Estimation par maximum de vraisemblance dans des modèles à blocs stochastiques dynamiques ou spatiaux

Maximum likelihood estimation in dynamic or spatial stochastic block models

Léa Longepierre

Laboratoire de Probabilités, Statistique et Modélisation - UMR 8001
Sorbonne Université

Thèse pour l'obtention du grade de :
Docteur de l'université Sorbonne Université

Sous la direction de : Catherine Matias

Rapportée par : Nial Friel
Christine Keribin

Présentée devant un jury composé de : Catherine Matias *(directrice)*
Nial Friel *(rapporteur)*
Christine Keribin *(rapporteure)*
Matthieu Latapy *(examinateur)*
Nathalie Peyrard *(examinatrice)*
Fabrice Rossi *(examinateur)*

École Doctorale de Sciences Mathématiques
de Paris-Centre
Section Mathématiques appliquées

2 juillet 2020

Remerciements

Voilà maintenant dix ans que j'étudie au campus Pierre et Marie Curie, et c'est avec un peu d'émotion que je réalise que cette étape de ma vie se termine avec cette thèse. De nombreuses personnes m'ont permis d'en arriver là, et je souhaite leur exprimer ma gratitude. Aux personnes que j'aurais oubliées, je suis désolée, et je vous remercie également.

La première personne que je souhaite remercier est bien évidemment Catherine, ma directrice de thèse. Catherine, merci pour la confiance que tu m'as accordée, pour ta patience, pour m'avoir laissé la liberté dont j'avais besoin, tout en étant disponible et en prenant le temps de lire et relire de nombreuses fois mon travail quand j'en avais besoin. J'ai eu la chance de pouvoir profiter pendant cette thèse de tes immenses connaissances et de ta rigueur, que j'avais déjà beaucoup appréciée lorsque j'ai suivi ton cours en master.

Je tiens également à remercier mes rapporteurs, Nial Friel et Christine Keribin. Thank you both for taking the time to read my work with a lot of attention, for your comments, and for accepting to be part of my jury. Je remercie également les autres membres de mon jury, Matthieu Latapy, Nathalie Peyrard et Fabrice Rossi. Je suis honorée de vous présenter ma thèse.

Je remercie également les personnes ayant participé à mon comité de suivi de thèse, Pierre Barbillon, Tabea Rebafka et Nicolas Verzelen.

Le LPSM a été un cadre idéal pour préparer ma thèse et je m'y suis sentie très bien. Je tiens donc à remercier l'ensemble de ses membres, et notamment Lorenzo Zambotti et Ismaël Castillo, et aux membres de l'équipe Modélisation aléatoire du vivant. Pour leur aide administrative considérable, je dois également remercier Florence, Serena, Louise (qui m'a bien aidée depuis le master !), Josette, Valérie, Nathalie, Fatime...

Pendant ma dernière année de thèse, j'ai également travaillé à l'Université Paris Dauphine, où j'ai été particulièrement bien accueillie. Pour cela, je remercie entre autres Marc Hoffmann, Vincent Rivoirard, Béatrice Baeza et Ana Drumea.

J'ai eu l'occasion d'enseigner à Sorbonne Université et à Dauphine, et j'ignorais que ça me plairait autant, ça a été une expérience très enrichissante. Je remercie les responsables des cours dans lesquels j'ai enseigné, Tabea Rebafka, Vincent Lemaire et Olivier Lopez à Sorbonne Université, et Justin Salez, Christian Robert et Julien Stoehr à Dauphine.

Merci également aux membres du groupe de travail réseaux de l'Agro, de l'ANR EcoNet, et aux personnes avec qui j'ai pu échanger sur mon travail et demander conseil, telles que Christophe Ambroise, Mahendra Mariadassou et Julien Stoehr. J'ai eu l'occasion d'assister à de nombreuses conférences pendant ma thèse, et je remercie les organisateurs, en particulier les membres de COSTNET, ainsi que les personnes rencontrées pendant ces conférences.

Certaines personnes m'ont aidée avant cela en me préparant à cette thèse, tels que les très bon enseignants que j'ai pu avoir, notamment en master et à l'ISUP, entre autres Philippe Saint-Pierre, Michel Delecroix, Olivier Zindy, Arnaud Guyader, Olivier Lopez, Jean-Patrick Baudry, Frédéric Guilloux.... D'ailleurs, aussi loin que je me rappelle, j'ai eu d'excellents enseignants en mathématiques, qui ont nourri mon intérêt pour les mathématiques et m'ont conforté dans l'idée de poursuivre dans cette voie.

J'adresse un immense merci aux collègues du bureau 203, le meilleur bureau. Vraiment, je ne pouvais pas espérer avoir de meilleurs collègues. Merci à Romain pour cette dernière année, pour m'avoir fait déculpabiliser malgré toi et pour ton ouverture d'esprit musicale. Yating, merci beaucoup pour ton soutien, et pour ton aide en probabilités et administrative. Guillermo, merci de nous avoir fait découvrir ta culture avec des playlists de flûte et des pisco sour. Merci également à Cécile, Pauline et Yassir. Armand, je comprends toujours pas la moitié de tes jeux de mots mais tu me fais rire quand même, merci à toi. Merci à Yoan d'avoir toujours une histoire absurde à raconter en arrivant le matin et une pire phobie administrative que moi. Enfin, merci à Robert, pour avoir donné une touche de vie à ce bureau. Je n'oublie pas bien sûr le bureau 201, et en particulier Thibaut et Rancy. Rancy, j'espère moi aussi un jour pouvoir découvrir le Liban avec toi. Thibaut, ces partages de cafés (je te dois beaucoup de café d'ailleurs), puis d'appels et de plaintes m'ont toujours fait beaucoup de bien. Merci également à Pauline. Je remercie également mes collègues doctorants (ou anciens doctorants) du LPSM, notamment Sarah, Eric, Carlo, Paul, Nicolas, Michel, Flaminia, Suzanna, Henri, Florian, Alexandra, et mes collègues du CEREMADE, notamment Grégoire, Pascal et Meryem.

Je n'aurais pas réussi sans le soutien moral de mes amis, qui m'ont changé les idées quand j'en avais besoin. Je remercie d'abord la Fac team, mes partenaires de bière, de malbouffe, de mauvais films, de Simpsons (merci à Matt Groening pour toutes ces heures de divertissement au passage) et parfois de course (pour les plus courageux). Kikouin, on s'est découvert beaucoup de passions communes pendant ce stage à Saint-Louis et j'adore parler musique avec toi, et pouvoir profiter de ton immense culture cinématographique (qui va de Barton Fink à Suicide Squad en passant par Cats... Tu connais vraiment tout). Marion, tu proposes toujours les meilleurs plans et j'ai hâte de refaire un barbecue dangereux ou une raclette caniculaire chez toi. Adrien, tu es un de mes premiers amis de l'ISUP, tu m'as fait beaucoup rire dès les premiers jours, et ça ne s'est pas arrêté depuis. Je remercie également Léo, notre président, ces quelques soirées avec toi et Kevin pendant ma thèse ont été incroyables. Agathe, j'aime toujours autant ricaner bêtement avec toi depuis tout ce temps que je te connais, et ta reconversion m'inspire et me motive beaucoup à être courageuse et à me lancer pour avoir ce que je veux. Joseph, tu as été comme un petit démon sur mon épaule plus d'une fois, mais je suis très heureuse de partager mon amour pour Hanoi Rocks et le fromage fondu avec toi depuis plus de dix ans.

Merci également à mes autres amis de l'ISUP, notamment Pierre (merci de prendre le relais pour être pointilleux et prouver à Max qu'il a tort quand je ne suis pas là pour le faire), Virginie, Gabriel, Bichot et toute la filière biostatistique (on a dû affronter beaucoup d'épreuves ensemble, et vous retrouver une fois par an à Noël me fait toujours plaisir — Mélaine, désolée, je n'ai jamais répondu à ton mail et je m'en veux toujours...). Je n'oublie pas mes amis que j'ai moins eu le temps de voir pendant ces années. Aurélien, Thibaut, j'ai hâte de vous revoir (promis j'aurai plus de temps bientôt — oui, je sais que j'ai déjà dit ça à de nombreuses reprises) et de revenir passer des soirées au Supersonic avec vous. Steven, on arrivera à se faire cet apéro un jour.

