, 2003), is a mixture model on the edges adapted to bipartite networks. It relies on a clustering of nodes in rows and a clustering of nodes in columns. The nodes which belong to the same cluster (or equivalently block) are assumed to share the same connectivity profile in the network and the probability of interaction between two nodes depends on the blocks they belong to. More precisely, each of the n_r row species is attributed to a block $k \in \{1, \dots, Q_r\}$ independently from the other species. Let Z_i be such that $Z_i = k$ if row species i belongs to block k. The Z_i's are assumed to be independent and identically distributed (i.i.d.) and:

$$\mathbb{P}(Z_i = k) = \pi_k \quad i \in \{1, \dots, n_r\}, \tag{4.7}$$

with $\sum_{k=1}^{Q_r} \pi_k = 1$. For $j = 1, \dots, n_c$, let W_j be such that $W_j = q$ if column species j belongs to block $q \in \{1, \dots, Q_c\}$. The W_j's are assumed to be i.i.d. and:

$$\mathbb{P}(W_j = q) = \rho_q \quad j \in \{1, \dots, n_c\}, \tag{4.8}$$

with $\sum_{q=1}^{Q_c} \rho_q = 1$. Then, conditionally to their respective latent blocks, the interactions between two species are distributed independently as:

$$A_{ij} | \{Z_i = k, W_j = q\} \sim \mathcal{B}(\delta_{kq}), \tag{4.9}$$

where $\mathcal{B}$ is the Bernoulli distribution and $\delta_{kq} \in (0, 1)$ for $(k, q) \in \{1, \dots, Q_r\} \times \{1, \dots, Q_c\}$. The parameter $\boldsymbol{\theta} = \{\rho_q, \pi_k, \delta_{kq}\}_{k=1,\dots,Q_r, q=1,\dots,Q_c}$ then encodes the topology of the network. Let $\boldsymbol{n} = (n_r, n_c)$ be the numbers of species (richness) in rows and in columns of the network: $biSBM(\boldsymbol{\theta}, \boldsymbol{n})$ denotes the distribution on the networks defined by equations (4.7), (4.8) and (4.9) for a given value of $\boldsymbol{\theta}$. Note that $d = \sum_{k,q} \pi_k \delta_{kq} \rho_q$ is the expected connection probability for any pair of species (i, j). d will be referred to as the expected density.

Remark. *A special case among the biSBM distributions is the one where $\delta_{kq} = d$ for any k, q , or equivalently –in terms of models on networks– where $Q_r = 1$ and $Q_c = 1$. In that case, all connections occur independently with the same probability d. This biSBM is the bipartite Erdős-Rényi network.*

Degree corrected stochastic block models The degree corrected stochastic block model (Karrer and Newman, 2011) is an extension of the SBM which accounts for the heterogeneity in the degree distribution. This heterogeneity does not only depend on the blocks but also on some parameters related to the nodes themselves. The extension to bipartite network is done in Larremore et al. (2014). In these models, the distribution on the dyads is a Poisson distribution even if the observed network consists of binary edges. This is a reasonable approximation in the case where the network is large and sparse. However, in ecological interaction networks, this assumption is hardly met. That is why we use a true binary definition of the degree corrected biSBM that we denote by DCbiSBM in the following. On top of the blocks for the nodes in rows and columns given by variables Zs and Ws as defined in Equations (4.7) and (4.8), two vector parameters $\gamma_r \in \mathbb{R}^{n_r}$ and $\gamma_c \in \mathbb{R}^{n_c}$ are associated to the nodes and control their degree. The larger $\gamma_{r,i}$ (resp. $\gamma_{c,j}$) the higher the probability of connection involving species i (resp. j), compared to other species within the same block. Then, Equation (4.9) is replaced with

$$A_{ij}|\{Z_i = k, W_j = q\} \sim \mathcal{B}(1/(1 + \exp(-(\delta_{kq} + \gamma_{r,i} + \gamma_{c,j}))) \,, \tag{4.10}$$

with $\gamma_{r,1} = \gamma_{c,1} = 1$ for identifiability issues.

We denote by $DCbiSBM(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{n})$ where $\boldsymbol{\gamma} = (\gamma_r, \gamma_c)$, this probabilistic model with a given set of parameters. If there is only one block, i.e. $Q_r = Q_c = 1$, this model only accounts for the heterogeneity of degrees and so corresponds to a bipartite version of the expected degree distribution (EDD) model (Chung and Lu, 2002).

4.3.2. Extinction sequence distributions adapted to bipartite Block Models

As described in Section 4.2, classically, the extinction sequences are either distributed uniformly on $\mathfrak{S}_{n_r}$ or conditionally to A. Positing a probabilistic distribution on A leads to a joint distribution on A and S. In the case of a uniform distribution on S: $S \sim \mathbb{U}$, the joint distribution on (A, S) consists of the product of the distributions for A and S. We will then denote by $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$ and $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{n},\mathbb{U}}$ the corresponding joint distributions when $A \sim biSBM(\boldsymbol{\theta}, \boldsymbol{n})$ or $A \sim DCbiSBM(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{n})$ respectively.

As described in Section 4.2, the extinction sequences may depend on A through its degree distributions. When A follows a probabilistic distribution, such an extinction sequence is defined conditionally on the realization of A. We propose another dependence between A and S through the row clustering variables Zs. More precisely, we consider that species of row block 1 disappear first, then block 2, etc. Formally, let $\mathbb{B}$ be the uniform probability on the row species following a given ordering of the blocks, then $S|Z \sim \mathbb{B}$ if:

$$\mathbb{P}(S = s|Z) = \frac{\mathbf{1}_{Z_{s(1)} \leq \cdots \leq Z_{s(n_r)}}}{\#\{s : Z_{s(1)} \leq \cdots \leq Z_{s(n_r)}\}}. \tag{4.11}$$

The support of $\mathbb{B}$ is restricted to $\prod_{k=1}^{Q_r} n_k!$ elements, where n_k is the cardinal of block k. Equations (4.7), (4.8), (4.9) and (4.11) define a joint probability distribution on

(A, S) such that A and S are independent conditionally to Z. We denote $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}$ this joint distribution on (A, S) when $A \sim biSBM(\boldsymbol{\theta}, \boldsymbol{n})$. Under this joint distribution, if $\boldsymbol{\theta}$ is such that $\delta_{k+} = \sum_{q=1}^{Q_c} \rho_q \delta_{kq}$ is a decreasing (resp. increasing) sequence, then $\mathbb{B}$ will generate sequences with an expected decreasing (resp. increasing) sequence of degrees, i.e. with decreasing (resp. decreasing) connectivity. In the following, these extinction sequences will be referred to as block-decreasing (resp. block-increasing). The blocks may also be used to generate extinctions by ecological traits that correspond to the blocks.

If $A \sim DCbiSBM(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{n})$, the extinction sequences are not related to expected degrees anymore but it may still make sense to generate extinction sequences linked to ecological traits.

4.4. Moments of the robustness statistic

Studying the distribution of $R(A, S, m)$ or $\overline{R}(A, S) = \frac{1}{n_r} \sum_{m=1}^{M} R(A, S, m)$ under the joint distribution of A and S could be done by simulation, using the fact that the biSBM and DCbiSBM are generative models. However, this comes at a computational cost. In this section, we prove that, when $(A, S) \sim \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$ or $(A, S) \sim \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}$ the first moments of the robustness are tractable in a closed-form. The proof relies on the exchangeability of the nodes under a biSBM and on the fact that the considered extinction sequences are adapted to this exchangeability. When A follows a DCbiSBM, the nodes are no longer exchangeable because of the parameters $\boldsymbol{\gamma}$ associated with each node. Although we are able to obtain a closed-form expression, it is not tractable. Therefore, we will rely on a Monte Carlo approximation for summing on the extinction sequences.

We now exhibit explicit expressions of $R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{S}}(m) = \mathbb{E}_A[R_{\mathbb{S}}(A, m)]$ when $(A, S) \sim \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$ or $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}$ and of $\mathbb{V}_A[R_{\mathbb{U}}(A, m))]$ when $A \sim biSBM(\boldsymbol{\theta}, \boldsymbol{n})$.

4.4.1. Expectation

Proposition 4.1. *Let $(A, S) \sim \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$ and set $\delta_{+q} = \sum_{k=1}^{Q_r} \pi_k \delta_{kq}$. Then,*

- $\forall m = 1, \ldots, n_r$:

$$R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m) \quad = \quad 1 - \sum_{q=1}^{Q_c} \rho_q (1 - \delta_{+q})^{n_r - m}, \tag{4.12}$$

- *Consequently, the robustness statistic is:*

$$\overline{R}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}} \quad = \quad \frac{1}{n_r} \sum_{m=0}^{n_r} R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m) = 1 - \frac{1}{n_r} \sum_{q=1}^{Q_c} \rho_q \frac{(1 - \delta_{+q}) - (1 - \delta_{+q})^{n_r+1}}{\delta_{+q}} \tag{4.13}$$

Proof. Since $(A, S) \sim \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$,

$$R_{A,S}(m) \;=\; \mathbb{E}_A[\mathbb{E}_S[\,R(A,S,m)\,]] = \mathbb{E}_A[\,R_{\mathbb{U}}(A,m)\,] = \frac{1}{n_r!}\sum_{s\in\mathfrak{S}_{n_r}}\left(1-\frac{1}{n_c}\sum_{j=1}^{n_c}\mathbb{E}_A\left[\mathbf{1}_{\sum_{i=m+1}^{n_r}A_{s(i)j}=0}\right]\right).$$

Using the property of exchangeability of the biSBM, we have that, for any permutation s, $\mathbb{E}_A\left[\mathbf{1}_{\sum_{i=m+1}^{n_r}A_{s(i)j}=0}\right] = \mathbb{E}_A\left[\mathbf{1}_{\sum_{i=m+1}^{n_r}A_{ij}=0}\right]$. Moreover, introducing the latent variables Z and W we have:

$$\begin{aligned}
\mathbb{E}_A\left[\mathbf{1}_{\sum_{i=m+1}^{n_r}A_{ij}=0}\right] &= \sum_{k_{m+1:n_r}\in\{1,\dots,Q_r\}^{n_r-m}}\sum_{q=1}^{Q_c}\mathbb{P}\left[\sum_{i=m+1}^{n_r}A_{ij}=0|Z_{m+1:n_r}=k_{m+1:n_r},W_j=q\right]\\
&\qquad\times\mathbb{P}\left(Z_{m+1:n_r}=k_{m+1:n_r},W_j=q\right)\\
&= \sum_{k_{m+1:n_r}\in\{1,\dots,Q_r\}^{n_r-m}}\sum_{q=1}^{Q_c}\prod_{i=m+1}^{n_r}(1-\delta_{k_iq})\left(\prod_{i=m+1}^{n_r}\pi_{k_i}\right)\rho_q\\
&= \sum_{q=1}^{Q_c}\rho_q\sum_{k_{m+1:n_r}\in\{1,\dots,Q_r\}^{n_r-m}}\prod_{i=m+1}^{n_r}(1-\delta_{k_iq})\pi_{k_i} = \sum_{q=1}^{Q_c}\rho_q\left(\sum_{k=1}^{Q_r}\pi_k(1-\delta_{kq})\right)^{n_r-m}.
\end{aligned}$$

As a consequence,

$$R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m) \;=\; 1-\sum_{q=1}^{Q_c}\rho_q\left(\sum_{k=1}^{Q_r}\pi_k(1-\delta_{kq})\right)^{n_r-m} = 1-\sum_{q=1}^{Q_c}\rho_q(1-\delta_{+q})^{n_r-m}$$

where $\delta_{+q}=\sum_{k=1}^{Q_r}\pi_k\delta_{kq}$. Then, averaging over m leads to:

$$\overline{R}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}} \;=\; \frac{1}{n_r}\sum_{m=0}^{n_r}R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m) = 1-\frac{1}{n_r}\sum_{q=1}^{Q_c}\rho_q\frac{(1-\delta_{+q})-(1-\delta_{+q})^{n_r+1}}{\delta_{+q}}.$$

□

Note that, if the network has no specific structure ($\delta_{kq}=d$ or $Q_r=Q_c=1$) then $R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m)=1-(1-d)^{n_r-m}$.

Proposition 4.2. *Let* $(A,S)\sim\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}$, *then*

$$R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}(m) = 1-\sum_{q=1}^{Q_c}\rho_q\sum_{n_1+\dots+n_{Q_r}=n_r}\frac{n_r!}{n_1!\dots n_{Q_r}!}\prod_{k=1}^{Q_r}\pi_k^{n_k}\left(1-\delta_{kq}\right)^{\min^+(n_k,\sum_{l\le k}n_l-m)}, \tag{4.14}$$

where $\min^+$ *is the positive part of the minimum function:* $\min^+(x,y) = \max(0,\min(x,y))$.

The proof of Proposition 4.2 is provided in the Appendix 4.A. The robustness statistic $\overline{R}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}$ is the mean of the $R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}(m)$'s: no simplified expression has been obtained. Note that, if $Q_r=1$ or if $\delta_{kq}=d$, then $R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}(m)=R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m)$.

In Equation (S-4.19), the summation over the partitions of the n_r row species into Q_r blocks may be burdensome if n_r or Q_r are large. In such cases, a Monte Carlo approximation may be used. Among the ecological networks we consider in Section 4.6, only a few of them require this approximation.

4.4.2. Variance

We now aim at computing the variance of the robustness $\mathbb{V}_A[R_{\mathbb{U}}(A,m)]$, when $A \sim biSBM(\boldsymbol{\theta}, \boldsymbol{n})$, thus quantifying the variability of the robustness for a sample of networks sharing the same block models parameters (i.e. the same mesoscale patterns of connectivity).

Proposition 4.3. *Let $A \sim biSBM(\boldsymbol{\theta}, \boldsymbol{n})$, $\eta_q = 1 - \delta_{+q}$ and $\eta_{qq'} = \sum_{k=1}^{Q_r} \pi_k (1 - \delta_{kq})(1 - \delta_{kq'})$.*

1. *Then:*

$$\begin{aligned}\mathbb{V}_A[R_{\mathbb{U}}(A,m)] &= \frac{1}{n_c} \sum_{l=m}^{\min(2m,n_r)} \frac{\binom{m}{l-m}\binom{n_r-m}{l-m}}{\binom{n_r}{m}} \sum_{q=1}^{Q_r} \rho_q \eta_q^l - (\sum_{q=1}^{Q_r} \rho_q \eta_q^m)^2 \\ &+ \frac{(n_c-1)}{n_c} \sum_{l=m}^{\max(2m,n_r)} \frac{\binom{m}{l-m}\binom{n_r-m}{l-m}}{\binom{n_r}{m}} \sum_{q,q'=1}^{Q_r} \rho_q \rho_{q'} (\eta_q \eta_{q'})^{l-m} \eta_{qq'}^{2m-l}.\end{aligned}$$

2. *The variance of the robustness statistic due to the network variability under a given biSBM is:*

$$\begin{aligned}\mathbb{V}_A[\overline{R}_{\mathbb{U}}(A)] &= \frac{1}{n_r^2 n_c^2} n_c \sum_{m,m'=0}^{n_r} \sum_{l=\max(m,m')}^{\min(m+m',n_r)} \frac{\binom{m}{l-m'}\binom{n_r-m}{l-m}}{\binom{n_r}{m'}} \sum_{q=1}^{Q_r} \rho_q \eta_q^l - (\frac{1}{n_r} \sum_{m=0}^{n_r} \sum_{q=1}^{Q_r} \rho_q \eta_q^m)^2 \\ &+ \frac{1}{n_r^2 n_c^2} n_c (n_c - 1) \sum_{m,m'=0}^{n_r} \sum_{l=\max(m,m')}^{\min(m+m',n_r)} \frac{\binom{m}{l-m'}\binom{n_r-m}{l-m}}{\binom{n_r}{m'}} \sum_{q,q'=1}^{Q_r} \rho_q \rho_{q'} \eta_q^{l-m'} \eta_{q'}^{l-m} \eta_{qq'}^{m+m'-l}.\end{aligned}$$

The proof is provided in the supplementary material 4.A. The expectations of the robustness under the biSBM given in Proposition 4.1 and 4.2 only rely on the model parameters $\boldsymbol{\theta}$ and the number of rows n_r, whereas the expression of the variance also involves the number of columns n_c. Note that for the block sequences $\mathbb{S} = \mathbb{B}$, the calculus is not so simple, but the value could still be estimated by simulations if required.

$\mathbb{V}_A(\overline{R}_{\mathbb{U}}(A))$ quantifies the variability of the robustness statistic among a population of networks distributed as biSBM. However, from an ecological point of view, it could also be interesting to quantify the variability of the robustness of a network with respect not only to the network but also to the extinction sequence. This is obtained by computing

$$\mathbb{V}_{A,S}\left(\frac{1}{n_r} \sum_{m=1}^{n_r} R(A,S,m)\right) = \mathbb{V}_{A,S}\left(\overline{R}(A,S)\right)$$

Remark. *This total variance can be reformulated as:*

$$\mathbb{V}_{A,S}\left(\overline{R}(A,S)\right) = \mathbb{E}_S\left(\mathbb{V}_{A|S}(\overline{R}(A,S))\right) + \mathbb{V}_S\left(\mathbb{E}_{A|S}(\overline{R}(A,S)\ )\right) = \mathbb{E}_S\left(\mathbb{V}_A(\overline{R}(A,S))\right) \tag{4.15}$$

$$= \mathbb{E}_A\left(\mathbb{V}_S(\overline{R}(A,S))\right) + \mathbb{V}_A\left(\mathbb{E}_S(\overline{R}(A,S)\ )\right) \tag{4.16}$$

because S and A are independent under $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$ and $\mathbb{E}_{A|S}(\overline{R}(A,S))$ does not depend on S by exchangeability. The total variance can be expressed explicitly for $\mathbb{S}=\mathbb{U}$ as

$$\mathbb{V}_{A,S}\left(\overline{R}(A,S)\right) = \frac{n_c}{n_r^2 n_c^2}\sum_{m,m'=0}^{n_r}\sum_{q=1}^{Q_r}\rho_q\eta_q^{n_r-\min(m,m')} - \left(\frac{1}{n_r}\sum_{q=1}^{Q_c}\rho_q\frac{\eta_q-\eta_q^{n_r+1}}{\delta_{+q}}\right)^2$$
$$+\frac{n_c(n_c-1)}{n_r^2 n_c^2}\sum_{m,m'=1}^{n_r}\sum_{q,q'=1}^{Q_r}\rho_q\rho_{q'}\eta_q^{\max(m,m')-\min(m,m')}\eta_{qq'}^{n_r-\max(m,m')},$$

through the computation of Equation (4.15). The second term in Equation (4.16) is provided in Proposition 4.3. Thus we are able to compute the remaining terms to understand the various sources of variability (due to S or A).