J'adresse un très grand merci à Pierre Foissy, le meilleur médecin que j'ai eu l'occasion de rencontrer dans ma vie.

Mes remerciements vont aussi bien évidemment à ma famille, sans qui je n'aurais pas pu en arriver là. Tout d'abord, merci à mes parents ("les vieux"), je sais que je peux compter sur votre soutien, qu'il soit moral ou matériel, et je vous en remercie. Cet été à Buno m'a fait un bien fou (ainsi que ces petits week-ends à Londres toujours hyper sympas !), j'en avais vraiment besoin. La vieille, tu m'as toujours bien conseillée et aiguillée, tout au long de mes études, et toujours soutenue autant que possible, et tu n'as pas idée à quel point je t'en suis reconnaissante. Je remercie également mes frères et

sœur, Julien, Nicolas et Anna, ainsi que Redouane et mes neveux Augustin et Florent. On ne s'est jamais ennuyé à la maison avec vous !

Merci à Clément, mon cousin préféré depuis notre plus jeune âge, l'époque où on s'imaginait la vie parfaite et où tu me parlais avec passion de tes warhammer. Merci également à Bertrand et à toute la famille Weill, à Françoise et grand-papa, aux Wallard, Charlotte, Jean, Lucie, Henri (c'est toi qui m'a encouragée à me lancer dans cette thèse) et aux Longepierre. Merci à mes deux chats préférés, Nessie (qui aime venir dormir sur moi quand je travaille) et Kitty.

Merci à la famille Paté, qui m'a accueillie à bras ouverts (et chez qui j'ai particulièrement bien mangé et bu...). Nicole, Daniel, vous êtes les meilleurs beaux-parents. Merci également à Samuel et Christine, Léna, Scylla, Olivier, Thomas, Béatrice, et le reste de la famille.

Max(imilien), merci. Pour ton aide informatique précieuse d'abord, mais surtout pour m'avoir soutenue et supportée pendant ces années. Ça fait plus de cinq ans que tu fais partie de ma vie, et malgré nos discussions mathématiques houleuses, c'est un bonheur de vivre avec toi. Je garde un très bon souvenir de cet été studieux place des Vosges avec toi. Je suis très fière de toi, et infiniment reconnaissante pour ce que tu fais pour moi.

"I can't believe I wrote the whole thing."

Abstract

This thesis deals with maximum likelihood estimation in dynamic and spatial extensions of the stochastic block model (SBM), based respectively on hidden Markov chains and fields. In the first part, we consider a dynamic version of the stochastic block model, suited for the observation of networks at multiple time steps. In this dynamic SBM, the nodes are partitioned into latent classes and the connection between two nodes is drawn from a Bernoulli distribution depending on the classes of these two nodes. The temporal evolution of the nodes memberships is modeled through a hidden Markov chain. We prove the consistency (as the numbers of nodes and time steps increase) of the maximum likelihood and variational estimators of the model parameters, and obtain upper bounds on the rates of convergence of these estimators. We also explore the case where the number of time steps is fixed and the connectivity parameters are allowed to vary. Besides, we obtain some results regarding parameter identifiability in this dynamic SBM. In the second part, we introduce a spatial version of the stochastic block model, suited for the observation of networks at different spatial locations. In this spatial SBM, as before, the nodes are partitioned into latent classes and the connection is drawn from a Bernoulli distribution depending on the classes of these two nodes. There, the spatial evolution of the nodes memberships is modeled through hidden Markov random fields. We first prove that the model parameter is generically identifiable under certain conditions. For the estimation of the parameters, we propose an algorithm based on the simulated field Expectation-Maximisation (EM) algorithm, which is a variation of the EM algorithm relying on a mean field like approximation thanks to the simulation of latent configurations.

Keywords: maximum likelihood estimation, dynamic network, dynamic stochastic block model, variational estimation, temporal network, Markov random field, Potts model, EM algorithm, spatial data, spatial network, mean field like approximation

Résumé

Cette thèse porte sur le maximum de vraisemblance dans des extensions dynamiques et spatiales du modèle à blocs stochastiques (SBM), fondées respectivement sur des chaînes et champs de Markov cachés. Dans une première partie, on considère une version dynamique du modèle à blocs stochastiques, adaptée à l'observation de réseaux à différents pas de temps. Dans ce SBM dynamique, les nœuds sont répartis dans des groupes latents et la connexion entre deux nœuds suit une loi de Bernoulli dont le paramètre dépend du groupe de ces deux nœuds. L'évolution temporelle des appartenances aux groupes des nœuds est modélisée par une chaîne de Markov cachée. On prouve la consistance (lorsque les nombres de nœuds et de pas de temps augmentent) des estimateurs du maximum de vraisemblance et variationnels des paramètres du modèle, et on obtient des bornes supérieures pour le taux de convergence de ces estimateurs. On explore aussi le cas où le nombre de pas de temps est fixé et les paramètres de probabilités de connexion peuvent varier dans le temps. On obtient également des résultats concernant l'identifiabilité des paramètres dans ce SBM dynamique. Dans une seconde partie, on introduit une version spatiale du modèle à blocs stochastiques, adaptée à l'observation de réseaux dans différentes localisations spatiales. Dans ce SBM spatial, comme précédemment, les nœuds sont répartis dans des groupes latents et la connexion entre deux nœuds suit une loi de Bernoulli dont le paramètre dépend du groupe de ces deux nœuds. L'évolution spatiale des appartenances aux groupes des nœuds est modélisée par des champs de Markov cachés. On montre d'abord que le paramètre de ce modèle est génériquement identifiable sous certaines conditions. Pour l'estimation des paramètres, on propose d'adapter à notre modèle une variante de l'algorithme Espérance-Maximisation (EM) reposant sur une approximation de type champ moyen grâce à la simulation de configurations latentes.

Keywords: estimation du maximum de vraisemblance, réseau dynamique, modèle à blocs stochastiques dynamique, estimation variationnelle, champ aléatoire de Markov, modèle de Potts, algorithme EM, données spatiales, réseau spatial, approximation de type champ moyen

Contents

Résumé détaillé

Les graphes aléatoires constituent un outil adapté pour représenter et modéliser des réseaux, composés d'entités intéragissant entre elles. On s'intéresse dans cette thèse plus particulièrement au modèle à blocs stochastiques (Stochastic Block Model, SBM). On étudie des versions dynamique et spatiale du SBM, les données disponibles pouvant être composées de plusieurs graphes dépendant les uns des autres, notamment de façon temporelle ou spatiale.

On définit un graphe $\mathcal{G} = (V, E)$ comme étant un ensemble V de nœuds et un ensemble E d'arêtes entre ces nœuds. On s'intéresse dans cette thèse à des graphes binaires, c'est-à-dire dans lesquels une arête est présente ou absente (en opposition aux graphes valués, dans lesquels les arêtes ont des poids).

Le modèle à blocs stochastiques (SBM) est un modèle de graphe dans lequel n nœuds sont répartis dans Q groupes, cette répartition étant inconnue. Les appartenances aux groupes $\{Z_i\}_{1\leq i\leq n}$ sont des variables aléatoires latentes i.i.d. à valeurs dans $\{1, \dots, Q\}$, chaque nœud ayant une probabilité α_q d'être dans le groupe q. La distribution des arêtes est caractérisée par les groupes auxquels appartiennent les deux nœuds. On considère le cas binaire, dans lequel une arête entre deux nœuds appartenant respectivement aux groupes q et l est présente avec une probabilité π_{ql}. Le graphe observé a donc une matrice d'adjacence $X = (X_{ij})_{1\leq i,j\leq n} \in \{0,1\}^{n^2}$, où X_{ij} représente l'arête entre les nœuds i et j. Un cas particulier est celui où les nœuds d'un même groupe sont fortement connectés entre eux, et peu connectés avec les nœuds des autres groupes. Une méthode d'estimation classique des paramètres du SBM repose sur l'utilisation de l'algorithme VEM (Variational Expectation-Maximisation), le maximum de vraisemblance ne pouvant pas être calculé directement, et l'algorithme EM (Expectation-Maximisation) ne pouvant pas être utilisé en raison de la complexité de la loi des observations X sachant les variables latentes $(Z_1, \dots, Z_n)$.