4.4.3. Illustration of the variability of the robustness function

In Figure 4.1, we illustrate the variations of the robustness function for a given structure encoded in

$$\boldsymbol{\theta} = \left(\delta = \begin{pmatrix} .4 & .15 \\ .25 & .05 \end{pmatrix}, \pi = (0.25, 0.75), \rho = (0.2, 0.8)\right).$$

This corresponds to a so-called core-periphery structure where the first blocks (in rows and columns) are well connected (the core) and the second blocks are less connected (the periphery). We represent the functions $m \mapsto R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{S}}(m)$ with $\mathbb{S}=\mathbb{U}$ and $\mathbb{S}=\mathbb{B}$ such that the blocks are ordered by decreasing or increasing connectivity (solid black lines in Figure 4.1). For the uniform distribution, we also plot the area given by $m \mapsto R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}(m) \pm 2\sqrt{\mathbb{V}_A(R_{\mathbb{U}}(A,m))}$ (grey ribbon). The colored dotted lines are Monte Carlo estimates of the robustness functions $\widehat{R}_{\mathbb{S}}(A,m)$ corresponding to 10 simulated networks for the same extinction sequence distributions. The Monte Carlo estimates are computed over 300 realizations of extinction sequences. The inflection points and the dispersion of the dotted lines observed respectively around an extinction rate of 0.25 for decreasing extinction and 0.75 for increasing extinction correspond respectively to the proportion of the core or the periphery blocks. When the rate of extinction exceeds one of these proportions, the extinctions which were only happening in one block then happen in the other block. Also notice that when $\mathbb{S}=\mathbb{B}$ ordered by increasing connectivity, some networks still have a robustness of 1 even after a large fraction of primary extinctions has occurred. The robustness for this primary extinction distribution is highly dependent on the degree of the most connected primary species.

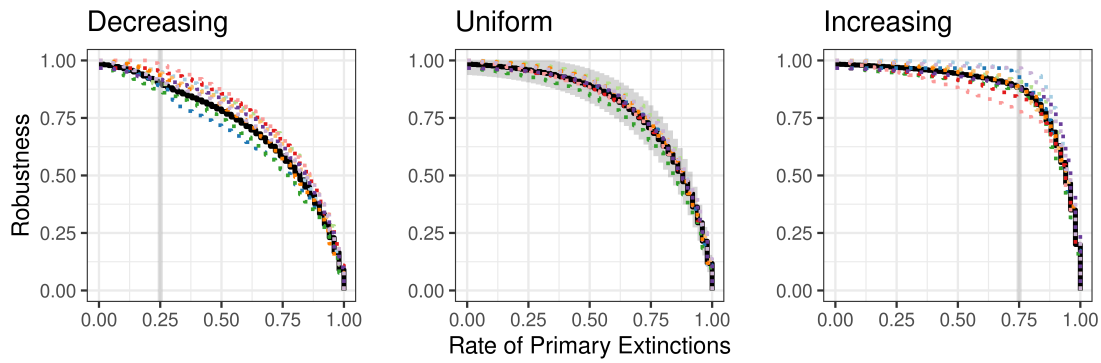


Figure 4.1 – Robustness function computed from a set of biSBM parameters (plain black) and by Monte Carlo for 10 networks generated from the same biSBM distribution (dotted) for block decreasing, uniform and block increasing primary extinction sequences. The grey ribbon on the uniform facet is twice the standard error given in Proposition 4.3.

4.5. Impact of the Network Structure on the Robustness

From the expressions derived in the previous section when $A \sim biSBM(\boldsymbol{\theta}, \boldsymbol{n})$, we are now able to study the average behavior of the robustness with respect to the mesoscale structure of the network encoded in $\boldsymbol{\theta}$.

4.5.1. Analytical Properties

First properties on $R_{\theta,\mathbf{n},\mathbb{S}}(m)$

We first derive the following straightforward but useful properties of the robustness function and statistic:

Properties 4.1. *1. Under the joint distribution $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{S}}$ where $\mathbb{S} \in \{\mathbb{U}, \mathbb{B}\}$, the following properties hold:*

(a) the function $m \in \{0, \dots, n_r\} \mapsto R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{S}}(m)$ is a strictly decreasing function provided that $\delta_{k+} > 0$ for all $k > 0$,

(b) $R_{\boldsymbol{\theta},n,\mathbb{S}}(0) \leq 1 - (1-d)^{n_r} \leq 1$ where $d = \sum_{k,q} \pi_k \rho_q \delta_{kq}$,

(c) For $\boldsymbol{\theta} = (\pi, \rho, \delta)$ and $\boldsymbol{\theta}' = (\pi', \rho', \delta')$ such that $\pi = \pi'$, $\rho = \rho'$ and $\forall (k,q) \in \{1, \dots, Q_r\} \times \{1, \dots, Q_c\}$ $\delta_{kq} \leq \delta'_{kq}$, we have

$$\forall m \in 0, \dots, n_r, R_{\boldsymbol{\theta},n,\mathbb{S}}(m) \leq R_{\boldsymbol{\theta}',n,\mathbb{S}}(m) \quad \text{and} \quad \overline{R}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{S}} \leq \overline{R}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{S}}\,.$$

2. Under the joint distribution $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$, if $n_r \leq n'_r$ then $\overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}} < \overline{R}_{\boldsymbol{\theta},(n'_r,n_c),\mathbb{U}}$.

Proof. Property *1.(a)* comes from the robustness definition and from the fact that $s(m+1:n_r) \subset s(m:n_r)$ for any extinction sequence s. Property *1.(b)* is a consequence of Proposition 4.4. Property *1.(c)* is true for all $m \in \{0,\dots,n_r\}$ from Equations (4.12) and (S-4.19).

For Property *2.*, first notice that $R_{\boldsymbol{\theta},(n_r+1,n_c),\mathbb{U}}(m+1) = R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(m)$. Then,

$$\begin{aligned}
\overline{R}_{\boldsymbol{\theta},(n_r+1,n_c),\mathbb{U}} &= \frac{1}{n_r+1}\Big(R_{\boldsymbol{\theta},(n_r+1,n_c),\mathbb{U}}(0) + \sum_{m=0}^{n_r} R_{\boldsymbol{\theta},(n_r+1,n_c),\mathbb{U}}(m+1)\Big) \\
&= \frac{1}{n_r+1}\Big(R_{\boldsymbol{\theta},(n_r+1,n_c),\mathbb{U}}(0) + \sum_{m=0}^{n_r} R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(m)\Big) \\
&= \frac{1}{n_r+1}\Big(R_{\boldsymbol{\theta},(n_r+1,n_c),\mathbb{U}}(0) + n_r\overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}\Big) \\
\text{(Property 1.(a))} \quad &> \frac{1}{n_r+1}\Big(R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(0) + n_r\overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}\Big) \\
&> \frac{1}{n_r+1}\Big(\overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}} + n_r\overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}\Big) = \overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}
\end{aligned}$$

□

Properties *1.(a)* and *1.(b)* supply a bound depending on the expected density and the number of row species n_r. Property *1.(d)* implies that for a given block structure, the robustness increases when each element δ_{kq} increases while Property *2.* states that the uniform robustness automatically increases with the size n_r of the network. However, it is important to note that these properties do not assess that the robustness is an increasing function of the connectance d.

Upper bound for the robustness under a uniform extinction sequence

We now aim at identifying mesoscale structures maximizing the average robustness under the joint distribution $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$. In order to remove the effect of the mean number of interactions a.k.a. the density $d = \sum_{k,q}\pi_k\delta_{kq}\rho_q$, we propose to compare structures encoded in $\boldsymbol{\theta}$ leading to the same density value. We also fix the number of row species n_r. Thus, for a given density d, we define the set $\Theta_d = \{\boldsymbol{\theta} = (\pi,\rho,\delta) : \sum_{k,q}\pi_k\delta_{kq}\rho_q = d\}$. The following proposition provides an upper bound for the expectation of the robustness function and of the robustness statistic as a function of the density. It also identifies a condition on $\boldsymbol{\theta}$ to achieve this upper bound. Eventually, it shows that the parametrization of an Erdős-Rényi distribution satisfies this condition and reaches the lowest variance of the robustness statistic among the parametrizations satisfying this condition.

Proposition 4.4. *Upper bound of robustness under* $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$.

1. *For all* $m \in \{0\dots,n_r\}$*:* $R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(m) \leq 1-(1-d)^{n_r-m}$.
2. *For all* $m \in \{0\dots,n_r-2\}$

$$\arg\max_{\boldsymbol{\theta}\in\Theta_d} R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(m) = \left\{\boldsymbol{\theta} : \sum_{k=1}^{Q_r}\pi_k\delta_{kq} = \sum_{k=1}^{Q_r}\pi_k\delta_{kq'}, \quad \forall(q,q')\in\{1,\dots,Q_c\}^2\right\} := \Theta^{\max}_{d,n_r},$$

Moreover, $\forall \boldsymbol{\theta} \in \Theta_d$ and $\forall n_c$:

$$\overline{R}_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}} \leq 1 - \frac{1}{n_r}\frac{(1-d)-(1-d)^{n_r+1}}{d}. \tag{4.17}$$

3. *Assume that E is a network with n_r rows and n_c columns following an Erdős-Rényi distribution with density parameter d. Then:*

$$\min_{A\sim biSBM(\boldsymbol{\theta},\boldsymbol{n}):\boldsymbol{\theta}\in\Theta^{\max}_{d,n_r}} \mathbb{V}_A(\overline{R}_{\mathbb{U}}(A)) = \mathbb{V}_E(\overline{R}_{\mathbb{U}}(E)).$$

In other words, among all parameters $\boldsymbol{\theta} \in \Theta^{\max}_{d,n_r}$, the one of the Erdős-Rényi network minimizes the variance of robustness statistic $\overline{R}_{\mathbb{U}}(A)$ given in Proposition 4.3.

Proof. 1. Recall that $R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(m) = 1 - \sum_{q=1}^{Q_c} \rho_q (1-\delta_{+q})^{n_r-m}$ where $\delta_{+q} = \sum_{k=1}^{Q_r} \pi_k \delta_{kq}$ for any $q \in \{1,\dots,Q_r\}$. So, $R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(n_r) = 0$ and for $R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(n_r-1) = 1-(1-d)$: the bound is true for $m = n_r$ and $m = n_r - 1$. Now, if $0 \leq m \leq n_r - 2$, then $x \mapsto x^{n_r-m}$ is a strictly convex function and the Jensen. inequality applies:

$$\begin{aligned} R_{\boldsymbol{\theta},(n_r,n_c),\mathbb{U}}(m) &= 1 - \sum_{q=1}^{Q_c} \rho_q (1-\delta_{+q})^{n_r-m} \leq 1 - \left(\sum_{q=1}^{Q_r} \rho_q (1-\delta_{+q})\right)^{n_r-m} \quad (4.18) \\ &= 1 - \left(1 - \sum_{q=1}^{Q_r} \rho_q \delta_{+q}\right)^{n_r-m} = 1-(1-d)^{n_r-m}, \end{aligned}$$

2. Equality at line (4.18) holds if and only if the term inside the strictly convex function is constant, ie. for any q, q', $\delta_{+q} = \delta_{+q'}$.

3. When we compute the variance given in Proposition 4.3(2) on any biSBM with $\theta \in \Theta^{\max}_{d,n_r}$, we notice that $\eta_q = \eta_{q'} = 1-d$ and that only the term $\sum_{q,q'} \rho_q \rho_{q'} \eta_q^{l-m'} \eta_{q'}^{l-m} \eta_{qq'}^{m+m'-l}$ varies. We can reformulate this quantity and use the Jensen inequality:

$$\begin{aligned} \sum_{q,q'} \rho_q \rho_{q'} \eta_q^{l-m'} \eta_{q'}^{l-m} \eta_{qq'}^{m+m'-l} &= (1-d)^{l-m'}(1-d)^{l-m} \sum_{q,q'} \rho_q \rho_{q'} \eta_{qq'}^{m+m'-l} \\ \text{(Jensen)} &\geq (1-d)^{2l-m'-m} \left(\sum_{q,q'} \rho_q \rho_{q'} \sum_k \pi_k (1-\delta_{kq})(1-\delta_{kq'})\right)^{m+m'-l} \\ \text{(Distributivity)} &= (1-d)^{2l-m'-m} \left(\sum_k \pi_k (1-\sum_q \rho_q \delta_{kq})(1-\sum_{q'} \rho_{q'} \delta_{kq'})\right)^{m+m'-l} \\ \text{(Jensen)} &\geq (1-d)^{2l-m'-m} \left(\left(\sum_k \pi_k (1-\sum_q \rho_q \delta_{kq})\right)^2\right)^{m+m'-l} \\ &= (1-d)^{m'+m} \qquad \text{(Initial term from ER)} \end{aligned}$$

□

Although the homogeneous distribution on the networks (Erdős-Rényi) leads to the maximum robustness in expectation for a given density, a particular realization of a network according to another distribution may be likely to have a larger robustness than a realization according to the Erdős-Rényi distribution. Indeed, the variance is larger when it corresponds to a distribution that represents a more complex structure. This behavior is illustrated in Figure 4.2.

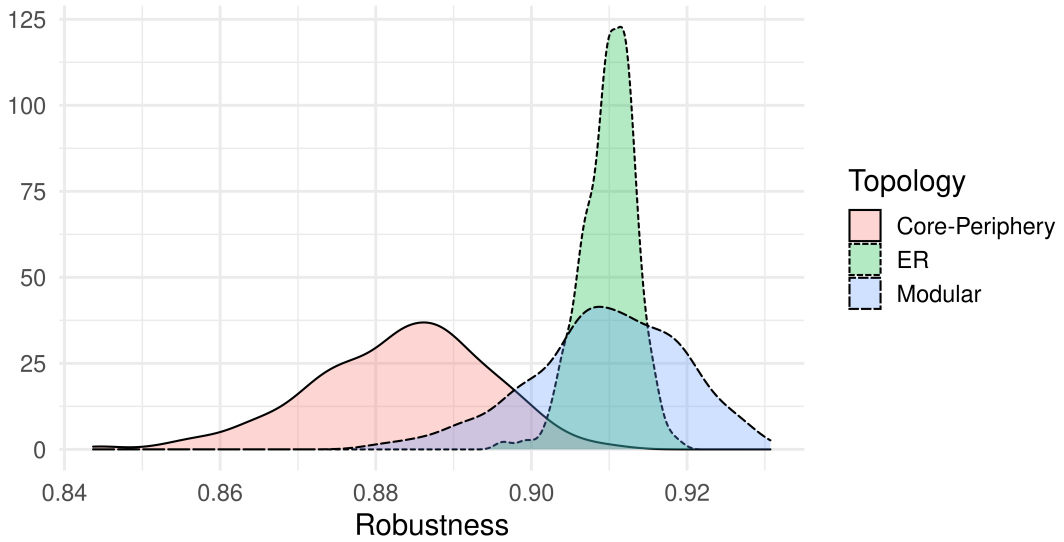


Figure 4.2 – Density curve of the distribution of the expectation of the robustness statistic under biSBMs of the same size and density with three different sets of parameters. Each network is simulated from a given set of parameters, then the robustness is computed by the Monte-Carlo approximation given in (4.4). 500 simulations each.

4.5.2. Analysis for Typical Structures

We now illustrate numerically how the robustness statistic varies with respect to the network topology [2]. For that purpose, we fix $n_r = n_c = 100$, $Q_c = Q_r = 2$, and $\pi = (.25, .75)$ and make ρ vary. For $j \in [1/8, 8]$, we consider the following connectivity matrices.

Modular: $\delta = \begin{pmatrix} j & 1 \\ 1 & j \end{pmatrix}$, each block of row species is strongly connected to a block of column species and lowly connected to the other.

Core-periphery: $\delta = \begin{pmatrix} j & j \\ j & 1 \end{pmatrix}$.

[2] All the simulations and estimations in this section are done using the `R` package `robber` (Chabert-Liddell, 2021b)available on `CRAN` and documented at https://chabert-liddell.github.io/robber/.

- For $j > 1$, the structure is nested, the core is strongly connected to the whole network while the periphery is lowly connected with itself.
- For $j < 1$, the core is strongly connected with itself but the rest of the network is lowly connected.

Each connectivity matrix is then normalized such that the density of the network is equal to 0.0156, by applying the following transformation: $\tilde{\delta}_{kq} = 0.0156 \frac{\delta_{kq}}{\sum_{k',q'} \pi_{k'} \delta_{k'q'} \rho_{q'}} \quad \forall k, q \in \{1, 2\}$. This value is chosen so the robustness statistic associated with an Erdős-Rényi distribution with density $d = 0.0156$ and $n_r = 100$ rows is approximately equal to 0.5. We recall this value is an upper bound of the expectation of the robustness statistic under the joint distribution $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{U}}$ where $\boldsymbol{\theta}$ is such that the network has the same density and same number of rows.