Dans une première partie, présentée dans le Chapitre 2 on étudie une version dynamique en temps discret du SBM (Yang et al., 2011; Matias and Miele, 2017). Dans ce modèle, l'évolution temporelle de l'appartenance à un groupe de chacun des nœuds est

modélisée par une chaîne de Markov (non observée). À chaque pas de temps, les arêtes sont modélisées par un SBM binaire. Cela correspond au modèle de Yang et al. (2011). En se basant sur Celisse et al. (2012), on prouve la consistance (lorsque les nombres de nœuds et de pas de temps augmentent) de l'estimateur du maximum de vraisemblance $\hat{\pi} = \{\hat{\pi}_{ql}\}_{1\leq q,l\leq Q}$ du paramètre de probabilité de connexion entre groupes sous certaines hypothèses sur ces paramètres, et de celui de la matrice de probabilité de transition de la chaîne de Markov cachée avec une hypothèse supplémentaire sur la vitesse de convergence de $\hat{\pi}$. On obtient également une borne pour la vitesse de convergence de ces estimateurs. En pratique, on ne peut pas calculer les estimateurs du maximum de vraisemblance, on prouve donc également la consistance (et on obtient des bornes sur la vitesse comme précédemment) des estimateurs variationnels, approximant le maximum de vraisemblance grâce à un algorithme VEM (maximisant en fait une borne inférieure de la vraisemblance). On prouve de même la consistance des estimateurs lorsque le nombre de nœuds tend vers l'infini (et on obtient des bornes sur la vitesse) dans le cas où le nombre de pas de temps est fixé et les probabilités de connexions peuvent changer au cours du temps (correspondant au modèle de Matias and Miele (2017)). On obtient également des résultats concernant l'identifiabilité des paramètres dans ce SBM dynamique dans la Section 1.6.1.

Dans une deuxième partie, présentée dans le Chapitre 3, on propose une version spatiale du SBM. On considère une dépendance spatiale des graphes, modélisée par des champs de Markov. Pour chaque nœud i, les appartenances aux groupes suivent un champ de Markov, et plus précisément un modèle de Potts, basé sur un graphe $\mathcal{G}_i = (V_i, E_i)$ sur les L localisations. La loi des variables latentes d'appartenances aux groupes $Z_i = (Z_i^1, \dots, Z_i^L)$ pour un nœud i s'écrit alors

$$\mathbb{P}_\psi(Z_i) = \mathbb{P}_\psi(Z_i^1, \dots, Z_i^L) = \frac{1}{S_i(\alpha_i, \beta_i)} \exp\left[\sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{Z_i^l = q} + \beta_i \sum_{(l,l')\in E_i} \mathbb{1}_{Z_i^l = Z_i^{l'}}\right]$$

avec $\psi = (\alpha, \beta)$ où $\alpha = (\alpha_{iq})_{1\leq i\leq n, 1\leq q\leq Q}$ est le paramètre du champ externe et $\beta = (\beta_i)_{1\leq i\leq n}$ le paramètre d'interaction entre les localisations voisines dans les graphes sur les localisations, et avec $S_i(\alpha_i, \beta_i)$ une constante de normalisation (qui n'est pas calculable, sauf pour un nombre L de localisations très petit). À chaque localisation l, les interactions suivent un SBM de paramètre de connexion $\pi^l = \{\pi^l_{qq'}\}_{1\leq q,q'\leq Q}$, c'est-à-dire que pour tout i et j dans $\{1, \dots, n\}$,

$$X_{ij}^l \mid Z_i^l, Z_j^l \sim \mathcal{B}(\pi^l_{Z_i^l Z_j^l}).$$

On prouve l'identifiabilité générique des paramètres de ce modèle sous certaines hypothèses, et on explore le cas particulier du modèle d'affiliation[1], ce cas étant exclu par une des hypothèses, et on montre que ses paramètres ne sont pas identifiables sans hypothèse supplémentaire. On s'intéresse également à l'estimation des paramètres ψ (du modèle de Potts) et $\pi = \{\pi^l\}_{1 \leq l \leq L}$ (probabilités de connexion). L'estimateur du maximum de vraisemblance n'est pas calculable en raison des constantes de normalisation inconnue, et car la vraisemblance est une somme sur les Q^{nL} configurations latentes possibles. L'algorithme EM ne peut pas non plus être utilisé, en raison de la complexité de la loi des variables latentes $\{Z_i^l\}_{1 \leq i \leq n, 1 \leq l \leq L}$ sachant les observations $\{X_{ij}^l\}_{1 \leq i,j \leq n, 1 \leq l \leq L}$ et des constantes de normalisation. On propose d'estimer les paramètres en se basant sur l'algorithme simulated EM (Celeux et al., 2003), une variante de l'algorithme EM reposant sur une approximation de type champ moyen grâce à la simulation de configurations latentes. Plus précisément, on approxime la loi des variables latentes sachant les observations par une loi factorisée en négligeant pour chaque nœud à chaque localisation les fluctuations des localisations voisines et des autres nœuds à la même localisation, en les considérant fixées aux valeurs d'une configuration simulée avec un échantillonneur de Gibbs. Pour résoudre le problème de la constante de normalisation, on approxime également la loi des variables latentes par une loi factorisée en négligeant pour chaque localisation les fluctuations des localisations voisines. La méthode est illustrée sur des jeux de données synthétiques.

[1] dans lequel le paramètre de connexion prend seulement deux valeurs, qui sont une probabilité de connexion intra-groupe et une probabilité de connexion inter-groupes

List of Figures

List of Algorithms

Chapter 1

Introduction

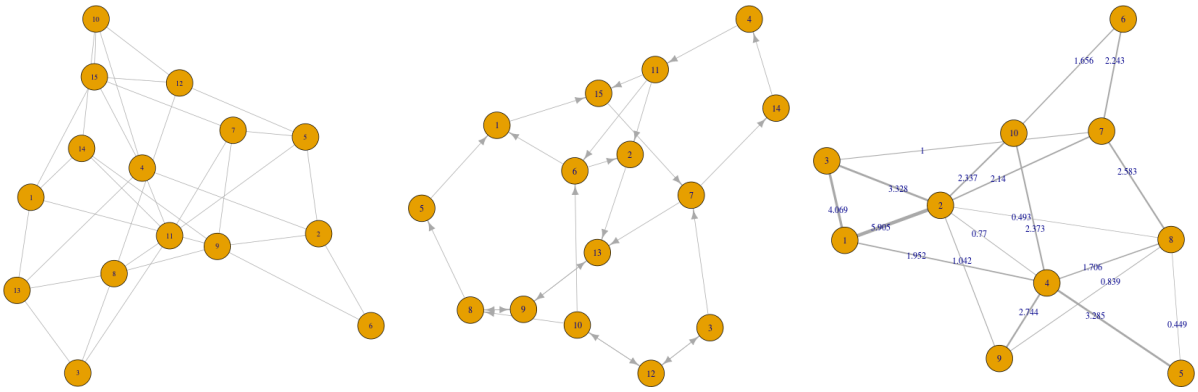
Random graphs are a suitable and widely used tool to model and describe interactions in many kinds of network datasets. A few examples are social networks (Facebook, Twitter, etc.), biological networks (neural networks, protein interaction networks, gene regulatory networks, etc.), ecological networks (food-webs, competition, interaction such as plant-pollinator networks, species contact networks), transport networks or computer networks.

Statistical analysis of random graphs has been intensively studied over the past decades. This research has focused on various methods, uses and applications, for example methods for sampling network data, describing characteristics of networks, inference problems or prediction.

In the real world, the data we observe may be more complex than just a single network. We often have to deal with multiple networks that are not independent, or graphs with interactions changing continuously for example, hence a need for statistical analysis methods and results for these kinds of complex networks. In this work, we are interested in the kind of complex data formed by a collection of graphs that are dependent, either over time or space. For example, some well-studied types of time-evolving networks are that of human proximity networks[1] or communication networks[2]. Regarding space-evolving networks, we could be interested in the relations between different categories of people (different socio-economic classes, different animal species) at different geographical locations.