The extinction sequence distributions are $\mathbb{S} \in \{\mathbb{U}, \mathbb{B}^{\uparrow}, \mathbb{B}^{\downarrow}\}$ where $\mathbb{B}^{\uparrow}$ (resp. $\mathbb{B}^{\downarrow}$) corresponds to the block increasing (resp. decreasing) extinction sequences distribution defined in Equation (4.11). For every $\mathbb{S}$, topology (modular or core-periphery) and value of j and ρ, we compute the expectations of the robustness statistics by using the expressions derived in Section 4.4. These expectations are displayed in the heat maps of Figure 4.3.

For modular networks, the plots are symmetric with respect to the central dot which corresponds to the case of an Erdős-Rényi distribution. The impact of the modular structure is slighter for $\mathbb{S} = \mathbb{U}$ than for $\mathbb{S} \in \{\mathbb{B}^{\uparrow}, \mathbb{B}^{\downarrow}\}$. For $\mathbb{S} = \mathbb{U}$, the strongest impact is observed when the modular structure is strong and the most connected blocks are slightly larger than the least connected ones (Fig. 4.3.◊). Keeping the same strong modular structure with the most connected blocks slightly smaller than the least connected ones (Fig. 4.3.△) leads to a negative impact on the robustness no matter $\mathbb{S}$.

When the network is highly modular and the most connected block is small (j and ρ both small or both large, Fig. 4.3.□) the robustness is strongly impacted. This impact is negative for $\mathbb{B}^{\downarrow}$ and positive with $\mathbb{B}^{\uparrow}$.

For a core-periphery structure, there is a clear asymmetry between $j < 1$ and $j > 1$. The robustness statistic tends to be smaller when the core is mainly connected to itself, especially when the core is small ($j < 1$ and large ρ: Fig. 4.3.■) no matter $\mathbb{S}$. On the other hand, core-periphery structure with small core that are highly connected to the whole network have the strongest impact on the robustness, negatively when $\mathbb{S} = \mathbb{B}^{\downarrow}$ and positively when $\mathbb{S} = \mathbb{B}^{\uparrow}$ (Fig. 4.3.▲). The effect of this structure (▲) is very slight on uniform extinction sequences whereas the effect tends to get larger when the blocks' sizes are more balanced (Fig. 4.3.♦).

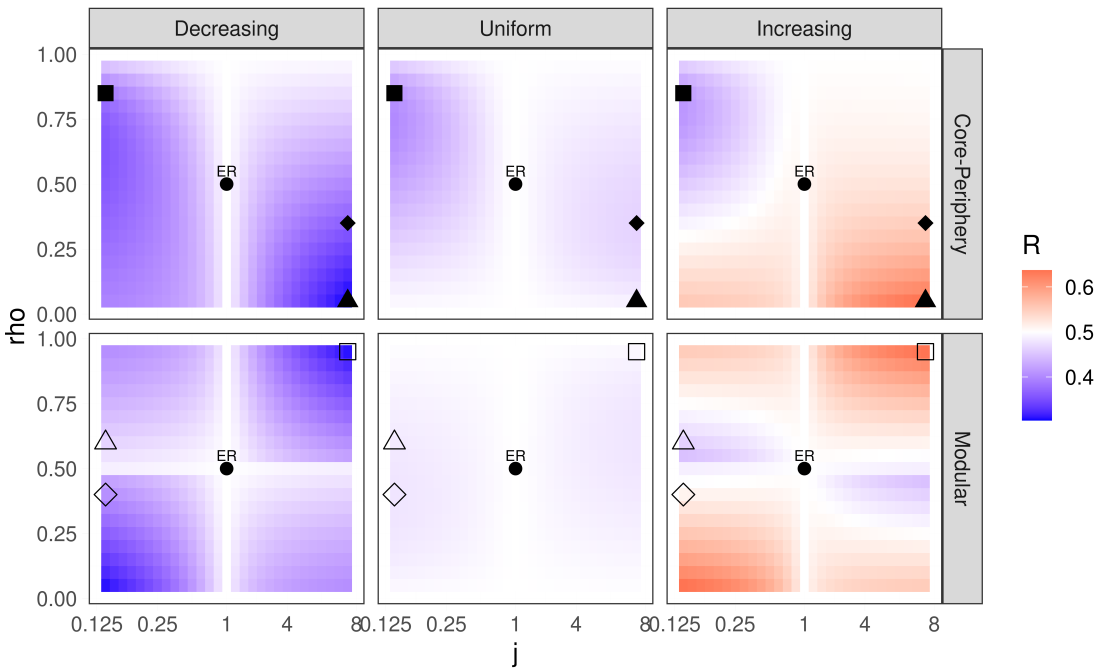


Figure 4.3 – Robustness for different biSBM topologies. Primary extinction sequences are displayed in rows and topologies in columns. In abscissa, j is a topology strength parameter. `rho` is the proportion of column species that belongs to the first column block. • represents a topology with no structure (Erdős-Rényi (ER)). Color gradient varies from blue (less robust than an ER), to white (as robust as an ER), to red (more robust than an ER). The symbols corresponds to the following mesoscale structures: ■ - large core, strongly connected only to itself, ▲ - small core, strongly connected to the whole network, ♦ - medium size core strongly connected only to itself, □ - highly modular with unbalanced blocks' sizes, ◊ - highly modular with slightly smaller highly connected blocks, △ - highly modular with slightly larger highly connected blocks.

4.6. Analysis of a collection of observed bipartite ecological networks

In this section, we analyse the robustness of a collection of 136 plant-pollinator, seed-dispersal or host-parasite networks, issued from the Web of Life dataset (www.web-of-life.es). The selected networks involve at least 10 row species and 10 column species.

When observing an interaction network, one can compute its empirical robustness as defined in Equation (4.3). However, understanding the behavior of the robustness statistic for any network with the same probabilistic distribution (i.e. with the same mesoscale structure) may also be attractive and informative. This can be done by positing a model on the network of interest, estimating the corresponding parameters (summarizing its structure) and deriving the moments of the robustness for these inferred parameters using Section 4.4 (either using a closed-form expression or a Monte Carlo integration, depending on the model). This expected version of the standard robustness may be considered as a new robustness indicator.

In what follows, we put in perspective the empirical robustness and its expected versions for the biSBM and DCbiSBM models (Subsection 4.6.1) and comment the systematic differences we observe. Then, we demonstrate the interest of the expected version when the network is partially observed (Subsection 4.6.2). Indeed, when inferring the bipartite SBM, the observational process that generates the observed network with possibly missing data could be taken into account (Tabouy et al., 2019). This which allow us to compensate for observational biases in the empirical robustness.

Inferring the biSBM and DCbiSBM The parameters of these two models can be inferred by a variational version of the Expectation-Maximization (EM) algorithm. The number of blocks is chosen according to an Integrated Classification Likelihood (ICL) criterion (Daudin et al., 2008). This variational EM algorithm is theoretically grounded (Bickel et al., 2013) and has proven its pratical efficiency (Daudin et al., 2008; Mariadassou et al., 2010). In practice for the inference of these models, we use the `blockmodels` `R` package (Leger et al., 2020) and to handle missing observations the `GREMLINS` `R` package (Bar-Hen et al., 2020), both available on CRAN .

4.6.1. Computation of the Robustness for the Web of Life Dataset

For each network A, on the one hand, we compute the empirical robustness $\widehat{\overline{R}}_{\mathbb{U}}(A)$ using 300 Monte Carlo realisations. On the other hand, for each model (biSBM and DCbiSBM) we denote $\hat{\boldsymbol{\theta}}$ (resp. $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}$) the estimated parameters and compute the

expectations of the robustness under each model. For the biSBM, we also supply their variance and define the ratio

$$Z_R(A) = \mid \widehat{\overline{R}}_{\mathbb{U}}(A) - \overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{U}} \mid / \sqrt{\mathbb{V}_{\hat{\boldsymbol{\theta}},n}(\,\mathbb{E}_{\mathbb{U}}[\,\overline{R}(A,S)|A\,]\,)}$$

which is the number of standard deviations separating the two computed robustness statistics. This quantity helps assess the goodness of fit of the biSBM to the network A with respect to the robustness statistic.

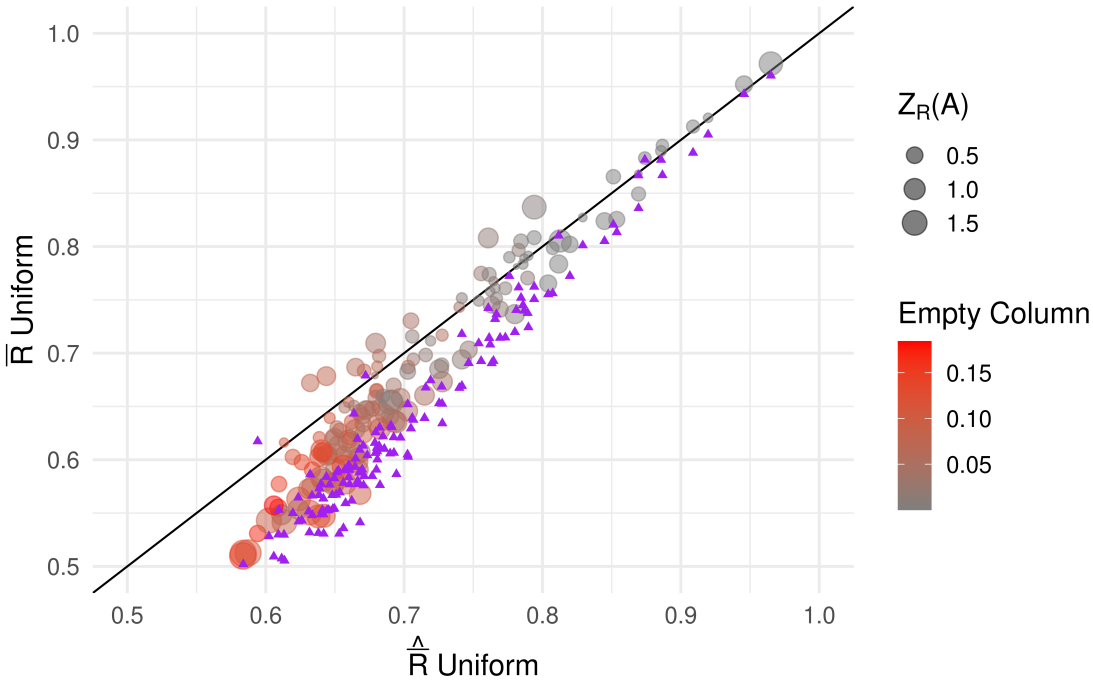


Figure 4.4 – Expected robustness statistic ($\overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{U}}$) under a biSBM (circles) and DCbiSBM (purple triangles) as a function of the standard robustness ($\widehat{\overline{R}}_{\mathbb{U}}(A)$) for uniform extinction sequences. The grey to red gradient stands for $(1-\hat{d})^{n_r}$, the probability to have an empty column under an Erdős-Rényi distribution. The size of the point is related to $Z_R(A)$.

biSBM versus DCbiSBM In Figure 4.4, we plot the points $(\widehat{\overline{R}}_{\mathbb{U}}(A), \overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{U}})$ for the two models (circles are for biSBM and purple triangles for DCbiSBM). For the biSBM, the color of the point depends on $(1-\hat{d})^{n_r}$ which is the estimated probability to observe a column with no interaction. The observed robustness indicators range from 0.5 to 1. The points roughly follow the identity line for the biSBM (circles) whereas the expected version seems systematically smaller than the empirical robustness for the DCbiSBM (purple triangles). We comment further on these points hereafter.

- Under a biSBM, the smaller $(1-\hat{d})^{n_r}$, the smaller $\mid \widehat{\overline{R}}_{\mathbb{U}}(A) - \overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{U}} \mid$ and $Z_R(A)$ (the number of standard deviations). When $(1-\hat{d})^{n_r}$ is large (red dots), the

expected version of the robustness under a biSBM underestimates the standard robustness (red dots at the bottom left of the plot). This phenomenon may be explained as follows. By construction, ecological networks only involve species that have been seen at least one time in interaction (species with no interaction were removed). As a consequence, since the biSBM does not take into account this phenomenon, the space of networks we integrate over when computing the robustness statistic under a biSBM is too large. This remark is especially true for small networks and highlights the limitation of the biSBM model for small ecological networks.

- On 127 networks (over the 136), the DCbiSBM selects 1 block, encoding solely, as the EDD model, the degree distribution and the size of the networks in their parameters. On these networks, the expectation of the robustness statistic under a DCbiSBM is smaller than the empirical robustness most of the time (purple triangles). We can conclude that the DCbiSBM seems to be unable to capture the additional structure beyond their degree distribution that makes them more robust to uniform random extinction than expected.

This first study highlights the limitation of the DCbiSBM model to mimic ecological networks (hence, hereafter, we focus our analysis solely on the biSBM). On the contrary, for not too small networks, the expected version of the robustness with biSBM supplies coherent robustness values with the empirical indicator $\widehat{\overline{R}}_{\mathbb{U}}(A)$. This new version of the robustness has the advantage to arise in a closed-form and a quantification of its variance is available. This comparison has been made for $\mathbb{S} = \mathbb{U}$, we now consider more elaborate extinction sequence distributions.

About $\mathbb{S}$. We compute the robustness statistics for decreasing and increasing primary extinction sequences on the degrees with the two methods described in Section 4.2: the one which strictly depends on the order of the degrees ($\widehat{\overline{R}}_{\mathbb{D}^{\uparrow}_{ord}}(A)$ and $\widehat{\overline{R}}_{\mathbb{D}^{\downarrow}_{ord}}(A)$) and the one where the nodes are weighted by a linear function of the degrees as in Equation (4.5) with $\alpha = 1$ ($\widehat{\overline{R}}_{\mathbb{D}^{\uparrow}_{lin}}(A)$ and $\widehat{\overline{R}}_{\mathbb{D}^{\downarrow}_{lin}}(A)$). We examine the fit with $\overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{S}}$ where $\mathbb{S} \in \{\mathbb{B}^{\downarrow}, \mathbb{B}^{\uparrow}\}$ mimics the degree increasing and decreasing extinction sequences. The comparison are plotted in Figure 4.5.

When considering primary extinction sequences by decreasing connectivity, there is a positive bias of $\overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{B}^{\downarrow}}$ on $\widehat{\overline{R}}_{\mathbb{D}^{\downarrow}_{ord}}(A)$ (top-left) which is attenuated when considering $\widehat{\overline{R}}_{\mathbb{D}^{\downarrow}_{lin}}(A)$ (bottom-left). For primary extinction sequences by increasing connectivity, the empirical robustness is highly dependent on the degrees of the most connected species, hence the fit of $\overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{B}^{\uparrow}}$ on $\widehat{\overline{R}}_{\mathbb{D}^{\uparrow}_{ord}}(A)$ is very poor with a negative bias (top-right). While the negative bias still remains, the fit on $\widehat{\overline{R}}_{\mathbb{D}^{\uparrow}_{lin}}(A)$ is correct for a much higher fraction of the networks (bottom-right).

As a conclusion, the degree decreasing sequences standardly used in the ecological field can be easily reproduced with our block decreasing connectivity sequences,

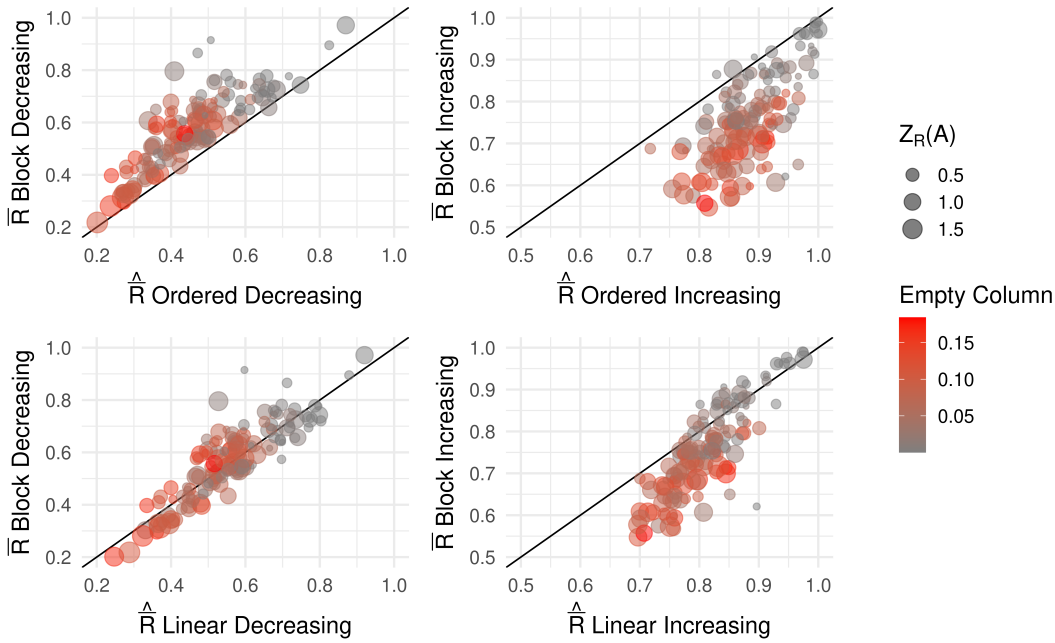


Figure 4.5 – $\overline{R}_{\hat{\boldsymbol{\theta}},n,\mathbb{S}}$ as a function of the classical robustness $\widehat{\overline{R}}_{\mathbb{D}}(A)$ for various $\mathbb{S}$. Block Decreasing (resp. increasing) $= \mathbb{B}^{\downarrow}$ (resp $\mathbb{B}^{\uparrow}$), Ordered Decreasing (resp. Increasing) $= \mathbb{D}^{\downarrow}_{ord}$ (resp. $\mathbb{D}^{\uparrow}_{ord}$). Linear Decreasing (resp. Increasing) $= \mathbb{D}^{\downarrow}_{lin}$ (resp. $\mathbb{D}^{\uparrow}_{lin}$). The grey to red gradient stands for $(1-\hat{d})^{n_r}$, the probability to have an empty column under an Erdős-Rényi distribution. The size of the point is the deviation between the biSBM and the empirical robustness in terms of the number of standard deviations of the robustness under a biSBM distribution for $\mathbb{S} = \mathbb{U}$.

leading to quite comparable values. Once again, our expected version of the robustness can be calculated in a closed-form while the empirical robustness relies on a computationally expensive Monte Carlo integration. The degree decreasing extinction sequences are very sensible to the most highly connected species leading to no agreement between the two versions of the robustness.

4.6.2. Correction for Partially Observed Networks

Although the ecological networks are often considered to describe all the possible interaction between species, the sampling may be incomplete (Blüthgen et al., 2008) which may bias the computed network statistics (Rivera-Hutinel et al., 2012) such as the robustness. By relying on a probabilistic model such as the biSBM which can be adjusted to account for the observation process, the sampling effect on the robustness can be corrected. We assume that we could have obtained the *true* interaction network if the sampling effort has been large enough. Instead of this true network, we have only a partially observed interaction network that corresponds to a subset of the true network. We assume that one of the two following frameworks may have generated missing data:

Partially observed species 25% of both the row species and the column species are removed uniformly at random, resulting in the following networks subset: $A^{obs} = \{A_{ij} : i \text{ and } j \text{ are observed}\}$. In this framework, we need to assume that some other data or some expert knowledge gives us the true number of species and the density of the true network. By relying on this expert knowledge, we are able to adjust the parameters of the biSBM.

Partially observed interactions The observation process consists of the observation of interaction on two different transects T_1 and T_2. Since not all the species are present on both transects, the interactions between the species which were not observed on the same transect are labelled as missing and encoded by `NA`s. This results in the following modified incidence matrix:

$$A^{obs} = \begin{bmatrix} & j \in \{T_1 \cap T_2\} & j \in \{T_1 \setminus T_2\} & j \in \{T_2 \setminus T_1\} \\ i \in \{T_1 \cap T_2\} & A_{ij} & A_{ij} & A_{ij} \\ i \in \{T_1 \setminus T_2\} & A_{ij} & A_{ij} & \texttt{NA} \\ i \in \{T_2 \setminus T_1\} & A_{ij} & \texttt{NA} & A_{ij} \end{bmatrix}.$$

More precisely, we consider that on average 50% of the species were observed on both transects while 25% were just observed on one of the two transects and select those species uniformly at random, resulting in 12.5% of missing interactions on average. Note that in this case, we do not need any expert or additional knowledge to have an unbiased estimation of the biSBM parameters. It is sufficient that the missing values are taken into account in the biSBM inference.

We performed a simulation study where we considered 98 networks from the web of life dataset as the *true* interaction networks. We kept the 98 networks where $(1-d)^{n_r} < .1$ and the error in terms of standard deviation is smaller than 1. For each of these 98 networks, we simulated the two observation processes described above 30 times, then we computed the standard robustness statistic and its expectation under a biSBM. For Figure 4.6, we computed the root mean squared error (RMSE) for each network and for different cases (in terms of observation process and extinction sequence) between the robustness computations on the true network (fully observed) and on the partially observed network. When the method for computing the robustness is based on MC computations, we considered the uniform primary extinction sequences (named Monte Carlo in Fig. 4.6) and the increasing and decreasing extinction sequences that depend strictly (named Ordered Monte Carlo in Fig. 4.6) or proportionally as in Equation (4.5) (named Linear Monte Carlo in Fig. 4.6) on the row degrees. More specifically, the RMSE for the methods based on MC computations are computed for any extinction sequences $\mathbb{S} = \{\mathbb{U}, \mathbb{D}^{\uparrow}_{ord}, \mathbb{D}^{\uparrow}_{lin}, \mathbb{D}^{\downarrow}_{ord}, \mathbb{D}^{\downarrow}_{lin}\}$ listed above, as:

$$\sqrt{\frac{1}{30}\sum_{b=1}^{30}(\hat{\bar{R}}_{\mathbb{S}}(A) - \hat{\bar{R}}_{\mathbb{S}}(\tilde{A}_b))^2}$$

for each of the 98 fully observed networks A and where the $\tilde{A}_b$'s are realizations of a partial observation of A. In order to get the robustness under a biSBM, the biSBM was first inferred by taking into account missing data and the robustness statistics

were then computed under the inferred biSBM. We computed the corresponding RMSE for the distributions on extinction sequences $\mathbb{S} \in \{\mathbb{U}, \mathbb{B}^{\uparrow}, \mathbb{B}^{\downarrow}\}$ as:

$$\sqrt{\frac{1}{30}\sum_{b=1}^{30}(\overline{R}_{\hat{\boldsymbol{\theta}},\boldsymbol{n},\mathbb{S}} - \overline{R}_{\tilde{\boldsymbol{\theta}}_b,\boldsymbol{n},\mathbb{S}})^2}$$

where $\hat{\boldsymbol{\theta}}$ is estimated from the fully observed network and the $\tilde{\boldsymbol{\theta}}_b$'s are estimated from partial observations of the network. Note that the numbers of species in row and in column are the same for the fully and the partially observed networks even in the case of missing species since we assumed that we had some additional knowledge which provided us with this information.

The errors in the prediction of the robustness statistics computed from partially observed networks are much smaller when computing the robustness under a biSBM than when using MC computations. Indeed, the missing data can be accounted for in the biSBM inference and the underlying structure of the network can still be recovered from partial information whereas the Monte Carlo simulations are more sensitive to perturbations in the network. The extinction sequences which depend strictly on the degrees are the most impacted by a partial observation of the network. This impact is rather strong although only 12.5% of the information is missing. Note that adding some randomness in the primary extinction sequences by using a distribution which depends linearly on the degrees instead of a strict order has a stabilizing effect.

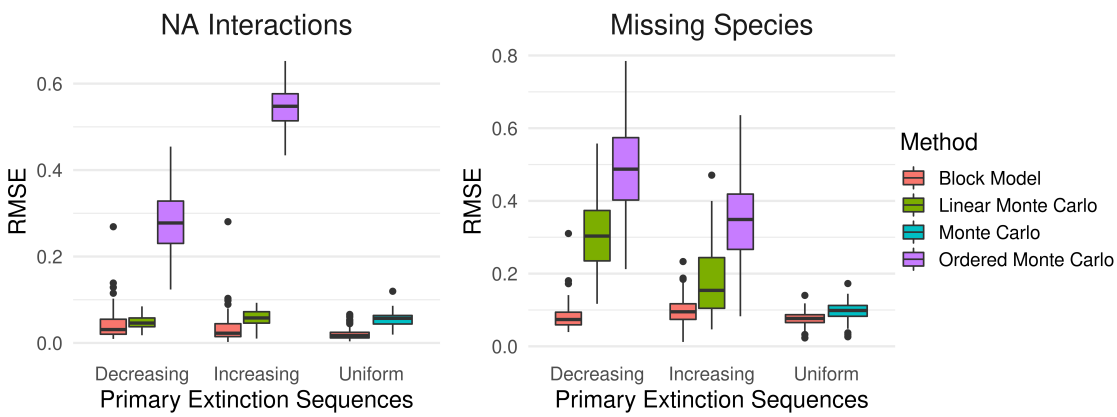


Figure 4.6 – Error (RMSE) in the prediction of the robustness of 98 fully observed networks computed from partially observed networks. On the left, 12.5% of possible interactions are recoded as `NA`. and on the right, 25% of species are missing.

In the framework with partially observed species we assumed that we had additional knowledge giving us access to the true number of species and network density. In practice, this is not always the case. When details about the sampling of networks are available, a few methods exist in the literature to estimate the number of species (Jiménez-Valverde et al., 2006; Gotelli and Colwell, 2011) and similar methods could be used to estimate the density. One could then plug these estimates in the formula

of robustness. If neither the sampling scheme nor the expert knowledge is available, the true number of species remains unknown. Note that in most available ecological interaction networks, the number of species is underestimated as species which have not been seen in interaction with other species are not included in the network. Finally, having access to the number of species, but not to the true number of interactions, leads to an underestimated robustness.

4.7. Discussion

We proposed an expected version of the empirical robustness for bipartite ecological interaction networks by considering a joint model on the network and on the primary extinction sequences. In particular, we obtained a closed and tractable form of the robustness when considering a biSBM as the network model for uniform and by blocks primary extinction sequences. We validated our method by showing that the obtained values were consistent with the empirical robustness classically used in the ecological network community. Having analytical forms allows us to better understand the impact on the robustness of the topology of the network in terms of number of species, density, and mesoscale connectivity patterns. Furthermore, we used the difference between the empirical and expected versions to show that the biSBM is better suited than the DCbiSBM for bipartite ecological networks. This could come from the fact that the DCbiSBM tends to favor community structure above core-periphery structure (Newman, 2018, 14.7.3). The core-periphery structure is indeed very common in bipartite ecological networks which has a strong impact on the robustness.

On real networks, this method can serve as an alternative to the traditional empirical approach, especially when networks are partially observed (because of the inference stability of the biSBM) or when the sampling effort is incomplete (because the model parameters can be easily corrected). This step could be improved with more precise information by considering specific sampling schemes in the model (Tabouy et al., 2019) or by obtaining more details on the sampling in order to better estimate the parameters block by block. Moreover, from the observation of a network, the impact of some hypotheses on the structure might be tested by tuning some parameters of the models and computing the corresponding expected robustness. It may also help study the impact of the structure beyond the effect of the number of species or density when comparing the robustness of several networks: indeed, once the biSBM parameters are estimated, the robustness statistics can be computed by setting the same numbers of species and the same density for all the networks.

Although we developed our method for bipartite networks, this approach can be extended to other types of networks in particular multipartite networks (Pocock et al., 2012) and food webs. Multilayer networks (multiplex or multipartite networks) are gaining a lot of attention with scholars studying ecological networks (Hutchinson et al., 2019). Extending this framework to multipartite networks by including cascading effect between layers is an interesting perspective once data will be more

readily available. The study of robustness for food webs is quite active but requires some ecological insight to properly model the food web as some species are basal species and thus do not prey on other species (in-degree is 0). So food webs which are usually blockmodeled by a directed stochastic block model, might be better handled in our case using a multipartite block model with basal species as a functional group of its own. In this case, it might be important to incorporate rewiring or cascading mechanisms to the modeling of the extinctions. A direction to look at in order to deal with the extinction of basal species and the incorporation of cascading mechanisms is the approach of Bayesian networks to model species extinction in food webs (Eklöf et al., 2013; Häussler et al., 2020).

Lastly, we believe that other ecological indicators could be estimated through a parametric model based approach. Especially, the EDD model and the DCbiSBM seem particularly well suited to study nestedness (see Mariani et al., 2019, for a review) which is a widely used statistic for ecological networks.

Aknowledgement

The authors would like to thank Sonia Kefi, François Massol and Vincent Miele for their helpful advice. This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. This work was partially supported by the grant ANR-18-CE02-0010-01 of the French National Research Agency ANR (project EcoNet).

4.A. Proof for Section 4.4 (Moments of the robustness statistic)

Proposition 4.2. *Let* $(A, S) \sim \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}$, *then*

$$R_{\boldsymbol{\theta},\boldsymbol{n},\mathbb{B}}(m) = 1 - \sum_{q=1}^{Q_c} \rho_q \sum_{n_1+\cdots+n_{Q_r}=n_r} \frac{n_r!}{n_1!\ldots n_{Q_r}!} \prod_{k=1}^{Q_r} \pi_k^{n_k} \left(1-\alpha_{kq}\right)^{\min^+(n_k, \sum_{l\leq k} n_l - m)}, \tag{S-4.19}$$

where $\min^+$ *is the positive part of the minimum function:* $\min^+(x,y) = \max(0, \min(x,y))$.

Proof. Let us assume without loss of generality that the primary extinction sequences go from block 1 to block Q_r. Under this assumption, $\mathbb{P}(S = s|Z) = \frac{\mathbf{1}_{Z_{s(1)}\leq\cdots\leq Z_{s(n_r)}}}{\#\{s: Z_{s(1)}\leq\cdots\leq Z_{s(n_r)}\}}$. Thus, an extinction sequence which does not maintain the block ordering has a null probability. Then, conditioning first by the blocks memberships then the primary extinction sequences:

$$\begin{aligned}
\mathbb{E}_{(A,S)}[R(A,S,m)] &= \sum_{q_{1:n_c}}^{Q_c} \rho_{1:n_c} \sum_{k_{1:n_r}}^{Q_r} \pi_{1:n_r} \mathbb{E}_{(A,S)}[R(A,S,m)|Z_{1:n_r} = k_{1:n_r}, W_{1:n_c} = q_{1:n_c}] \\
&= 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \sum_{q}^{Q_c} \rho_q \sum_{k_{1:n_r}}^{Q_r} \pi_{1:n_r} \mathbb{E}_{(A,S)} \left[\prod_{i=m+1}^{n_r} (1 - A_{S(i)j}) | Z_{1:n_r} = k_{1:n_r}, W_j = q \right] \\
&= 1 - \frac{1}{n_c} \sum_{j=1}^{n_c} \sum_{q}^{Q_c} \rho_q \sum_{k_{1:n_r}}^{Q_r} \pi_{1:n_r} \sum_{s\in\mathfrak{S}_{n_r}} \mathbb{P}(S = s \mid Z_{1:n_r} = k_{1:n_r}) \left(\prod_{i=m+1}^{n_r} (1 - \alpha_{k_{s(i)}q}) \right).
\end{aligned} \tag{S-4.20}$$

We have for all sequences s, s' in the support of $S \mid Z$, that for all i, $Z_{s(i)} = Z_{s'(i)}$. Hence, using the exchangeability properties of the biSBM for species belonging to the same block: $\prod_{i=m+1}^{n_r} \left(1 - \alpha_{k_{s(i)}q}\right) = \prod_{i=m+1}^{n_r} \left(1 - \alpha_{k_{s'(i)}q}\right)$. Furthermore,

$$\prod_{i=m+1}^{n_r} \left(1 - \alpha_{k_{s(i)}q}\right) = \prod_{i=m+1}^{n_r} \prod_{k=1}^{Q_r} (1 - \alpha_{kq})^{\mathbf{1}_{\{k_{s(i)}=k\}}} = \prod_{k=1}^{Q_r} (1 - \alpha_{kq})^{\sum_{i=m+1}^{n_r} \mathbf{1}_{\{k_{s(i)}=k\}}},$$

with for all s such that $k_{s(1)} \leq \cdots \leq k_{s(n_r)}$:

$$\sum_{i=m+1}^{n_r} \mathbf{1}_{\{k_{s(i)}=k\}} = \begin{cases} 0 & \text{if } \sum_{l\leq k} n_l \leq m, \\ \sum_{l\leq k} n_l - m & \text{if } \sum_{l\leq k} n_l - n_k < m < \sum_{l\leq k} n_l, \\ n_k & \text{if } m \leq \sum_{l\leq k} n_l - n_k. \end{cases}$$

This leads to:

$$\begin{aligned}
&\sum_{s\in\mathfrak{S}_{n_r}} \mathbb{P}(S=s\,|Z_{1:n_r}=k_{1:n_r}) \prod_{i=m+1}^{n_r}(1-\alpha_{k_{s(i)}q}) \\
&\qquad = \sum_{s\in\mathfrak{S}_{n_r}} \mathbb{P}(S=s\mid Z_{1:n_r}=k_{1:n_r}) \prod_{k=1}^{Q_r}(1-\alpha_{kq})^{\min^+(n_k,\sum_{l\leq k} n_l - m)} \\
&\qquad = \prod_{k=1}^{Q_r}(1-\alpha_{kq})^{\min^+(n_k,\sum_{l\leq k} n_l - m)}. \qquad \text{(S-4.21)}
\end{aligned}$$

Inputting Equation (S-4.21) in Equation (S-4.20), we obtain the result:

$$\mathbb{E}_{(A,S)}[R(A,S,m)] = 1-\sum_{q=1}^{Q_c}\rho_q \sum_{n_1+\dots+n_{Q_r}=n_r} \frac{n_r!}{n_1!\dots n_{Q_r}!} \times \prod_{k=1}^{Q_r}\pi_k^{n_k}(1-\alpha_{kq})^{\min^+(n_k,\sum_{l\leq k} n_l - m)}.$$

□

Proposition 4.3. *Let $A \sim biSBM(\boldsymbol{\theta},\boldsymbol{n})$, $\eta_q = 1-\alpha_{+q}$ and $\eta_{qq'} = \sum_{k=1}^{Q_r}\pi_k(1-\alpha_{kq})(1-\alpha_{kq'})$.*

1. *Then:*

$$\begin{aligned}
\mathbb{V}_A[R_{\mathbb{U}}(A,m)] &= \frac{1}{n_c}\sum_{l=m}^{\min(2m,n_r)} \frac{\binom{m}{l-m}\binom{n_r-m}{l-m}}{\binom{n_r}{m}} \sum_{q=1}^{Q_r}\rho_q\eta_q^l - (\sum_{q=1}^{Q_r}\rho_q\eta_q^m)^2 \\
&\quad + \frac{(n_c-1)}{n_c}\sum_{l=m}^{\max(2m,n_r)} \frac{\binom{m}{l-m}\binom{n_r-m}{l-m}}{\binom{n_r}{m}} \sum_{q,q'=1}^{Q_r}\rho_q\rho_{q'}(\eta_q\eta_{q'})^{l-m}\eta_{qq'}^{2m-l}.
\end{aligned}$$

2. *The variance of the robustness statistic due to the network variability under a given biSBM is:*

$$\begin{aligned}
\mathbb{V}_A[\overline{R}_{\mathbb{U}}(A)] &= \frac{1}{n_r^2 n_c^2} n_c \sum_{m,m'=0}^{n_r}\sum_{l=\max(m,m')}^{\min(m+m',n_r)} \frac{\binom{m}{l-m'}\binom{n_r-m}{l-m}}{\binom{n_r}{m'}} \sum_{q=1}^{Q_r}\rho_q\eta_q^l - (\frac{1}{n_r}\sum_{m=0}^{n_r}\sum_{q=1}^{Q_r}\rho_q\eta_q^m)^2 \\
&\quad + \frac{1}{n_r^2 n_c^2} n_c(n_c-1)\sum_{m,m'=0}^{n_r}\sum_{l=\max(m,m')}^{\min(m+m',n_r)} \frac{\binom{m}{l-m'}\binom{n_r-m}{l-m}}{\binom{n_r}{m'}} \sum_{q,q'=1}^{Q_r}\rho_q\rho_{q'}\eta_q^{l-m'}\eta_{q'}^{l-m}\eta_{qq'}^{m+m'-l}.
\end{aligned}$$

Proof. The calculus of the following proof relies on the following lemma:

Combinatorial lemma 4.1. *Let m, m' the first terms of two permutations of $\mathfrak{S}_n$, the the proportion of couple of permutation (s,s') that have exactly l unique terms is:*

$$\frac{\#\{(s,s')\colon s(1:m)\cup s'(1:m')=l\}}{\#(\mathfrak{S}_n,\mathfrak{S}_n)} = \begin{cases} \frac{\binom{m}{m+m'-l}\binom{n-m}{l-m}}{\binom{n}{m'}} & \text{if } \max\{m,m'\}\leq l\leq \min\{m+m',n\} \\ 0 & \text{otherwise} \end{cases},$$

where $s(1:m)=\{s(1),\dots,s(m)\}$.