In this chapter, we are going to give definitions, basic concepts and a brief overview of methods and results on random graphs. We will focus on the methods of node clustering and particularly on the Stochastic Block Model (SBM) (Holland et al., 1983; Nowicki and

[1] recording when two people are close to each other
[2] such as e-mails or phone calls between people



Undirected binary graph Directed binary graph Undirected weighted graph

Fig. 1.1 Representation of different types of graphs

Snijders, 2001). Then we will talk about dynamic graphs, and particularly a dynamic version of the SBM, and finally give some context about Markov random fields in order to introduce a space dependency between graphs.

1.1 Graphs: definitions and notations

We are first going to give in this section a definition and basic concepts of random graphs, that will be useful in the following. Let us define a *graph* $\mathcal{G}$ by

$$\mathcal{G} = (V, E)$$

where V is a set of *vertices* (also called *nodes*) and E is a set of *edges*, of which the elements are pairs of vertices, representing the connections between these vertices. In the following, we will consider a vertex set of the form $V = \{1, \dots, n\}$ with $n := |V|$ the cardinality of the set V (i.e. the number of nodes). The edges may be directed (or oriented) or not, and may be binary or weighted. The edges are said to be *directed* if an edge (i, j) is different from the edge (j, i) (with $i, j \in V$), and the considered graph is then called a directed graph. If the edges are not directed, the graph is called *undirected*. A graph is said to be *binary* if its edges are either present or absent. It is said to be *weighted* if the edges $e \in E$ have a weight, i.e. a value associated to it (usually in $\mathbb{R}$ or even in $\mathbb{R}^k$). See Figure 1.1 for the representation of three different types of graphs. A graph can be represented by its *adjacency matrix* $X = (X_{ij})_{1 \leq i,j \leq n}$, which is a $n \times n$

matrix defined as follows for a binary graph

$$X_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise,} \end{cases}$$

and as follows for a weighted graph

$$X_{ij} = \begin{cases} w_{ij} & \text{if } (i,j) \in E \text{ with an associated weight } w_{ij} \\ 0 & \text{if } (i,j) \notin E. \end{cases}$$

This matrix is symmetric for an undirected graph. Note that in this work, we will consider graphs with no *loops*, i.e. no edge (i,i) with $i \in V$ (leading to an adjacency matrix whose diagonal is equal to zero), and with no multiple edges, i.e. no more than one edge between two nodes. Such graphs are called *simple*. The graphs we will focus on will also be binary.

Let us introduce some vocabulary and define a few characteristics of graphs. The number of nodes n and the number of edges $|E|$ are sometimes called respectively the *order* and the *size* of the graph. Note that a graph with no loops (i.e. no edge (i,i)) has at most $n(n-1)/2$ edges if it is undirected and $n(n-1)$ if it is directed. We can then define the *density* ρ of a graph as the proportion of existing edges, i.e. the size of the graph over the maximum number of edges, that is $\rho = |E|/(n(n-1)/2)$ in the undirected case and $\rho = |E|/(n(n-1))$ in the directed case. Two vertices $i, j \in V$ are said to be *adjacent* (or *neighbours*) if there is an edge between i and j. Two edges are said to be *adjacent* if they have an endpoint in common. A vertex $i \in V$ is said to be *incident* on an edge $e \in E$ if i is an endpoint of e.

We define for a binary graph the *degree* d_i of a vertex i as the number of edges incident on this vertex. For directed graphs, we talk about *in-degree* (d_i^{in}) and *out-degree* (d_i^{out}), respectively counting the number of edges pointing to a vertex and pointing out from a vertex. Note that we can obtain the degrees by summing rows or columns of the adjacency matrix of a binary graph as follows

$$d_i = \sum_{j=1}^{n} X_{ij} = \sum_{j=1}^{n} X_{ji} \text{ for an undirected graph,}$$
$$d_i^{in} = \sum_{j=1}^{n} X_{ji} \text{ and } d_i^{out} = \sum_{j=1}^{n} X_{ij} \text{ for a directed graph.}$$

For any $i, j \in V$, we also define a *path* from i to j as a sequence of edges $e_1, \ldots, e_K \in E$ such that for every $1 \leq k \leq K-1$, the edges e_k and e_{k+1} share a common endpoint, and

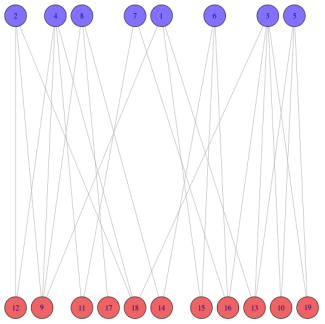


Fig. 1.2 Bipartite graph

i (resp. j) is an endpoint of e_1 (resp. e_K). A *cycle* is a path from a node i to itself. In particular, a cycle of size 3 (i.e. with 3 edges) is called a *triangle*.

Remark 1.1.1. The notion of path defined here does not take into account the direction of edges in the case of a directed graph. We can naturally define the notion of directed path in such graphs.

A graph $\mathcal{H} = (V_H, E_H)$ is a *subgraph* of the graph $\mathcal{G} = (V_G, E_G)$ if it is a graph such that $V_H \subseteq E_G$ and $V_H \subseteq E_G$.

The *connected components* of a graph are defined as maximally connected subgraphs (i.e. such that there is a path between any two nodes of the same connected component, and there is no path between a node in a connected component and a node outside of this connected component).

Types of graphs A graph $\mathcal{G} = (V, E)$ is said to be *complete* if all the vertices are connected one to another, i.e. $E = \{(i, j)\}_{i,j \in V, i \neq j}$. We can then define a *clique* of a graph $\mathcal{G}$, that is a complete subgraph of $\mathcal{G}$. Note that a clique of 3 nodes is a triangle, as defined above. A graph is said to be *connected* if it contains a unique connected component, i.e. if there exists a path between any two nodes $i, j \in V$. A graph is said to be *bipartite* if the nodes are of one of two types, and they can be connected only to nodes of the other type, as in Figure 1.2

Regarding the density of a graph, we will talk about dense graphs for graphs with many edges, when the number of edges is of the order of the maximal number of edges (typically $|E| \sim cn^2$ with c a positive constant) and of sparse graphs for graphs with relatively few edges, in which the number of edges is "much less" than the possible number of edges, for example is linear with respect to the number of nodes (typically $|E| = O(n)$).

We will talk about *random graphs* when considering a collection of graphs with a probability distribution on this collection.

In the following, we will present some statistical uses and applications of random graphs in Section 1.2, some random graph models in Section 1.3, and will then focus on node clustering techniques in Section 1.4, which is the main interest of this work. See Kolaczyk (2009) or Kolaczyk (2017) for more details on statistical analysis of graphs.

1.2 Some statistical uses of random graphs

As mentioned before, graphs can be used in many contexts, to tackle various problems and using many different methods. Their use goes back as far as the late 1930's with Moreno and Jennings (1938) in the context of social relationships (sociometry). We give a few (not exhaustive) examples.

Descriptive analysis One can simply be interested in some structural properties or characteristics of an observed graph in order to answer questions about it. For example, to try to answer the question "Do friends (i.e. neighbours in the graph) of a given individual tend to be friends of one another?", we can compute *transitivity measures*. Such measures can be the *transitivity coefficient*, counting the proportions of triangles (i.e. sets of three nodes connected by three edges) among connected triplets (i.e. sets of three nodes connected by at least two edges). A high transitivity coefficient means that the connected subgraphs of order three tend to form triangles, i.e. "the friend of your friend tends to be your friend". More generally we can describe the tendency of the nodes in a network to form cliques or groups of highly connected nodes based on *clustering* measures. A *clustering coefficient* can be defined in many ways, for example based on a local density measure, defined for any node i as the proportion of present edges among possible edges between the neighbours of node i. Taking the mean of these local density measures for every node gives a clustering coefficient. The transitivity coefficient introduced above can also be used as a clustering coefficient.

Another question can be "How important in the network is a given node?", important meaning that it has a lot of connections, and this characteristic can be described by *centrality* measures. Many centrality measures have been defined, and can be based for example on the degrees of the nodes or on the distance between a node and all the other nodes.