Proof. There are $n!$ permutations of size n. For the first permutation of size n, we look at the first m . In order to create the second permutation we must among the first m' terms take:

- $l - m$ terms that are not common with the first permutation (among $n - m$)
- $m + m' - l$ terms that are common with the first permutation (among m).

Those m' terms can be reordered into $m'!$ possible arrangements and the $n - m'$ resting terms into $n - m'!$ permutations, giving:

$$\#\{(s, s') \colon s(1:m) \cup s'(1:m') = l\} = n! \binom{m}{m + m' - l} \binom{n-m}{l-m} m'!(n - m')!$$

We then divide by the $(n!)^2$ couple of permutations possible.

$$\frac{\#\{(s, s') \colon s(1:m) \cup s'(1:m') = l\}}{\#(\mathfrak{S}_n, \mathfrak{S}_n)} = \frac{\binom{m}{m+m'-l}\binom{n-m}{l-m}}{\binom{n}{m'}}$$

The bound on l is straightforward. □

We first prove the result for the robustness statistic. The variance of the robustness function is a straightforward derivation from it.

$$\mathbb{V}_A(\,\mathbb{E}_S[\,\overline{R}(A, S))|A\,]\,) = \underbrace{\mathbb{E}_A[\,\mathbb{E}_S^2[\,1 - \overline{R}(A, S)|A\,]\,]}_{B} - \underbrace{\mathbb{E}_A^2[\,\mathbb{E}_S[\,1 - \overline{R}(A, S)|A\,]\,]}_{C} \quad \text{(S-4.22)}$$

Using Equation (4.12) and Equation (4.4), we have: $C = \left(\frac{1}{n_r}\sum_{m=0}^{n_r}\sum_{q=1}^{Q_c}\rho_q\eta_q^{n_r-m}\right)^2$.

We divide B into 2 terms, based on the column index:

$$
\begin{aligned}
B &= \mathbb{E}_A\left[\,(1-\frac{1}{n_r}\sum_{m=0}^{n_r-1}\mathbb{E}_S[\,R(A,S,m)|A]\,)^2\right] \\
&= \frac{1}{n_r^2}\sum_{m,m'=0}^{n_r-1}\mathbb{E}_A\left[\,1-\mathbb{E}_S[\,R(A,S,m)|A]\cdot\mathbb{E}_S[\,1-R(A,S,m')|A]\,\right] \\
&= \frac{1}{n_r^2}\sum_{m,m'=0}^{n_r-1}\mathbb{E}_A\left[\left(\frac{1}{n_r!}\sum_{s\in\mathfrak{S}_{n_r}}\frac{1}{n_c}\sum_{j=1}^{n_c}\mathbf{1}_{\{\sum_{i=m+1}^{n_r}A_{s(i)j}=0\}}\right)\right. \\
&\qquad\left.\cdot\left(\frac{1}{n_r!}\sum_{s'\in\mathfrak{S}_{n_r}}\frac{1}{n_c}\sum_{j'=1}^{n_c}\mathbf{1}_{\{\sum_{i=m'+1}^{n_r}A_{s'(i)j'}=0\}}\right)\right] \\
&= \frac{1}{n_r^2}\sum_{m,m'=0}^{n_r-1}\frac{1}{n_c^2}\sum_{j\neq j'=1}^{n_c}\underbrace{\frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\mathbb{E}_A\left[\mathbf{1}_{\{\sum_{i=m+1}^{n_r}A_{s(i)j}=0\}}\mathbf{1}_{\{\sum_{i=m'+1}^{n_r}A_{s'(i)j'}=0\}}\right]}_{B1} \\
&\quad+\frac{1}{n_r^2}\sum_{m,m'=0}^{n_r-1}\frac{1}{n_c^2}\sum_{j=1}^{n_c}\underbrace{\frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\mathbb{E}_A\left[\mathbf{1}_{\{\sum_{i=m+1}^{n_r}A_{s(i)j}=0\}}\mathbf{1}_{\{\sum_{i=m'+1}^{n_r}A_{s'(i)j}=0\}}\right]}_{B2}.
\end{aligned}
\tag{S-4.23}
$$

To compute $B1$, we separate the set of the union of the extinction sequences into 3 disjoint sets in order to distribute the mixture parameter π:

$$
\begin{aligned}
B1 &= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\mathbb{E}_{\pi,\rho}\left[\mathbb{E}_\alpha\left[\mathbf{1}_{\{\sum_{i=m+1}^{n_r}A_{s(i)j}=0\}}\mathbf{1}_{\{\sum_{i=m'+1}^{n_r}A_{s'(i)j'}=0\}}|Z,W\right]\right] \\
&= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\mathbb{E}_{\pi,\rho}\left[\prod_{i=m+1}^{n_r}(1-\alpha_{Z_{s(i)}W_j})\prod_{i=m'+1}^{n_r}(1-\alpha_{Z_{s'(i)}W_{j'}})\right] \\
&= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\sum_{q,q'}\rho_q\rho_{q'}\sum_{k_{1:n_r}}\pi_{k_{1:n_r}}\prod_{i=m+1}^{n_r}(1-\alpha_{k_{s(i)}q})\prod_{i=m'+1}^{n_r}(1-\alpha_{k_{s'(i)}q'}) \\
&= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\sum_{q,q'}\rho_q\rho_{q'}\sum_{k_{1:n_r}}\pi_{k_{1:n_r}}\prod_{i\in s(m+1:n_r)\setminus s'(m'+1:n_r)}(1-\alpha_{k_iq}) \\
&\qquad\prod_{i\in s'(m'+1:n_r)\setminus s(m+1:n_r)}(1-\alpha_{k_iq'})\prod_{i\in s(m+1:n_r)\cap s'(m'+1:n_r)}(1-\alpha_{k_iq})(1-\alpha_{k_iq'}) \\
(Lem.4.1) &= \sum_l\frac{\binom{m}{m+m'-l}\binom{n-m}{l-m}}{\binom{n}{m'}}\sum_{q,q'}\rho_q\rho_{q'}\eta_q^{l-(n_r-m')}\eta_{q'}^{l-(n-m)}\eta_{qq'}^{2n-m-m'-l}
\end{aligned}
$$

For $B2$, the column indices are the same, hence some entries of A are taken twice:

$$\begin{aligned}
B2 &= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\sum_q \rho_q \sum_{k_{1:n}} \pi_{k_{1:n}} \mathbb{E}_\alpha\Big[\mathbf{1}_{\{\sum_{i\in s(m+1:n_r)} A_{ij}=0\}} \\
&\qquad \cdot \mathbf{1}_{\{\sum_{i'\in s'(m'+1:n_r)} A_{i'j}=0\}} | Z_{1:n_r}=k_{1:n_r}, W_j=q\Big] \\
&= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\sum_q \rho_q \sum_{k_{1:n}} \pi_{k_{1:n}} \\
&\qquad \cdot \mathbb{E}_\alpha\Big[\mathbf{1}_{\{\sum_{i\in s(1:n_r-m)\cup s'(1:n_r-m')} A_{ij}=0\}} | Z_{1:n_r}=k_{1:n_r}, W_j=q\Big] \\
&= \frac{1}{n_r!n_r!}\sum_{s,s'\in\mathfrak{S}_{n_r}}\sum_q \rho_q \sum_{k_{1:n}} \pi_{k_{1:n}} \prod_{i\in s(1:n_r-m)\cup s'(1:n_r-m')} (1-\alpha_{k_i q}) \\
\text{(Lem. 4.1)} &= \sum_l \frac{\binom{m}{m+m'-l}\binom{n-m}{l-m}}{\binom{n}{m'}} \sum_q \rho_q \eta_q^l
\end{aligned}$$

Finally we use the symmetry between m and $n_r - m$ and input first $B1$ and $B2$ into Equation (S-4.23) then B and C into Equation (S-4.22) to obtain the variance of the robustness statistic.

To obtain the variance of the robustness function, we fix m, and set $m = m'$ in Lemma 4.1, and Equation (S-4.23).

□

4.B. Comparing the robustness of bipartite ecological networks with different interaction types

In this appendix we aim at understanding the link between the robustness and different basic statistics such as the number of species and the network density on the networks issued from the Web of Life database presented in Section 4.6. Some of these links has been studied theoretically on Section 4.5 of this chapter. We will illustrate how the package `robber` may help us to compare the robustness of a collection of networks.

4.B.1. Robustness and Analysis of the block model of the Web of Life dataset

The collection we consider is comprised of 3 types of interaction bipartite networks encoded as :

A_HP for hosts-parasites with parasitism (antagonistic) interactions,

M_PL for plants-pollinators with pollination (mutualistic) interactions,
M_SD for plants-birds networks with seed dispersal (mutualistic) interactions.

We load the `robber` packages and the dataset which includes the 199 networks database.

```
library(robber)
data("web_of_life")
```

Then we can compute the empirical robustness statistic and the one by *biSBM* for each networks, we give an example for uniform primary extinction sequences. To do so, we must first obtain the list of the *biSBM* parameters for each network:

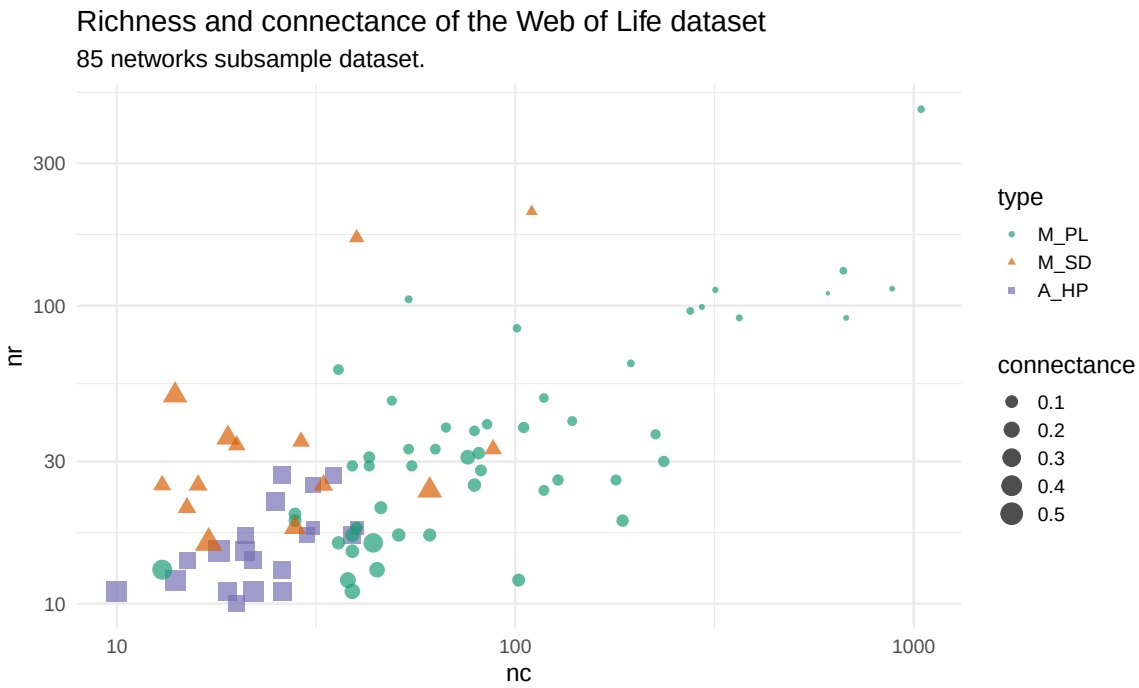
```
library(tidyverse)
rob_emp <- lapply(
  web_of_life,
  function(wol) robustness_emp(A = wol$net,
                               ext_seq = "uniform")$auc)
lbm_param <- lapply(web_of_life,
                    function(wol) get_lbm_param(wol$net))
rob_lbm <- lapply(
  lbm_param,
  function(param) do.call(param,
                          robustness_lbm,
                          ext_seq = "uniform")$auc)
```

Among those networks, we keep 85 of them, the ones with $n_r \geq 10$, $n_c \geq 10$ and with $Q_r > 1$ and $Q_c > 1$ after fitting a *biSBM*.

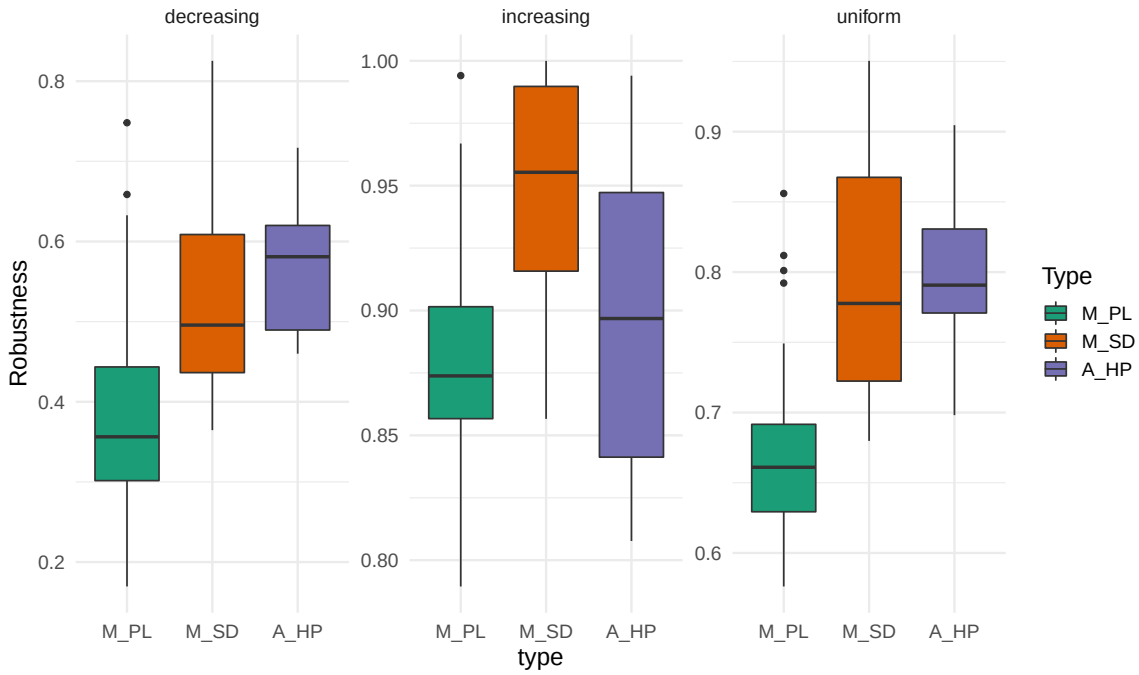
```
index10ER <- which(map_dbl(web_of_life, "nr") >= 10 &
                   map_dbl(web_of_life, "nc") >= 10  &
                   map_dbl(lbm_param, "QR") > 1 &
                   map_dbl(lbm_param, "QC") > 1)
```

4.B.2. Analyzing the link between richness, connectance and interaction type for empirical robustness

As can be seen in the next figure, the richness and connectance of the networks are highly dependent on the type of interaction. Parasitism networks tend to be small with high connectance and have a squared shape incidence matrix ($n_r \approx n_c$), while pollination networks tend to have much more pollinator species than plant species ($n_c >> n_r$), also the largest networks have a low connectance. Finally seed-dispersal networks may have more plant species than bird species ($n_r > n_c$).



In the next figure, we show boxplots of the empirical robustness for the three interaction types and three distributions of the primary extinction sequences ($\mathbb{D}^{\downarrow}_{ord}, \mathbb{D}^{\uparrow}_{ord}, \mathbb{U}$). We notice that pollination networks have a much lower robustness to plant extinctions than seed-dispersal networks. Also, the distribution of the robustness of parasitism networks does not seem to vary much depending on the extinction sequence. We will explore this further by fitting linear models on these statistics.