Some *connectivity* measures can also be defined, such as the number of connected components or their relative order with respect to the whole graph. The *average shortest*

path length, answering the question "What is the typical distance between two nodes?", can in particular determine if the observed network exhibit the *small world property*. This property, based on the suggestion of Milgram (1967) that we are separated from any other person on the planet by at most roughly six other people, refers to the fact that in some networks (even large ones), the distance between any two nodes is relatively small. It has been shown to hold for many types of networks by Watts and Strogatz (1998).

For more details on descriptive analysis of graphs, see Kolaczyk (2009).

Comparison to a null model Using generative models of graph, we can assess network topology. We can compare the observed network with graphs generated from random models, for some chosen features. The models used can be for example the simple Erdős-Rényi graph model (see Section 1.3) which assumes that every edge has the same probability of presence, or a model of random graphs with preserved degree sequence (see the fixed degree sequence configuration model in Section 1.3). The compared characteristics can be for example the reciprocity (as in Moreno and Jennings (1938) who compared the fraction of reciprocated links in their network with a random model), a clustering coefficient or the importance of occurrence of certain *motifs* in the graphs (for example triangles, cliques or cycles of a given size). It is possible to compute empirical p-values to perform hypothesis testing for the characteristic of interest, with H_0 the null hypothesis corresponding to the null model. See for example Zweig et al. (2016) for more details and references for the use of null models for random graphs.

Disease transmission In epidemiology, we can use graphs to study the spread of an infectious disease. In such context, the graph represents the contacts between people or subpopulations (and thus a possible contamination), that can evolve over time or not. For example, the SIR (susceptible-infectious-recovered)[3] model is widely used for that purpose. In such a model, if an infectious individual has a contact with a susceptible one, then the susceptible individual can become infectious, and will cease to be contagious after a certain period of time and be transferred to the recovered group that are then immune to the disease. Based on that model, some problematics can be tackled, such as the prediction of the existence and size of an outbreak. We can also study the effect of potential control measures on the epidemics. The use of graphs for the study of transmission dynamics has been applied for example to sexually transmitted infections (Haraldsdottir et al., 1992; Watts and May, 1992), to the 2009 H1N1 influenza pandemic,

[3]Other models of that type exist, such as SIS (susceptible-infectious-susceptible) or SEIR (susceptible-exposed-infected-recovered).

using a subpopulation network where connections among subpopulations represent the individual fluxes (Balcan et al., 2009; Bajardi et al., 2009), and more recently to the Covid-19 (Prasse et al., 2020).

Link prediction The task of predicting the absence or presence of an edge in a network (see for example Martínez et al. (2016)) is useful for example for recommendation systems, aiming at predicting probable links between nodes from an existing network, such as friendship (for example social media/Facebook) or the interest in a product (for example Amazon). Note that in the latter, i.e. in the case of product recommendation, the considered graphs are bipartite, the interest being the "interaction" between people and products, representing the interest in a product or the purchase of a product. The task of link prediction can be used when wanting to predict future links, or when a portion of the adjacency matrix is unobserved/missing (for example due to sampling issues). The methods used for link prediction can be based for example on score functions (that can be based on similarity measures, distance between the nodes, common neighbourhood...), by predicting the presence of an edge when the score is above a certain threshold, or on classification methods (for example based on logistic regression) to predict the missing values of the adjacency matrix with a classifier built from the observed values.

Node clustering A common interest when considering network data is the recovery of a clustering, i.e. a partition of the nodes into groups sharing the same connection behaviours. Indeed, the behaviour of entities (nodes) in a network is usually heterogeneous, and this approach allows to describe this heterogeneity and to obtain a summary of the network through groups with different behaviours. A particular case is that of community detection, when one wants to find groups of nodes that are highly connected between them, and less connected to nodes from other groups. We will talk about network clustering later.

Model fitting and testing One can be interested in fitting a model to the observed network, and testing the goodness of fit of this model. This allows, if the model is interpretable, to describe some characteristics of the network. Moreover, fitting a model to different networks can enable us to compare these networks. In this work, we will be interested in particular in parameter estimation, and will give more details about it later.

1.3 Random graph models

Many random graph models have been introduced. We present here some of the commonly used ones. The simplest one is the Erdös-Rényi graph model, assuming a constant probability of presence for every possible edge. We also present the configuration model that is based on the degree sequence (either fixed or following a power-law distribution), the exponential random graph model (ERGM), based on exponential families of distributions, and the Barabási–Albert model that is based on a preferential attachment mechanism. We then talk about latent variables graph models, that are the stochastic block model, the latent block model and the W-graph model.

In this section, we consider graph models for the observation of a single network.

1.3.1 Erdős-Rényi graph model

The simplest graph model is the one introduced by Erdös and Rényi (Erdős and Rényi, 1959; Erdős and Rényi, 1960). Denoted by $G(n, p)$, this is a model of undirected binary graph with n vertices, where edges are present between pairs of nodes independently with probability p. The expected number of edges in this model is $p\binom{n}{2} = pn(n-1)/2$. For any node i, its degree D_i follows a binomial distribution with parameter $(n-1, p)$,

$$P(D_i = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

For large n and small p (such that np is approximately constant), each degree D_i then follows approximately a Poisson distribution of parameter $(n-1)p$. Note that we can allow $p := p_n$ to vary with the number of edges n (usually to converge to 0 as the number of nodes increases), otherwise the obtained graphs are dense, whereas most observed large real-world networks are sparse.

This model is very convenient and easy to manipulate, in particular its parameter p can be simply estimated by the density of the observed graph, i.e. the proportion of existing edges (i.e. the number of edges in the graph over the maximum number of edges $n(n-1)/2$). However, a drawback of this model is that it is often unrealistic to assume that the edges are independent and equally likely. This idea is supported by the fact that, as we will talk about later, real-world networks usually exhibit degree sequences fitting a power-law distribution (see Section 1.3.2). Moreover, contrarily to most real-world networks, Erdős-Rényi graphs do not exhibit much clustering behaviour, the connection behaviour being too homogeneous among nodes. For example, using the transitivity coefficient defined in Section 1.2, that takes its values in $[0, 1]$ (a coefficient of 0 meaning

that the graph contains no triangle and a coefficient of 1 meaning that two adjacent edges always form a triangle), the expectation of this coefficient in an Erdős-Rényi graph is equal to p. As mentioned before, most large real-world networks are not dense and we usually consider $p \coloneqq p_n$ converging to 0 (as the number of nodes increases) to model them. This leads to a clustering coefficient converging to 0, while the values have been found to be quite large in real-world networks.

Remark 1. Actually, this model is the version introduced by Gilbert (1959), it was a slightly different model that was originally introduced by Erdős and Rényi. The original model $G(n, M)$ is the collection of all the simple undirected graphs of order n and size M (i.e. with n nodes and M edges), with a uniform distribution on this collection. This collection contains $\binom{n(n-1)/2}{M}$ different graphs. While these two models are not identical[4], they are equivalent under certain conditions for large n and if $M \sim pn(n-1)/2$ (see Luczak (1990) or Frieze and Karoński (2016)).

1.3.2 Configuration model

As we just said, the Erdős-Rényi graph model leads to unrealistic graphs, for instance regarding their degree sequence. One possible way to consider a graph is to consider directly its degree distribution or degree sequence. We first state the Erdős–Gallai theorem (Erdős and Gallai (1961), see also Berge (1976)) that gives necessary and sufficient conditions for a finite sequence of natural numbers to be the degree sequence of a simple undirected graph.

Theorem (Erdős–Gallai theorem). A sequence of non-negative integers $d_1 \geq \cdots \geq d_n$ can be represented as the degree sequence of a finite simple graph on n vertices if and only if $d_1 + \cdots + d_n$ is even and

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{n} \min(d_i, k)$$

holds for every k in $\{1, \ldots, n\}$.

Fixed degree sequence We can consider a fixed degree sequence $(d_1 \geq \cdots \geq d_n)$ for the n nodes, and consider the collection of all the graphs of order n with this degree sequence[5], with a uniform probability. Note that not any sequence of nonnegative

[4] In particular, in $G(n, M)$, we fix the number of edges, whereas in the case of $G(n, p)$, the number of edges follows a binomial distribution of parameter $(n(n-1)/2, p)$.