Linear Model for Empirical Robustness

From now on, we just keep 71 networks. Some networks come from time series, where we just keep the most complete network for each time series. Others were outliers on the linear model described below, breaking the Gaussian linear model assumptions. We note by `index10ERno` this new subset of networks. We make multiple linear regression by fitting a linear model with 5 covariates, which are the species richnesses: n_r and n_c the connectance of the network d and the probability to have an empty row $(1-d)^{n_c}$ or column $(1-d)^{n_r}$ under an Erdős-Rényi model to explain one robustness statistic at a time. The response variables are the empirical robustness statistic with $\mathbb{S} \in \{\mathbb{D}^{\downarrow}_{ord}, \mathbb{D}^{\uparrow}_{ord}, \mathbb{U}\}$ as the distribution of the primary extinction sequences. In order to assess the effect of the type of interactions beyond these basic statistics, we fit an ancova model were we add the type of interactions as a qualitative covariate.

```
fit.lm <- lm(robustness ~ connectance + nr + nc  +
             empty_row + empty_col,
            subset = index10ERno,data = tb_emp_stat)
fit.ancova <- lm(robustness ~ type*(connectance + nr + nc  +
             empty_row + empty_col),
            subset = index10ERno,data = tb_emp_stat)
```

The link between the basic statistic and the empirical robustness is significant for all 3 primary extinction sequence distributions (not shown here). In Table 4.1 we show for each type of robustness statistic, an analysis of variance table of the multiple linear model versus the ancova model, testing for the added effect of the type of interaction. We see that the adjusted R^2 is very high ($\approx .9$) for the multiple linear regression model for $\mathbb{S} = \mathbb{U}$. This show further evidence that the empirical robustness usually computed is well explained by the connectance and species richness of the network. This is also true for $\mathbb{S} = \mathbb{D}^{\downarrow}_{ord}$, but not so much for $\mathbb{D}^{\uparrow}_{ord}$. We think that the last results is due to the importance of the most connected species in the robustness, an information which is not directly included in the covariates we used. When we test the two nested models, we see that adding the effect of the interaction type is significant for $\mathbb{S} \in \{\mathbb{D}^{\uparrow}_{ord}, \mathbb{U}\}$, but not for $\mathbb{S} = \mathbb{D}^{\downarrow}_{ord}$.

Table 4.1 – Analysis of Variance Table for Empirical Robustness

Extinction	Res.Df	RSS	adj.R^2	F	Pr(>F)
Uniform	76	0.0708	0.892	NA	NA
Uniform	64	0.03208	0.938	6.438	2.413e-07
Increasing	76	0.1603	0.303	NA	NA
Increasing	64	0.1052	0.421	2.795	0.003991
Decreasing	76	0.3128	0.778	NA	NA
Decreasing	64	0.2583	0.768	1.124	0.3571

4.B.3. Normalized robustness

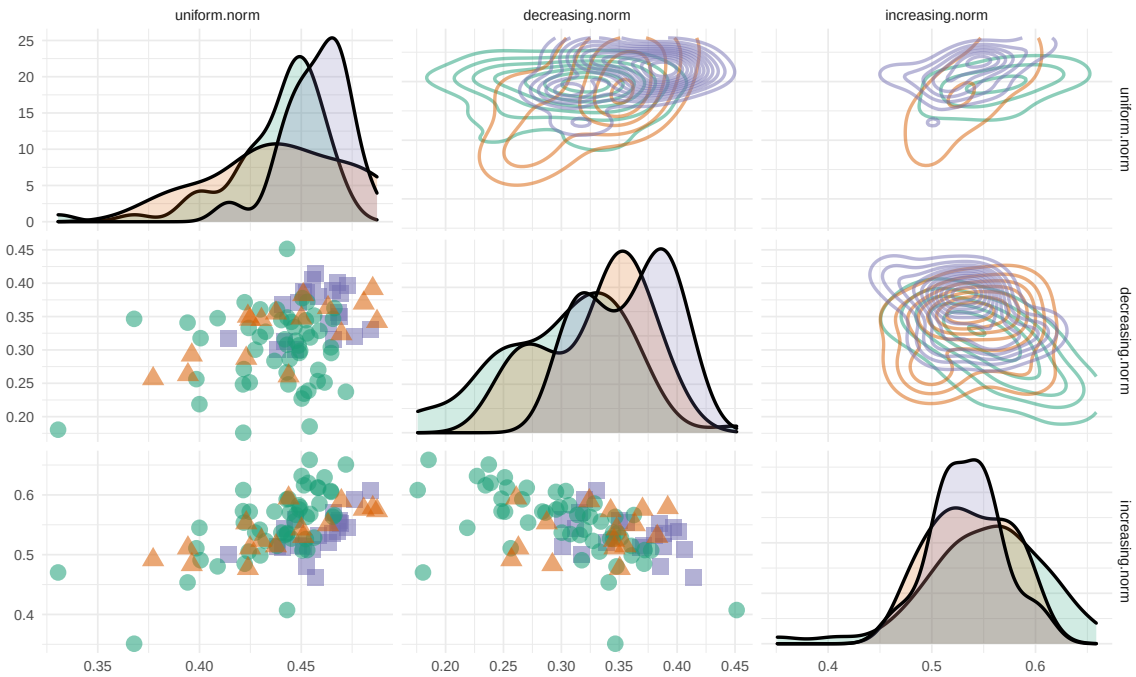
In order to try removing the effect of species richness and connectance from the robustness and to understand which networks have a more robust structure we can reparametrized the parameters of the $biSBM$ so that each network has the same new richness $\tilde{\mathbf{n}}$ and connectance $\tilde{d}$, by doing:

$$\tilde{n}_r = \tilde{n}_c = 100 \quad \tilde{\delta}_{kq} = \frac{.0156}{\hat{d}}\hat{\delta}_{kq}, \quad \forall (k,q) \in \{1,\dots,\hat{Q}_r\} \times \{1,\dots,\hat{Q}_c\},$$

where $\tilde{d} = .0156$ has been chosen so that the robustness statistic under an Erdős-Rényi model is about .5. We then compute the normalized $biSBM$ robustness statistic with $\mathbb{S} \in \{\mathbb{B}^{\downarrow}, \mathbb{B}^{\uparrow}, \mathbb{U}\}$ as the distribution of the primary extinction sequences.

```
tb_robustness_norm <-
  tibble("uniform.norm" = unlist(compare_robustness(lbm_param)),
         "decreasing.norm" = unlist(compare_robustness(lbm_param,
            ext_seq = "decreasing")),
         "increasing.norm" = unlist(compare_robustness(lbm_param,
            ext_seq = "increasing")))
```

We show above that the empirical robustness is highly dependant on the richness and connectance. Also, there seems to have an additional effect of the type of interaction for decreasing and uniform extinction sequences. Using the $biSBM$ robustness with normalized parameters should remove a part of this dependence. On the next figure, we plot the normalized biSBM robustness for all 3 types of extinction sequences. The shape of the distribution seems to differ depending on the type of interaction but the center of the distribution looks close.



Linear Model for Normalized *biSBM* Robustness

We fit linear models using the same covariates as for the empirical robustness (before normalization) but putting the normalized robustness statistics as the response variables instead. We still expect some effects of the connectance and richness on the normalized robustness, as larger and/or sparser networks may have different structures than smaller and/or denser ones.

```
fit <- lm(uniform.norm ~ connectance + nr + nc  +
             empty_row + empty_col,
            subset = index10ERno,data = tb_norm_stat)
fit_type <-
  lm(uniform.norm ~ type*(connectance + nr + nc  +
             empty_row + empty_col),
            subset = index10ERno,data = tb_norm_stat)
```

Let first notice in Table 4.2 that the adjusted R^2s are much lower than the ones for the models explaining the empirical robustness. The additional effect of the interaction type is still present for the uniform primary extinction sequences but for the extinction sequences by block it is much harder to conclude ($p - value \approx .07$).

Table 4.2 – Analysis of Variance Table for Normalized *biSBM* Robustness

Extinction	Res.Df	RSS	adj.R^2	F	Pr(>F)
Uniform	76	0.0224	0.452	NA	NA
Uniform	64	0.01415	0.562	3.11	0.00161
Increasing	76	0.1207	0.283	NA	NA
Increasing	64	0.09094	0.358	1.748	0.07715
Decreasing	76	0.1184	0.498	NA	NA
Decreasing	64	0.08885	0.523	1.771	0.07241

4.B.4. Examining if networks are more robust to plant or animal extinctions

We can also compute the model parameters for the transposed networks, ie. the one for which primary extinctions occur on the column species and look at their impact on the row species. This makes sense for multualistic networks, e.g. pollination networks, as the extinction of a plant has an impact on the pollinators but also the extinction of a pollinator has an impact on the ability of a plant to be pollinized. Let $(\mathbf{n}^\intercal, \boldsymbol{\theta}^\intercal)$ be the paramters of the transposed model, then:

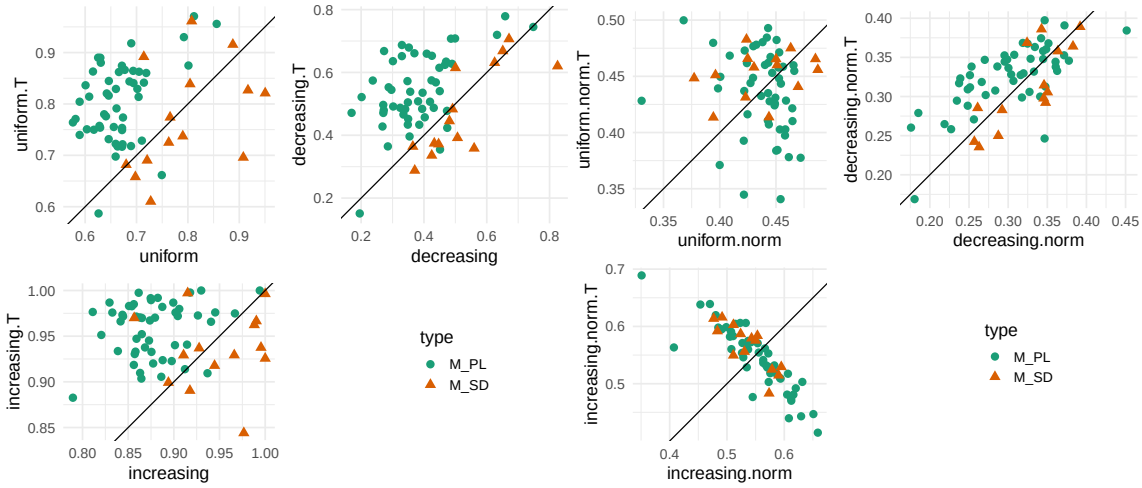
$$\delta^\intercal_{qk} = \delta_{kq}, \quad \rho^\intercal_k = \pi_k, \quad \pi^\intercal_q = \rho_q \text{ and } n^\intercal_c = n_r, \; n^\intercal_r = n_c.$$

The normalized robustness of the transposed networks could be easily compute with

the `robber` package as follows:

```
lbm_param.T <- lapply(lbm_param, function(par)
   list(con = t(par$con), pi = par$rho, rho = par$pi))
tb_robustness_norm.T <-
  tibble("uniform.norm.T" = unlist(compare_robustness(lbm_param.T)),
         "decreasing.norm.T" = unlist(compare_robustness(lbm_param.T,
             ext_seq = "decreasing")),
         "increasing.norm.T" = unlist(compare_robustness(lbm_param.T,
             ext_seq = "increasing")))
```

In the next figure, we notice on the left a clear asymmetry for pollination networks in the empirical robustness depending on wether the primary extinction sequences is on the plants or on the pollinators. This is an effect of the asymmetry in the species richness, ie. in the number of plants and pollinators in these networks. In most of them, there are more pollinators than plants ($n_c >> n_r$) and the networks are more robust to pollinator extinction than to plant extinction. When looking at the normalized robustness, we notice that networks are evenly distributed around the first bissectrice, except for decreasing by block primary extinction sequences where there seems to be a slight bias for pollination networks.



Chapter 5

Conclusions et perspectives

Cette thèse est divisée en trois travaux principaux. Dans un premier chapitre, consacré aux réseaux multiniveaux, nous avons développé un SBM adapté à ce type de données. Dans un autre travail, nous avons proposé des extensions du SBM adaptées à la modélisation conjointe d'une collection de réseau. Un des objectifs était de trouver une structure à l'échelle mésoscopique commune aux réseaux de la collection, et de partitionner la collection en groupes de réseaux ayant la même structure méso-échelle. Enfin, dans le dernier chapitre nous nous sommes intéressés à la relation entre la robustesse d'un réseau bipartite et sa structure. En particulier, dans le cas où le réseau est supposé être la réalisation d'un SBM bipartite, nous avons réussi à exprimer la robustesse du réseau en une fonction des paramètres de ce SBM.

Dans ce qui suit, je propose dans un premier temps, quelques pistes dans la continuation des travaux sur la modélisation jointe de réseaux. Puis, je me concentrerai sur des extensions liées à la robustesse et sur l'intêret d'utiliser la modélisation jointe lorsque nous étudions une collection de plusieurs réseaux.

Motivés par des applications en sociologie et en écologie, les travaux sur les réseaux multiniveaux et les collections de réseaux ont montré l'intérêt de prendre en compte les réseaux collectivement plutôt qu'individuellement. En particulier, pour comprendre si la modélisation conjointe était pertinente nous avons raisonné en termes respectivement de dépendance entre structures de connections des différents niveaux pour les réseaux multiniveaux et de similarité entre les réseaux pour les collections de réseaux. Ces dépendances et similarités peuvent être expliquées par d'autres facteurs que la structure des interactions. Ainsi, intégrer des covariables (Mariadassou et al., 2010) permettrait de mieux séparer ce qui relève proprement de la structure du réseau, de ce qui peut être expliqué par des caractéristiques extérieures.

De plus, la dépendance et les changements de structure de connexion gagneraient à être considérés de manière plus locale. Je propose ci-dessous quelques pistes incorporant ces notions pour analyser des réseaux multi-omiques.

Modélisation d'une collection de réseaux multi-omiques

Accroître la vitesse et l'efficacité de la sélection végétale par l'utilisation de données omiques est l'un des impératifs pour résoudre les problèmes alimentaires dans le cadre du changement climatique (Scossa et al., 2021). La prédiction génomique, qui met en relation le génome et le phénotype, permet de rationaliser la sélection végétale. En

outre, l'utilisation d'autres données omiques telles que le métabolome, l'ionome et le transcriptome peut non seulement améliorer la précision de la prédiction, mais aussi révéler les mécanismes génétiques et physiologiques qui sous-tendent la prédiction.

L'étude de la structure des relations entre les éléments de différentes couches omiques, comme l'ionomique et la métabolomique, peut fournir des indices pour comprendre les mécanismes à l'origine des prédictions. Dans la suite, je propose une nouvelle méthodologie adéquate pour analyser ces données omiques internes et identifier où des changements significatifs dans leur structure se produisent en modélisant l'influence génétique et environnementale sur la structure du réseau.

Un réseau multi-omique peut se modéliser par un SBM pour réseaux multipartites (Bar-Hen et al., 2020) permettant de définir l'existence d'une classification spécifique entre les nœuds d'un réseau. Par exemple, il est possible de s'assurer que les nœuds ionomiques et métabolomiques n'appartiennent pas au même bloc, ce qui est adapté à la modélisation des réseaux inter- et intra-multi-omiques.

Je souhaite étendre ce modèle au cas d'une population de réseaux multiomiques où le comportement des différents blocs dépend de covariables environnementales (stress de sécheresse...) et phénotypiques. Comme proposé par Pavlović et al. (2020) pour des données d'imagerie IRMf, dans le cadre d'un SBM, il est possible d'identifier les blocs dont le comportement change en réponse à ces covariables.

Plus formellement, je propose le modèle suivant. Soit $\{X^1, \dots, X^M\}$, une collection de réseau multi-omiques. Chaque réseau indicé par $m \in \{1, \dots, M\}$ est composé de n_m nœuds répartis en K groupes représentant les couches omiques. On fait l'hypothèse que la correspondance des noeuds entre les réseaux est connue et que la répartition des nœuds dans les groupes est la même pour chaque réseau. Alors, pour tout $k \in \{1, \dots, K\}$,

$$\mathbb{P}(Z_i^k = q) = \pi_q, \quad \forall i \in \{1, \dots, n\}, q \in \mathcal{Q}_k = \{1, \dots, Q_k\},$$

tel que $\sum_{q \in \mathcal{Q}_k} \pi_{q_k} = 1$.

Le réseau m est composé d'une suite de matrice $X^{m,k,k'}$, où $(k \leq k') \in \mathcal{K} \subset \{1, \dots, K\}^2$ indiquent les différents types de relations omiques d'intêret. Soit $\mathbf{e}_m$, le vecteur de taille p_e des covariables environnementales et $\mathbf{p}_m$, le vecteur de taille p_p des covariables phénotypiques du réseau m, alors la loi d'une arête conditionnellement à l'appartenance des nœuds aux blocs suit un modèle de régression logistique pour tout $(i \neq j) \in \{1, \dots, n\}^2$:

$$\mathbb{P}(X_{ij}^{m,k,k'} = 1 | Z_i^k = q, Z_j^{k'} = r) = \alpha_{qr}^{m,k,k'}, \quad (q, r) \in \mathcal{Q}_k \times \mathcal{Q}_{k'},$$

où les covariables peuvent agir à différents niveaux de précision sur la connectivité des réseaux.