[5] Note that all such graphs have a fixed number of edges $\sum_{i=1}^{n} d_i/2$.

integers can be used for this definition. Indeed, the sequence must satisfy the conditions stated in Erdős–Gallai theorem above to be the degree sequence of a simple undirected graph. It is possible to generate graphs from this model with a sequence satisfying the Erdős–Gallai theorem, for example with a matching or rewiring algorithm (Algorithms 1 and 2 respectively). More efficient algorithms have also been introduced (see for example Viger and Latapy (2005)). The matching algorithm starts from an empty graph (with no edges) and adds an edge at each iteration between two nodes whose degrees are not yet equal to their fixed degrees (in the considered degree sequence). It is not very efficient as it can create graphs with multiple edges or loops, "forcing" us to start over. The rewiring

Algorithm 1: Matching algorithm

input : A sequence of degrees $d = (d_1, \ldots, d_n)$
output : A list of edges

do
  Initialise empty node and edge lists V and E;
  for $i = 1$ **to** n **do**
    while $d_i \geq 1$ **do**
      $V \leftarrow \text{concatenate}(V, i)$;
      $d_i \leftarrow d_i - 1$;
    end
  end
  while V *is not empty* **do**
    Draw i, j uniformly from v without replacement;
    $E \leftarrow \text{concatenate}(E, \{i, j\})$;
  end
while E *contains loops or multiple edges*;
return E

algorithm starts from an initial graph with the considered degree sequence and at each iteration, replaces an edge with another without changing the degrees (provided that these new edges do not create multiple edges or loops). It is more efficient than the matching algorithm, but requires an initial graph with the expected degree sequence.

As mentioned earlier, to assess the significance of topological characteristics of an observed network, one can use a null model to compare the features of interest. The fixed degree configuration model is widely used for that, by generating random graphs with the same degree sequence as the observed one[6] and comparing the feature in the observed and simulated graphs.

[6] using Algorithm 1 or 2 for example.

Algorithm 2: Rewiring algorithm

input : A list of edges E of a graph with the considered degree sequence and a number of iterations T

output : An updated list of edges E

for $t = 1$ **to** T **do**
  Draw $e_1 = \{i_1, j_1\}$ and $e_2 = \{i_2, j_2\}$ uniformly from E;
  if $i_1 \neq j_2$, $i_2 \neq j_1$ *and* $\{i_1, j_2\}, \{i_2, j_1\} \notin E$ **then**
    Replace e_1 and e_2 with $\{i_1, j_2\}$ and $\{i_2, j_1\}$ in E;
  end
end
return E

Power-law degree distribution We can also assume that the degrees $D_1, \dots, D_n$ are i.i.d. random variables following a power-law distribution with some parameter $\gamma > 0$, i.e. such that $\forall i \in [\![1, n]\!]$

$$\mathbb{P}(D_i = d) \propto d^{-\gamma},$$

where $\propto$ means "proportional to". To generate a graph from this model, we could start by drawing a degree sequence from the power-law distribution, and then generate a graph with this degree sequence, but the drawn degree sequence has no reason to satisfy the Erdős–Gallai theorem. Such a method could be very inefficient. Britton et al. (2006) propose for example to circumvent this problem by using the matching algorithm (Algorithm 1) and removing loops and merging multiple edges into single edges in the generated graph to obtain a simple graph, thus obtaining a degree distribution slightly different from the wanted one, but with asymptotically the right degree distribution under certain conditions on the moments of the degree distribution (thus on the power law exponent). Some graph models exhibiting a power law degree distributions have however been introduced, for example generative network growth models have been introduced (Barabási and Albert, 1999; Kleinberg et al., 1999; Kumar et al., 1999, 2000; Aiello et al., 2001), mainly for the analysis of the World Wide Web, such as the Barabábasi-Albert model (that we will describe in Section 1.3.3). The purpose of such models is to obtain graphs exhibiting a power law degree sequence for a large number of nodes n. Some other methods were introduced, such as an extension of the Erdős-Rényi graph model (referred to as the generalised random graph) with random edge probabilities, based on the introduction of node-specific random variables (Britton et al. (2006), Chapter 6 of Hofstad (2016), Lee et al. (2017)). In particular, Lee et al. (2017) propose to use Bertoin–Fujita–Roynette–Yor random variables that satisfy the required conditions on the

node-specific variables to lead to a power law distribution, and they propose a variational Bayesian inference approach to estimate the parameter.

Regarding the estimation of the exponent γ, different methods are used. A simple method is to perform linear regression on the logarithm of the observed proportion of vertices with degree d for any d (i.e. of the empirical degree distribution), with respect to the logarithm of d. However, such methods have been shown to yield inaccurate estimators for different reasons (for details, see Goldstein et al. (2004); Bauke (2007) or Clauset et al. (2009)). Estimation of the exponent can also be based on the maximum likelihood (Bauke, 2007; Gao and van der Vaart, 2017; Nettasinghe and Krishnamurthy, 2019), based on either discrete or continuous data (see details in Clauset et al. (2009)). In particular, the widely used Hill estimator introduced by Hill (1975) is equivalent to the maximum likelihood estimator (MLE) when considering the data as continuous. Such methods have been proved to be consistent (Gao and van der Vaart, 2017; Wang and Resnick, 2019) and asymptotically normal (Gao and van der Vaart, 2017).

Other methods exist, for example based on the Kolmogorov-Smirnov statistic that is used for comparing a sample with a reference probability distribution (Klaus et al., 2011), or based on mean degrees (Ikeda, 2009). One can see for example Clauset et al. (2009) for some methods for the estimation of the exponent of power-law distributions.

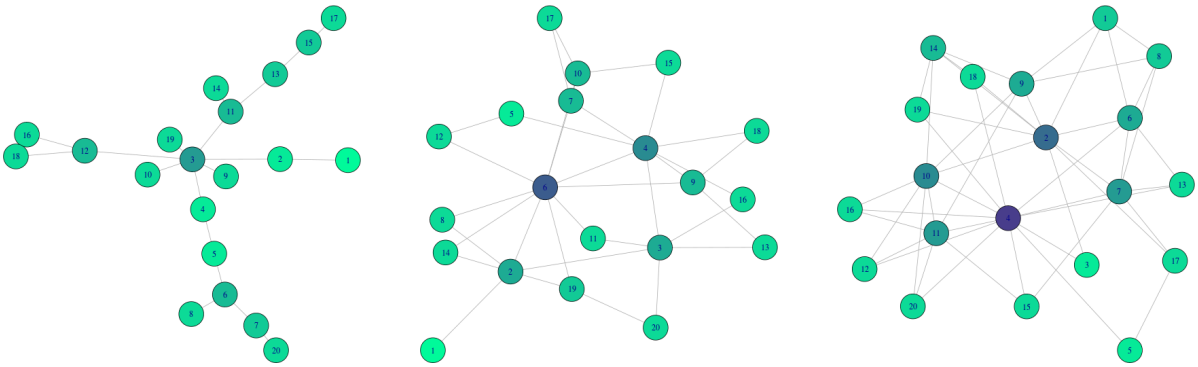
Note that one could choose other distributions for the degree sequences, but the power-law is a reasonable choice, as a lot of real networks have been shown to exhibit such a distribution. Some example are citation networks, some social networks including the collaboration network, networks in cell biology (Albert, 2005), the World Wide Web, etc. See Barabási and Albert (1999) for more detailed examples.

However, note that recently the adequacy of the power-law to many real-world networks has been questioned (Clauset et al., 2009; Latapy et al., 2017; Broido and Clauset, 2019).

1.3.3 Barabási–Albert model (Preferential Attachment)

As mentioned in the previous section, the Barabási–Albert model, based on a preferential attachment mechanism, is a generative network growth model, the graph growing at each step of the algorithm. It was introduced by Barabási and Albert (Barabási and Albert, 1999) and motivated by the growth of the World Wide Web.

The generation of a graph is as follows. We start with an initial graph $\mathcal{G}_0 = (V_0, E_0)$. Then, at each step, a new node of degree $m \geq 1$ is added to the network. It is connected to m existing nodes with a probability that is proportional to the degrees of the existing nodes.



$m = 1$ $m = 2$ $m = 3$

Fig. 1.3 Barabási–Albert model for different values of m, all starting from an initial graph of 5 nodes.

This model illustrates the concept *Rich get richer*, i.e. a positive feedback phenomenon in which the more connected a node is, the more likely it is to receive new connections. This hence tends to increase the difference of degrees between the nodes, and leads to a few nodes (called hubs) being very highly connected compared to the others. This is consistent with what we observe in numerous real world networks, including citation networks, some social networks, or the World Wide Web as mentioned above. Moreover, both phenomena of growth and preferential attachment are widely observed in real networks. This is rather rational, as when a new person (or entity) enters the network, it is more likely to become acquainted with one of the most visible people rather than with people with few connections.

This model exhibits an important property, which is that as the number of steps (and then of nodes) increases to infinity, the graphs generated by the Barabási–Albert model have degree distributions that tend to a power-law, i.e. $\mathbb{P}(D = d) \propto d^{-\gamma}$ (Barabási and Albert, 1999), as most large real-world networks have been shown to exhibit, as mentioned above. Barabási and Albert (1999) also show that in their model, both growth (of the number of nodes) and preferential attachment are needed to observe this degree distribution.

In practice, we work with a finite number of time steps/nodes, and the choice of the parameters can have an influence on the obtained graph. For example, see Figure 1.3 for the influence of m (nonetheless, note that the visualisation of a graph can be misleading, as the same graph can be represented in many different ways).

Note that other preferential attachment models exist. For example Gao et al. (2017b) and Gao and van der Vaart (2017) use a more general model than the Barabási–Albert model, in which the degree of the new nodes m depends on the time step t, and a new

node is connected to existing nodes with a probability that is proportional to a function of the degrees of the existing nodes (i.e. proportional to $f(d_i)$ with a certain function f instead of d_i for each existing node i).

1.3.4 Exponential Random Graph Models (ERGM)

ERGMs, also called p^* *models*, (see Strauss and Ikeda (1990); Wasserman and Pattison (1996); Hunter and Handcock (2006); Snijders et al. (2006), and also Frank and Strauss (1986) who introduced Markov random graphs, a particular sub-class of exponential random graph models) are based on exponential families of distributions. For a survey of this model, one can refer to Robins et al. (2007a); Wasserman and Robins (2005); Goldenberg et al. (2010), or to van der Pol (2019) for a more recent reference. This model assumes that the probability of a graph (i.e. of its adjacency matrix X) can be explained by a statistic $S(X)$ as follows

$$\mathbb{P}_\theta(X) = \frac{1}{c(\theta)} \exp\left(\theta^\top S(X)\right),$$

where $c(\theta)$ is a normalising constant, that we generally cannot compute in practice. ERGMs are flexible models in the sense that they can be based on different types of statistics, according to the type of pattern or behaviour we are interested in. For example S can be or can include density-related statistics (for example the number of edges), degree-based statistics, number of triangles or cycles, etc. These models are in particular widely used in social networks (Robins et al., 2007a,b; Lusher et al., 2013). Indeed, an important feature of ERGMs is that the edges are dependent, which is appropriate to model social networks, in which it is not reasonable to assume independence of the relations. Note that such a model can be defined for both directed and undirected graphs.

Note that an advantage of the ERGM is that it allows the inclusion of covariates (on the nodes), also called *actor attributes* (Robins et al., 2001), influencing the nodes connection behaviour. It is also possible to add covariates on pairs of nodes. An extended version of the ERGM can also be defined for weighted graphs (Robins et al., 1999; Desmarais and Cranmer, 2012; Krivitsky, 2012), and for multigraphs (Pattison and Wasserman, 1999).

When interested in the estimation of the parameter for the ERGM, it can be done by approximating the maximum likelihood, but there is a number of issues with this method, namely degeneracy (or *near degeneracy*) issues (Handcock, 2003; Handcock et al., 2003; Hunter et al., 2008; Rinaldo et al., 2009; Schweinberger, 2011; Chatterjee et al., 2013). Handcock et al. (2003) defines this issue as occurring when only a few

graphs do not have a very low probability, these graphs often being the full graph and the empty graph. Then, such models are not interesting for modeling real networks, and in addition, degeneracy is often associated with poor properties of estimation methods based on the likelihood, such as Markov chain Monte Carlo (MCMC) procedures.

An approximation of the maximum likelihood is required, because it is not tractable in this model. Some methods have been introduced, such as the use of the pseudolikelihood (Besag, 1975; Strauss and Ikeda, 1990) where the joint distribution is replaced by the product of the conditional distributions, but this method has been shown to not behave well, depending on the network (Wasserman and Robins, 2005; Robins et al., 2007b; Van Duijn et al., 2009).

Another type of methods for this purpose, perhaps more used, is the use of MCMC techniques (Snijders, 2002; Hunter and Handcock, 2006; Handcock, 2003) in order to approximate the maximum likelihood estimator, for example by estimating likelihood ratios, using a Metropolis-Hastings algorithm to generate a sample of networks simulated from the ERGM. As mentioned earlier, such methods can behave badly, converging to degenerate graphs[7] or failing to converge. New faster sampling techniques have been introduced (Stivala et al., 2020), allowing to perform these MCMC procedures more efficiently and on larger networks. Other methods for large graphs include computing estimators on snowball samples (Goodman, 1961) from the network (Pattison et al., 2013; Stivala et al., 2016).

Some methods have been introduced to tackle the difficulties due to degeneracy, for example the use of curved ERGMs (Hunter and Handcock, 2006; Hunter, 2007) which generalises the ERGM (see Efron (1975, 1978)). In particular Snijders et al. (2006) introduced a particular specification of ERGMs for the analysis of social networks, with new statistics to represent structural properties such as transitivity and heterogeneity of degrees, and which solves some of the problems of degeneracy. Caimo and Friel (2011) propose a MCMC algorithm in a Bayesian framework and give empirical evidence that the method quickly converges. Schweinberger and Handcock (2015), Schweinberger and Stewart (2020) and Schweinberger (2020) propose ERGMs with local dependence, adding an additional structure to the network by assuming that the graph nodes can be partitioned into subgraphs, and that dependence exists within subgraphs but not between subgraphs. They obtained consistency results in that context. Some other theoretical results have been obtained, such as Mukherjee (2020) who obtained a sufficient criterion for non-degeneracy for ERGMs on sparse graphs.

[7] either complete or empty

1.3.5 Stochastic Block Model (SBM)

1.3.5.1 Definition

We will focus on the Stochastic Block Model (SBM) (Holland et al., 1983; Frank and Harary, 1982; Nowicki and Snijders, 2001), a widely used latent variables model. In this model, the vertices are partitioned into Q groups (or classes), the group memberships being represented by latent variables $Z = (Z_1, \dots, Z_n)$, and the connection between two nodes is drawn from a distribution depending on the classes of these two nodes. The latent variables are independent and identically distributed (i.i.d.) in $\{1, \dots, Q\}$, following the distribution $\alpha = (\alpha_1, \dots, \alpha_Q)$ with $\alpha_q \in (0, 1)$ for every $q \in \{1, \dots, Q\}$. We distinguish two types of SBM, the binary SBM and the weighted SBM, which has been introduced later (Jiang et al., 2009; Mariadassou et al., 2010). In the binary SBM (which will be our interest), conditional on the latent groups $\{Z_i\}_{1 \leq i \leq n}$, the edges $\{X_{ij}\}_{1 \leq i,j \leq n}$ are independent Bernoulli random variables

$$X_{ij} \mid Z_i = q, Z_j = l \sim \mathcal{B}(\pi_{ql}),$$

where $\pi = (\pi_{ql})_{1 \leq q,l \leq Q} \in [0, 1]^{Q^2}$ is the connectivity parameter. Note that the matrix π is symmetric if we consider undirected graphs. In the weighted case, the weight associated with an edge follows a given parametric probability distribution, for example a Poisson or Gaussian distribution. We will denote by $\theta = (\alpha, \pi)$ the parameter of the SBM.