De manière la plus générale, on introduit de l'hétérogénéité dans la régression en faisant dépendre les paramètres de régression des blocs :

$$\alpha_{qr}^{m,k,k'} = \frac{1}{1 + \exp\left(-(\mathbf{e}_m^\intercal \beta_{qr}^{k,k'} + \mathbf{p}_m^\intercal \gamma_{qr}^{k,k'} + \delta_{q,r}^{k,k'})\right)},$$

où β (resp. γ) est le vecteur des paramètres de la régression pour les covariables environnementales (resp. phénotypiques) et δ est un facteur constant à tous les réseaux.

Les covariables environnementales (resp. phénotypiques) agissent de manière homogène sur l'ensemble du réseau multi-omique :

$$\beta_{qr}^{k,k'} = \beta_0 \quad (\text{resp. } \gamma_{qr}^{k,k'} = \gamma_0), \quad \forall (k, k', q, r) \in \mathcal{K} \times \mathcal{Q}_k \times \mathcal{Q}_{k'},$$

ou bien sur une famille de relation omique particulière :

$$\beta_{qr}^{k,k'} = \beta_0^{k,k'} \quad (\text{resp. } \gamma_{qr}^{k,k'} = \gamma_0^{k,k'}), \quad \forall (q, r) \in \times \mathcal{Q}_k \times \mathcal{Q}_{k'}.$$

Analyser les valeurs de β (resp. γ) peut permettre alors de comprendre quels types d'interactions omiques sont influencées par les covariables environnementales (resp. phénotypiques).

Robustesse de collections de réseaux

Je souhaiterais également étendre les travaux sur les collections de réseaux à des collections de réseaux bipartites et multipartites. L'extension aux réseaux multipartites est un moyen naturel de relier ces travaux à ceux effectués sur la robustesse. Nous avons vu au chapitre 3 que la modélisation par collection de réseaux a un effet stabilisateur sur le clustering obtenu et que considérer des informations provenant d'autres réseaux pouvait améliorer la prédiction d'arêtes, et cela, même sur des réseaux faiblement bruités. L'échantillonnage des réseaux écologiques étant connu pour être incomplet (Blüthgen et al., 2008), il serait intéressant d'analyser l'effet de stabilisation du modèle apporté par la modélisation jointe sur la robustesse par SBM bipartite. Étudier des collections de réseaux bipartites permettrait également en intégrant des covariables pour le types d'interaction, de déterminer l'effet du type d'interaction écologique sur la structure du réseau. Cela nous aiderait enfin à quantifier l'effet du type de l'interaction sur la robustesse.

Lien entre robustesse et autres caractéristiques du réseau Ce type d'analyse peut-être étendu à d'autres indices communément utilisés pour analyser des réseaux. Toujours dans le cadre des réseaux d'interactions écologiques bipartites, Vanbergen et al. (2017) relient, une fois l'effet du nombre d'espèces corrigé, la robustesse de réseaux plantes-pollinisateurs à l'emboîtement, au degré de spécialisation des espèces, à la vulnérabilité du réseau (nombre d'insectes visitant chaque espèce de plantes)... Trouver, à la manière des travaux sur la robustesse du chapitre 4, une expression de ces différents indicateurs sous des modèles de graphes probabilistes, permettrait toujours suivant ces modèles de relier ces différents indices à la structure du réseau et de comprendre comment ils évoluent les uns par rapport aux autres en fonction de cette structure.

Robustesse de réseaux tripartites Les travaux sur la robustesse peuvent également être étendus aux réseaux multipartites. Par exemple, Domínguez-García and Kéfi (2021) ont agrégé une collection de réseaux tripartites, mutualiste–mutualiste, mutualiste–antagoniste et antagoniste–antagoniste, et étudient leurs robustesses. Dans ces réseaux, lorsque les extinctions ont lieu sur un groupe fonctionnel en particulier, il est nécessaire d'incorporer un système de cascade dans les extinctions pour étudier les conséquences de ces extinctions sur des espèces qui ne sont pas directement en interaction avec ce groupe fonctionnel. Je donne ci-dessous quelques détails sur l'incorporation des cascades d'extinctions dans le cadre d'un SBM tripartite. Supposons que les extinctions primaires aient lieu sur la première couche, impliquent des extinctions secondaires sur la seconde couche et que les espèces de la troisième couche ne soient en interaction avec celles de la première couche. Alors, l'impact des extinctions primaires sur la troisième couche ne se fait que via les extinctions secondaires. Alors, on intègre un système de cascade sur les espèces de la troisième couche de la manière suivante. Soit

$$S_j^{(2)}(m) := \mathbb{1}\{\sum_{i=m+1}^{n_1} X_{S(i)j}^{1,2} > 0\},$$

l'indicatrice de l'évènement de non extinction secondaire de l'espèce j après m extinctions primaires. Alors, $\mathbb{P}_{(\mathbf{X},S)}(S_j^{(2)}(m) = 1)$ est directement calculable à partir des résultats sur les réseaux bipartites du chapitre 4. Cela nous donne pour chaque espèce de la seconde couche, une probabilité d'extinction secondaire à partir de laquelle nous pouvons conditionner les extinctions de la dernière couche. Ainsi, en notant

$$S_k^{(3)}(m) := \mathbb{1}\{\sum_{j=1}^{n_2} X_{jk}^{2,3} S_j^{(2)}(m) > 0\},$$

l'indicatrice de l'évènement de non extinction de l'espèce k de la troisième couche après m extinctions primaires, par indépendance des extinctions secondaires,

$$\begin{aligned}\mathbb{P}_{(\mathbf{X},S)}(S_k^{(3)}(m) = 1) \quad = & \sum_{s_{1:n_2}^{(2)} \in \{0,1\}^{n_2}} \mathbb{P}_{\mathbf{X}}(S_k^{(3)}(m) = 1 | S_{1:n_2}^{(2)}(m) = s_{1:n_2}^{(2)}(m)) \\ & \prod_{j=1}^{n_2} \mathbb{P}_{(\mathbf{X},S)}(S_j^{(2)}(m) = s_j^{(2)}(m)).\end{aligned}$$

Dans le cadre d'un SBM tripartite, ces probabilités sont calculables et ne dépendent que des paramètres du modèle comme pour le SBM bipartite. Nous pouvons alors, comme pour les réseaux bipartites, utiliser la robustesse par SBM tripartite pour étudier le lien entre la structure des réseaux et sa robustesse ou bien corriger la robustesse de réseau partiellement observé.

Concernant ce dernier point, la qualité de la correction dépend de la manière de redresser les paramètre du SBM. La méthode que nous avons utilisée dans le cadre des travaux sur la robustesse est naïve (nous avons supposé que les espèces manquantes étaient réparties uniformément entre les blocs et que nous observions une proportion constante des interactions dans chaque bloc du réseau). La thèse d'Emré Anakok, en cours à MIA Paris, porte sur l'estimation des paramètres du SBM à partir de réseaux

incomplets, ce qui permettrait de redresser les paramètres du modèle de manière plus réaliste, et donc d'améliorer notre méthode d'estimation de la robustesse.

Bibliography

[1] Abbe, E. (2016). Community detection and the stochastic block model. *ISIT notes*, pages 1–11.

[2] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9.

[3] Allesina, S. and Pascual, M. (2009). Food web models: a plea for groups. *Ecology Letters*, 12(7):652–662.

[4] Bane, M. S., Pocock, M. J., and James, R. (2018). Effects of model choice, network structure, and interaction strengths on knockout extinction models of ecological robustness. *Ecology and Evolution*, 8(22):10794–10804.

[5] Bar-Hen, A., Barbillon, P., and Donnet, S. (2020). Block models for generalized multipartite networks: Applications in ecology and ethnobiology. *Statistical Modelling*, page 1471082X20963254.

[6] Barabás, G., Pásztor, L., Meszéna, G., and Ostling, A. (2014). Sensitivity analysis of coexistence in ecological communities: theory and application. *Ecology Letters*, 17(12):1479–1494.

[7] Barbillon, P., Donnet, S., Lazega, E., and Bar-Hen, A. (2017). Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(1):295–314.

[8] Bartolucci, F., Marino, M. F., and Pandolfi, S. (2018). Dealing with reciprocity in dynamic stochastic block models. *Computational Statistics & Data Analysis*, 123:86–100.

[9] Bascompte, J., Jordano, P., Melián, C. J., and Olesen, J. M. (2003). The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences*, 100(16):9383–9387.

[10] Baudry, J.-P. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics*, 9(1):1041–1077.

[11] Bianconi, G. (2018). *Multilayer Networks: Structure and Function*. Oxford University Press.

[12] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.

[13] Bickel, P. J., Chen, A., and Levina, E. (2011). The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280 – 2301.

[14] Bickel, P. J., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943.

[15] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

[16] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[17] Blüthgen, N., Fründ, J., Vázquez, D. P., and Menzel, F. (2008). What do interaction network metrics tell us about specialization and biological traits. *Ecology*, 89(12):3387–3399.

[18] Blüthgen, N., Menzel, F., and Blüthgen, N. (2006). Measuring specialization in species interaction networks. *BMC ecology*, 6(1):1–12.

[19] Bollobás, B. (1980). A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316.

[20] Brailly, J. (2016). Dynamics of networks in trade fairs—a multilevel relational approach to the cooperation among competitors. *Journal of Economic Geography*, 16(6):1279–1301.

[21] Brailly, J., Comet, C., Delarre, S., Eloire, F., Favre, G., Lazega, E., Mounier, L., Montes-Lihn, J., Oubenal, M., Penalva-Icher, E., and Piña-Stranger, A. (2017). Neo-structural economic sociology beyond embeddedness. *economic sociology_the european electronic newsletter*, 19(3):36–49.

[22] Brailly, J., Favre, G., Chatellet, J., and Lazega, E. (2016). Embeddedness as a multilevel problem: A case study in economic sociology. *Social Networks*, 44:319–333.

[23] Brault, V. (2014). *Estimation et sélection de modèle pour le modèle des blocs latents*. PhD thesis, Université Paris Sud-Paris XI.

[24] Brault, V., Keribin, C., and Mariadassou, M. (2020). Consistency and asymptotic normality of Latent Block Model estimators. *Electronic Journal of Statistics*, 14(1):1234–1268.

[25] Burgos, E., Ceva, H., Perazzo, R. P., Devoto, M., Medan, D., Zimmermann, M., and Delbue, A. M. (2007). Why nestedness in mutualistic networks? *Journal of Theoretical Biology*, 249(2):307–313.

[26] Cai, Q. and Liu, J. (2016). The robustness of ecosystems to the species loss of community. *Scientific Reports*, 6:35904.

[27] Celeux, G., Fruewirth-Schnatter, S., and Robert, C. P. (2018). Model Selection for Mixture Models - Perspectives and Strategies. *Handbook of Mixture Analysis*, pages 117–154.

[28] Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

[29] Chabert-Liddell, S.-C. (2021a). *MLVSBM: A Stochastic Block Model for Multilevel Networks*. R package version 0.2.2.

[30] Chabert-Liddell, S.-C. (2021b). *robber: Using block model to estimate the robustness of ecological networks*. R package version 0.2.2.

[31] Chabert-Liddell, S.-C., Barbillon, P., and Donnet, S. (2022). Impact of the mesoscale structure of a bipartite ecological interaction network on its robustness through a probabilistic modeling. *Environmetrics*, 32(2):e2709.

[32] Chabert-Liddell, S.-C., Barbillon, P., Donnet, S., and Lazega, E. (2021). A stochastic block model approach for the analysis of multilevel networks: An application to the sociology of organizations. *Computational Statistics & Data Analysis*, 158:107179.

[33] Channarond, A., Daudin, J.-J., and Robin, S. (2012). Classification and estimation in the Stochastic Blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601.

[34] Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

[35] Chen, K. and Lei, J. (2018). Network Cross-Validation for Determining the Number of Communities in Network Data. *Journal of the American Statistical Association*, 113(521):241–251.

[36] Chiquet, J., Donnet, S., and Barbillon, P. (2021). *sbm: Stochastic Blockmodels*. R package version 0.4.3.

[37] Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145.

[38] Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98.

[39] 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.

[40] Corneli, M., Latouche, P., and Rossi, F. (2016). Exact icl maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks. *Neurocomputing*, 192:81–91. Advances in artificial neural networks, machine learning and computational intelligence.

[41] Curtsdotter, A., Binzer, A., Brose, U., de Castro, F., Ebenman, B., Eklöf, A., Riede, J. O., Thierry, A., and Rall, B. C. (2011). Robustness to secondary extinctions: comparing trait-based sequential deletions in static and dynamic food webs. *Basic and Applied Ecology*, 12(7):571–580.

[42] Danon, L., Díaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(9):219–228.

[43] Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and computing*, 18(2):173–183.

[44] De Bacco, C., Power, E. A., Larremore, D. B., and Moore, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317.

[45] de Manincor, N., Hautekèete, N., Mazoyer, C., Moreau, P., Piquot, Y., Schatz, B., Schmitt, E., Zélazny, M., and Massol, F. (2020). How biased is our perception of plant-pollinator networks? a comparison of visit-and pollen-based representations of the same networks. *Acta Oecologica*, 105:103551.

[46] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106.

[47] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

[48] Domínguez-García, V. and Kéfi, S. (2021). The structure and robustness of tripartite ecological networks. *bioRxiv*.

[49] Donnat, C. and Holmes, S. (2018). Tracking network dynamics: A survey using graph distances. *The Annals of Applied Statistics*, 12(2):971–1012.

[50] Doreian, P., Batagelj, V., and Ferligoj, A. (2005). *Generalized blockmodeling*, volume 25. Cambridge university press.

[51] Dormann, C. F., Gruber, B., and Fründ, J. (2008). Introducing the bipartite package: analysing ecological networks. *R News*, 8(2):8–11.

[52] DuBois, C., Butts, C. T., and Smyth, P. (2013). Stochastic blockmodeling of relational event dynamics. In *AISTATS*.

[53] Dunne, J. A., Williams, R. J., and Martinez, N. D. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, 5(4):558–567.

[54] Eklöf, A., Tang, S., and Allesina, S. (2013). Secondary extinctions in food webs: a bayesian network approach. *Methods in Ecology and Evolution*, 4(8):760–770.

[55] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

[56] Favre, G., Brailly, J., Chatellet, J., and Lazega, E. (2016). Inter-organizational network influence on long-term and short-term inter-individual relationships: The case of a trade fair for tv programs distribution in sub-saharan africa. In *Multilevel network analysis for the social sciences*, pages 295–314. Springer.

[57] Fortuna, M. A., Stouffer, D. B., Olesen, J. M., Jordano, P., Mouillot, D., Krasnov, B. R., Poulin, R., and Bascompte, J. (2010). Nestedness versus modularity in ecological networks: two sides of the same coin? *Journal of Animal Ecology*, 79(4):811–817.

[58] Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

[59] Funke, T. and Becker, T. (2019). Stochastic block models: A comparison of variants and inference methods. *PloS one*, 14(4):e0215296.

[60] Gallagher, R. J., Young, J.-G., and Welles, B. F. (2021). A clarified typology of core-periphery structure in networks. *Science Advances*, 7(12):eabc9800.

[61] Ghasemian, A., Hosseinmardi, H., and Clauset, A. (2020a). Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1722–1735.

[62] Ghasemian, A., Hosseinmardi, H., Galstyan, A., Airoldi, E. M., and Clauset, A. (2020b). Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences*, 117(38):23393–23400.

[63] Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

[64] Giordano, G., Ragozini, G., and Vitale, M. P. (2019). Analyzing multiplex networks using factorial methods. *Social Networks*, 59:154–170.

[65] Gotelli, N. J. and Colwell, R. K. (2011). *Estimating species richness*, pages 39–54. Oxford University Press.

[66] Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

[67] Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245.

[68] Grainger, T. N., Levine, J. M., and Gilbert, B. (2019). The invasion criterion: a common currency for ecological research. *Trends in Ecology & Evolution*, 34(10):925–935.

[69] Guimerà, R. and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078.

[70] Han, Q., Xu, K., and Airoldi, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning*, pages 1511–1520. PMLR.

[71] Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354.

[72] Hanneke, S., Fu, W., and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic journal of statistics*, 4:585–605.

[73] Hayashi, K., Konishi, T., and Kawamoto, T. (2016). A tractable fully bayesian method for the stochastic block model. *arXiv preprint arXiv:1602.02256*.

[74] Hayashi, K., Maeda, S. I., and Fujimaki, R. (2015). Rebuilding factorized information criterion: Asymptotically accurate marginal likelihood. *32nd International Conference on Machine Learning, ICML 2015*, 2:1358–1366.

[75] Hileman, J. and Lubell, M. (2018). The network structure of multilevel water resources governance in central america. *Ecology and Society*, 23(2).

[76] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

[77] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137.

[78] Hric, D., Darst, R. K., and Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E*, 90:062805.

[79] Hu, J., Qin, H., Yan, T., and Zhao, Y. (2020). Corrected Bayesian Information Criterion for Stochastic Block Models. *Journal of the American Statistical Association*, 115(532):1771–1783.

[80] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

[81] Hutchinson, M. C., Bramon Mora, B., Pilosof, S., Barner, A. K., Kéfi, S., Thébault, E., Jordano, P., and Stouffer, D. B. (2019). Seeing the forest for the trees: Putting multilayer networks to work for community ecology. *Functional Ecology*, 33(2):206–217.

[82] Häussler, J., Barabás, G., and Eklöf, A. (2020). A bayesian network approach to trophic metacommunities shows that habitat loss accelerates top species extinctions. *Ecology Letters*, 23(12):1849–1861.

[83] Jiménez-Valverde, A., Mendoza, S. J., Cano, J. M., and Munguira, M. L. (2006). *Comparing relative model fit of several species-accumulation functions to local Papilionoidea and Hesperioidea butterfly inventories of Mediterranean habitats*, volume 1 of *Topics in Biodiversity and Conservation*, pages 163–176. Springer Netherlands.

[84] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

[85] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

[86] Kawamoto, T. and Kabashima, Y. (2017). Cross-validation estimate of the number of clusters in a network. *Scientific Reports*, 7(1):3327.

[87] Kéfi, S., Miele, V., Wieters, E. A., Navarrete, S. A., and Berlow, E. L. (2016). How structured is the entangled bank? the surprisingly simple organization of multiplex ecological networks leads to increased persistence and resilience. *PLoS Biology*, 14(8):1–21.

[88] Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216.

[89] Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018). A review of dynamic network models with latent variables. *Statistics Surveys*, 12(0).

[90] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

[91] Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition.

[92] Kuhn, E., Matias, C., and Rebafka, T. (2020). Properties of the stochastic approximation EM algorithm with mini-batch sampling. *Statistics and Computing*, 30(6):1725–1739.

[93] Labeyrie, V., Thomas, M., Muthamia, Z. K., and Leclerc, C. (2016). Seed exchange networks, ethnicity, and sorghum diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 113(1):98–103.

[94] Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, 80:056117.

[95] Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C., and Dieckmann, U. (2018). Complexity and stability of ecological networks: a review of the theory. *Population Ecology*, 60(4):319–345.

[96] Larremore, D. B., Clauset, A., and Jacobs, A. Z. (2014). Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805.

[97] Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336.

[98] Latouche, P., Birmele, E., and Ambroise, C. (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115.

[99] Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

[100] Lazega, E. (2001). *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press on Demand.

[101] Lazega, E. (2009). Theory of cooperation among competitors: A neo-structural approach. *Sociologica*, 1:1–34.

[102] Lazega, E. (2020). *Bureaucracy, Collegiality and Social Change: Redefining Organizations with Multilevel Relational Infrastructures*. Edward Elgar Publishing.

[103] Lazega, E., Bar-Hen, A., Barbillon, P., and Donnet, S. (2016). Effects of competition on collective learning in advice networks. *Social Networks*, 47:1–14.

[104] Lazega, E. and Brailly, J. (2021+). Judges, lawyers, priests and scientists: Comparing networks with social contexts, social processes and social skills. Under review.

[105] Lazega, E. and Jourda, M.-T. (2016). The structural wings of matthew effects: The contribution of three-level network data to the analysis of cumulative advantage. *Methodological Innovations*, 9:2059799115622764.

[106] Lazega, E., Jourda, M.-T., Mounier, L., and Stofer, R. (2008). Catching up with big fish in the big pond? multi-level network analysis through linked design. *Social Networks*, 30(2):159–176.

[107] Lazega, E. and Mounier, L. (2002). Interdependent entrepreneurs and the social discipline of their cooperation: a research programme for structural economic sociology in a society of organizations. *Conventions and structures in economic organization: markets, networks, and hierarchies, Cheltenham, Edward Elgar Publishing*.

[108] Lazega, E., Sapulete, S., and Mounier, L. (2011). Structural stability regardless of membership turnover? the added value of blockmodelling in the analysis of network evolution. two definitions of collegiality and their inter-relation: The case of a roman catholic diocese. *Quality & Quantity*, 45:129–144.

[109] Lazega, E. and Snijders, T. A. (2015). *Multilevel network analysis for the social sciences: Theory, methods and applications*, volume 12. Springer.

[110] Lazega, E. and Wattebled, O. (2011). Two definitions of collegiality and their inter-relation: The case of a roman catholic diocese. *Sociologie du Travail*, 53:e57–e77.

[111] Le, C. M., Levin, K., and Levina, E. (2018). Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics*, 12(2):4697–4740.

[112] Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50.

[113] Leger, J.-B., Barbillon, P., and Chiquet, J. (2020). *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm.* R package version 1.1.4.

[114] Liu, M., Chen, X., Zhang, L., and Han, X. (2019). Node attacks on two-mode networks. *IEEE Access*, 7:108316–108330.

[115] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957.

[116] Mariadassou, M. and Matias, C. (2015). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573.

[117] Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *The Annals of Applied Statistics*, 4(2):715–742.

[118] Mariadassou, M. and Tabouy, T. (2020). Consistency and asymptotic normality of stochastic block models estimators from sampled data. *Electronic Journal of Statistics*, 14(2):3672–3704.

[119] Mariani, M. S., Ren, Z.-M., Bascompte, J., and Tessone, C. J. (2019). Nestedness in complex networks: observation, emergence, and implications. *Physics Reports*, 813:1–90.

[120] Martínez, V., Berzal, F., and Cubero, J.-C. (2016). A survey of link prediction in complex networks. *ACM Comput. Surv.*, 49(4).

[121] Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

[122] Matias, C., Rebafka, T., and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680.

[123] Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM: Proceedings and Surveys*, 47:55–74.

[124] McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013). Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis*, 60:12–31.

[125] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics Section. John Wiley & Sons.

[126] Medan, D., Perazzo, R. P., Devoto, M., Burgos, E., Zimmermann, M. G., Ceva, H., and Delbue, A. M. (2007). Analysis and assembling of network structure in mutualistic systems. *Journal of Theoretical Biology*, 246(3):510–521.

[127] Memmott, J., Waser, N. M., and Price, M. V. (2004). Tolerance of pollination networks to species extinctions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1557):2605–2611.

[128] Michalska-Smith, M. J. and Allesina, S. (2019). Telling ecological networks apart by their structure: A computational challenge. *PLOS Computational Biology*, 15(6):e1007076.

[129] Michalska-Smith, M. J., Sander, E. L., Pascual, M., and Allesina, S. (2018). Understanding the role of parasites in food webs using the group model. *Journal of Animal Ecology*, 87(3):790–800.

[130] Miele, V., Picard, F., and Dray, S. (2014). Spatially constrained clustering of ecological networks. *Methods in Ecology and Evolution*, 5(8):771–779.

[131] Miele, V., Ramos-Jiliberto, R., and Vázquez, D. P. (2020). Core–periphery dynamics in a plant–pollinator network. *Journal of Animal Ecology*, 89(7):1670–1677.

[132] Mukherjee, S. S., Sarkar, P., and Lin, L. (2017). On clustering network-valued data. *Advances in neural information processing systems*, 30.

[133] Newman, M. (2018). *Networks*. Oxford university press.

[134] Newman, M. E. J. and Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications*, 7(1):11863.

[135] Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

[136] Olhede, S. C. and Wolfe, P. J. (2013). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(41):14722–14727.

[137] Pachoud, C., Labeyrie, V., and Polge, E. (2019). Collective action in localized agrifood systems: An analysis by the social networks and the proximities. study of a serrano cheese producers' association in the campos de cima da serra/brazil. *Journal of Rural Studies*, 72:58–74.

[138] Paul, S. and Chen, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2):3807–3870.

[139] Paul, S. and Chen, Y. (2018). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *Annals of Applied Statistics*, 14(2):993–1029.

[140] Paul, S. and Chen, Y. (2020). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230–250.

[141] Pavlović, D. M., Guillaume, B. R., Towlson, E. K., Kuek, N. M., Afyouni, S., Vértes, P. E., Yeo, B. T., Bullmore, E. T., and Nichols, T. E. (2020). Multi-subject Stochastic Blockmodels for adaptive analysis of individual differences in human brain network cluster structure. *NeuroImage*, 220:116611.

[142] Pearl, J. (1982). Reverend bayes on inference engines: a distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence*, pages 133–136.

[143] Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548.

[144] Peixoto, T. P. (2014a). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1).

[145] Peixoto, T. P. (2014b). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047.

[146] Peixoto, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807.

[147] Peixoto, T. P. (2016). Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E*.

[148] Pilosof, S., Porter, M. A., Pascual, M., and Kéfi, S. (2017). The multilayer nature of ecological networks. *Nature Ecology and Evolution*, 1(4).

[149] Pocock, M. J., Evans, D. M., and Memmott, J. (2012). The robustness and restoration of a network of ecological networks. *Science*, 335(6071):973–977.

[150] Polge, E. and Torre, A. (2018). Territorial governance and proximity dynamics. the case of two public policy arrangements in the brazilian amazon. *Papers in Regional Science*, 97(4):909–929.

[151] Reyes, P. and Rodriguez, A. (2016). Stochastic blockmodels for exchangeable collections of networks. *arXiv preprint arXiv:1606.05277*.

[152] Rivera-Hutinel, A., Bustamante, R., Marín, V., and Medel, R. (2012). Effects of sampling completeness on the structure of plant–pollinator networks. *Ecology*, 93(7):1593–1603.

[153] Robins, G., Pattison, P., and Wang, P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks*, 31(2):105–117.

[154] Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878 – 1915.

[155] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[156] Saldaña, D. F., Yu, Y., and Feng, Y. (2017). How Many Communities Are There? *Journal of Computational and Graphical Statistics*, 26(1):171–181.

[157] Sander, E. L., Wootton, J. T., and Allesina, S. (2015). What can interaction webs tell us about species roles? *PLoS Computational Biology*, 11(7):e1004330.

[158] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464.

[159] Scossa, F., Alseekh, S., and Fernie, A. R. (2021). Integrating multi-omics data for crop improvement. *Journal of plant physiology*, 257:153352.

[160] Signorelli, M. and Wit, E. C. (2020). Model-based clustering for populations of networks. *Statistical Modelling*, 20(1):9–29.

[161] Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395.

[162] Snijders, T. A. (2017). Stochastic actor-oriented models for network dynamics. *Annual Review of Statistics and Its Application*, 4(1):343–363.

[163] Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100.

[164] Song, C., Rohr, R. P., and Saavedra, S. (2018). A guideline to study the feasibility domain of multi-trophic and changing ecological communities. *Journal of Theoretical Biology*, 450:30–36.

[165] Stanley, N., Shai, S., Taylor, D., and Mucha, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE transactions on network science and engineering*, 3(2):95–105.

[166] Sweet, T. M., Thomas, A. C., and Junker, B. W. (2013). Hierarchical network models for education research: Hierarchical latent space models. *Journal of Educational and Behavioral Statistics*, 38(3):295–318.

[167] Sweet, T. M., Thomas, A. C., and Junker, B. W. (2014). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. *Handbook on mixed membership models and their applications*, pages 463–488.

[168] Tabouy, T., Barbillon, P., and Chiquet, J. (2019). Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, pages 1–23.

[169] Tarres-Deulofeu, M., Godoy-Lorite, A., Guimera, R., and Sales-Pardo, M. (2019). Tensorial and bipartite block models for link prediction in layered networks and temporal networks. *Physical Review E*, 99(3).

[170] Thomas, M. and Caillon, S. (2016). Effects of farmer social status and plant biocultural value on seed circulation networks in Vanuatu. *Ecology and Society*, 21(2).

[171] Thompson, R. M. and Townsend, C. R. (2003). Impacts on stream food webs of native and exotic forest: An intercontinental comparison. *Ecology*, 84(1):145–161.

[172] Vacchiano, M., Lazega, E., and Spini, D. (2020+). What multilevel networks reveal about the life course: Insights on cumulative (dis)advantages. Under review.

[173] Vallès-Català, T., Massucci, F. A., Guimera, R., and Sales-Pardo, M. (2016). Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks. *Physical Review X*, 6(1):011036.

[174] Vallès-Català, T., Peixoto, T. P., Guimera, R., and Sales-Pardo, M. (2017). Consistencies and inconsistencies between model selection and link prediction in networks. *Physical Review E*.

[175] Vanbergen, A. J., Woodcock, B. A., Heard, M. S., and Chapman, D. S. (2017). Network size, structure and mutualism dependence affect the propensity for plant–pollinator extinction cascades. *Functional Ecology*, 31(6):1285–1293.

[176] Vissault, S., Cazelles, K., Bergeron, G., Mercier, B., Violet, C., Gravel, D., and Poisot, T. (2020). *rmangal: An R package to interact with Mangal database*. R package version 2.0.2.

[177] Vizentin-Bugoni, J., Debastiani, V. J., Bastazini, V. A., Maruyama, P. K., and Sperry, J. H. (2020). Including rewiring in the estimation of the robustness of mutualistic networks. *Methods in Ecology and Evolution*, 11(1):106–116.

[178] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[179] Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2012). Model-based clustering of large networks. *The Annals of Applied Statistics*, 7(2):1010–1039.

[180] Wang, P., Robins, G., Pattison, P., and Lazega, E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, 35(1):96–115.

[181] Wang, Y. X. and Bickel, P. J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, 45(2):500–528.

[182] Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61(3):401–425.

[183] White, H. C., Boorman, S. A., and Breiger, R. L. (1976). Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4):730–780.

[184] Xu, K. S. and Hero, A. O. (2013). Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7812 LNCS, pages 201–210. Springer, Berlin, Heidelberg.

[185] Yan, X. (2016). Bayesian model selection of stochastic block models. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 323–328. IEEE Press.

[186] Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks - A Bayesian approach. *Machine Learning*, 82(2):157–189.

[187] Yin, M., Wang, Y. R., and Sarkar, P. (2020). A theoretical case study of structured variational inference for community detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3750–3761. PMLR.

[188] Young, J.-G., St-Onge, G., Desrosiers, P., and Dubé, L. J. (2018). Universality of the stochastic block model. *Phys. Rev. E*, 98:032309.

[189] Zhang, L. and Peixoto, T. P. (2020). Statistical inference of assortative community structures. *Physical Review Research*, 2(4):043271.

[190] Zhang, X., Martin, T., and Newman, M. E. (2015). Identification of core-periphery structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 91(3).

[191] Zhao, Y. (2018). A Survey on Theoretical Advances of Community Detection in Networks. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5).

[192] Žiberna, A. (2014). Blockmodeling of multilevel networks. *Social networks*, 39:46–61.

[193] Žiberna, A. (2019). Blockmodeling linked networks. *Advances in Network Clustering and Blockmodeling*, pages 267–287.

[194] Žiberna, A. (2020). k-means-based algorithm for blockmodeling linked networks. *Social Networks*, 61:153–169.

[195] Zijlstra, B. J., Van Duijn, M. A., and Snijders, T. A. (2006). The multilevel p2 model. *Methodology*, 2(1):42–47.

[196] Žnidaršič, A., Doreian, P., and Ferligoj, A. (2019). Treating missing network data before partitioning. *Advances in Network Clustering and Blockmodeling*, pages 189–224.

[197] Žnidaršič, A., Ferligoj, A., and Doreian, P. (2012). Non-response in social networks: The impact of different non-response treatments on the stability of blockmodels. *Social Networks*, 34(4):438–450.

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre : Apprentissage statistique de collections de réseaux avec applications en écologie et en sociologie

Mots clés : Réseaux, Graphes aléatoires, Données hétérogènes, Modèle à blocs stochastiques, Méthodes variationnelles, Apprentissage non supervisé

Résumé : Cette thèse porte sur le développement de méthodes statistiques pour l'analyse de collections de réseaux d'interactions à travers trois contributions originales. Les réseaux d'interactions constituent une façon naturelle de représenter sous forme de graphe les échanges ou relations existant entre un ensemble de nœuds représentant des espèces ou des individus. Considérer des collections de réseaux permet d'étudier des systèmes hétérogènes, composés de plusieurs sortes d'interactions impliquant différents types de nœuds. Lorsque les différents réseaux de la collection sont liés par une relation hiérarchique, nous parlerons de réseaux multiniveaux. Le modèle à blocs stochastiques a prouvé sa pertinence pour modéliser l'hétérogénéité du comportement des nœuds dans un unique réseau. Des extensions aux collections de réseaux et aux réseaux multiniveaux sont proposées. Elles permettent d'obtenir un *clustering* des nœuds des réseaux en fonction de leur rôle dans l'écosystème ou le système social, et de résumer la structure du système à l'échelle mésoscopique à travers un faible nombre de paramètres. L'inférence de ces modèles est complexe et des méthodes variationnelles sont adaptées à cette fin. Des méthodes de sélection de modèles permettent également de déterminer la dépendance entre les niveaux pour les réseaux multiniveaux et la similarité entre les structures pour les collections de réseaux. Une dernière partie de cette thèse propose une nouvelle méthode pour étudier la robustesse de réseaux d'interactions écologiques. Chaque réseau est modélisé par un modèle probabiliste dont les paramètres représentent la structure du réseau. Cela permet de faire le lien entre la structure de l'écosystème et sa robustesse, mais aussi de comparer les robustesses d'une collection de réseaux et de corriger la robustesse d'un réseau dont l'échantillonage serait incomplet. Les méthodes développées sont implémentées dans des packages R et appliquées sur des données issues des sciences sociales et de l'écologie.

Title: Statistical learning of collections of networks with applications in ecology and sociology

Keywords: Networks, Random graphs, Heterogeneous data, Stochastic block model, Variational method, Unsupervised learning

Abstract: This thesis deals with the development of statistical methods for the analysis of collections of interaction networks through three original contributions. Interaction networks are a natural way to represent in graph form the exchanges or relationships existing between a set of nodes representing species or individuals. Considering collections of networks allows to study heterogeneous systems, composed of several kinds of interactions involving different types of nodes. When the different networks of the collection are linked by a hierarchical relationship, we speak of multilevel networks. The stochastic block model has proven its relevance to model the heterogeneity of the behavior of nodes in a single network. Extensions to collections of networks and to multilevel networks are proposed. They allow to obtain a clustering of the nodes of the networks according to their role in the ecosystem or social system, and to summarize the structure of the system at the mesoscopic scale through a small number of parameters. The inference of these models is complex and variational methods are adapted for this purpose. Model selection methods are also used to determine the dependence between levels for multilevel networks and the similarity between structures for collections of networks. A last part of this thesis proposes a new method to study the robustness of ecological interaction networks. Each network is modeled by a probabilistic model whose parameters represent the network structure. This allows to make the link between the structure of the ecosystem and its robustness, but also to compare the robustness of a collection of networks and to correct the robustness of a network whose sampling would be incomplete. The developed methods are implemented in R packages and applied on data from social sciences and ecology.