The SBM allows to directly model the heterogeneity of the connection behaviours of the nodes. It is flexible in the sense that it can model any type of network that is characterised by the connection behaviours of groups of nodes, encompassing very different structures of networks. In particular, it can model networks with a community structure, i.e. composed of groups of nodes that are highly connected between them and less connected to nodes from other groups, as on the left of Figure 1.4. This kind of network exhibiting a community structure can be modeled by the affiliation model that is a particular case of the SBM, and that is described in the next paragraph. Other types of graphs that can be modeled by the SBM are graphs with hubs and peripheral nodes as in the middle of Figure 1.4 or graphs with groups of nodes interacting mainly with nodes from other groups (and in particular bipartite graphs) as on the right of Figure 1.4. It can model more generally any network structure with groups of nodes sharing a similar connection behaviour towards other nodes (in the same class or in any other class). We will talk in Section 1.4.4 about the use of SBM for clustering purposes, and we will talk in the same section and in the introduction of Chapter 2 about the estimation of the

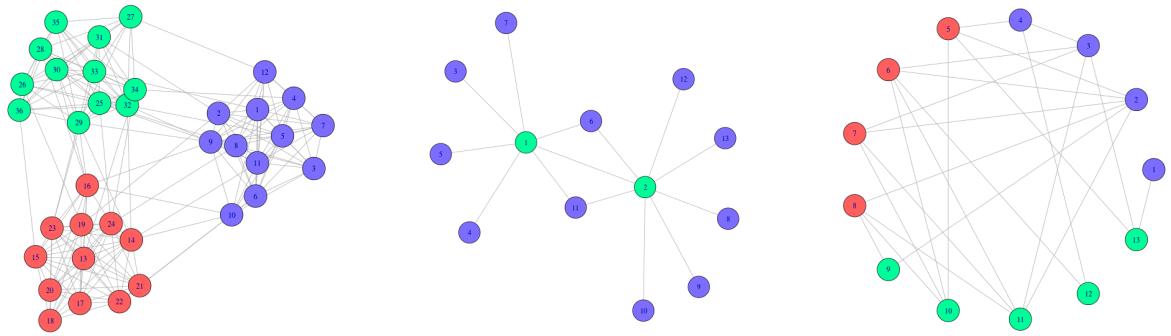


Fig. 1.4 Community structure on the left, hubs and peripheral nodes in the middle and groups interacting mainly with other groups on the right

parameters in the SBM, and particularly about the variational EM algorithm (VEM), which will be the interest of our work.

Particular case: the affiliation model A particular case of the SBM is the affiliation model, in which the connection parameter $\{\pi_{ql}\}_{1\leq q,l\leq Q}$ only takes two different values, that are the between-groups and the within-group connection probabilities. We then have, for any $q, l \in \{1, \ldots, Q\}$,

$$\pi_{ql} = \begin{cases} \pi_{\text{in}} & \text{if } q = l \\ \pi_{\text{out}} & \text{if } q \neq l. \end{cases}$$

In particular, assuming that π_{in} is large and π_{out} is small, the graphs generated by the affiliation model exhibit a community structure, such as the graph on the left in Figure 1.4. Inversely, assuming that we have π_{in} small and larger π_{out}, we obtain graphs such as the one on the right in Figure 1.4.

Weighted SBM Let us give more details about the weighted version of the SBM. It consists in replacing the Bernoulli distribution for the distribution of $X_{ij} \mid Z_i, Z_j$ by any parametric distribution with a parameter depending on Z_i, Z_j. To control the density of the graph[8], we use a zero-inflated distribution, i.e. the values taken by the edges are defined as a mixture of a Dirac mass at zero and the considered probability distribution. Explicitly, defining the model parameter $\theta = (\{\alpha_q\}_{1\leq q\leq Q}, \{\eta_{ql}\}_{1\leq q,l\leq Q}, \{\pi_{ql}\}_{1\leq q,l\leq Q})$, we

[8]For example, if the considered distribution is absolutely continuous with respect to the Lebesgue measure, all the edges are present in the graph, which may not be appropriate.

write $X_{ij} \mid Z_i = q, Z_j = l \sim F(\cdot; \eta_{ql}, \pi_{ql})$ with

$$F(x; \eta_{ql}, \pi_{ql}) = (1 - \pi_{ql})\delta_0(x) + \pi_{ql} G(x; \eta_{ql}),$$

where $G(\cdot; \eta_{ql})$ is the conditional distribution of X_{ij} given $Z_i = q$ and $Z_j = l$ and given the presence of the edge, and δ_0 denotes the Dirac distribution. We assume that G has no point mass at zero for identifiability purposes. If the considered distribution initially has some mass at zero, we then use its zero-truncated version. The distribution G can be for example a zero-truncated Poisson distribution for discrete weights, or a Gaussian or Laplace distribution for continuous weights. Note that the π_{ql} are density parameters of the graphs. For parsimony purposes, we can assume that these parameters are constant, i.e. $\pi_{ql} \coloneqq \pi$ for every q, l, meaning that the density is homogeneous between groups.

1.3.5.2 Identifiability

Identifiability results have been obtained for the SBM (Allman et al., 2009, 2011; Celisse et al., 2012). We will give more details about the work of Allman et al. (2009) and Allman et al. (2011), as one of our result will be based on it.

Allman et al. (2009) and later Allman et al. (2011) proved some identifiability results for latent structure models, including the SBM, in the context of undirected graphs. The proofs of their results are based on an application of a theorem by Kruskal (Kruskal (1976, 1977) or see for example Rhodes (2010) or Theorem 16 in Allman et al. (2011)) that states the identifiability (up to label swapping) of the parameters of a latent variable model with three observed discrete random variables, under some conditions based on the notion of Kruskal rank of matrices, and assuming in particular that the three observed random variables are independent given the latent one. This result is applied with an appropriate decomposition into three pairwise disjoint subsets of the complete set of edges, these three variables then being conditionally independent (and with each matrix having full row rank, implying full Kruskal rank). In Allman et al. (2009), they prove that the binary SBM with two groups has identifiable parameters as long as the three connection parameters are distinct, and for a number of nodes $n \geq 16$. In Allman et al. (2011), they prove the following main result for identifiability of the binary SBM. This result states the generic identifiability only, meaning that the nonidentifiable parameters form a set of Lebesgue measure zero. The form of the subspace of non identifiable parameters is not specified in this result, and it is important to keep in mind that when we impose a constraint on the parameter reducing the parameter space to a subspace of smaller dimension, parameter identifiability is no longer guaranteed.

Theorem 1.3.1 (Allman et al. (2011)). *The group proportion parameters α_q and the connection parameters π_{ql} for $q, l \in [\![1, Q]\!]$ are generically identifiable up to label permutation from the distribution of X when $Q \geq 3$ and $n \geq m^2$, with*

$$\begin{cases} m \geq Q - 1 + \left(\frac{Q+2}{2}\right)^2 & \text{if } Q \text{ is even,} \\ m \geq Q - 1 + \frac{(Q+1)(Q+3)}{4} & \text{if } Q \text{ is odd.} \end{cases}$$

Moreover, the result remains valid when the group proportions α_q are fixed.

As we said before, the constraints on the parameter leading to non identifiability are not specified in this result. However, the generic part of the proof of this result concerns only the connection parameters π_{ql}. In particular, in this proof, these parameters need to be distinct, so this result does not apply to the affiliation model. Some results have been derived in Allman et al. (2011) for this particular model. The proofs of these results are based on arguments on moments of the distribution, as these moments may be obtained explicitly in terms of model parameters for a small number of nodes. They obtain an identifiability result for $Q = 2$.

Corollary 1.3.1 (Allman et al. (2011)). *The group proportion parameters $(\alpha_1, \alpha_2 = 1-\alpha_1)$, up to label swapping, and the parameters $(\pi_{\text{in}}, \pi_{\text{out}})$ of the random graph affiliation mixture model with $Q = 2$ groups and binary edge state variables are strictly identifiable from the distribution of X if $n \geq 3$ and provided $\pi_{\text{in}} \neq \pi_{\text{out}}$.*

Other partial results are obtained in this case. When the group proportions are fixed and in $(0, 1)$, they prove that π_{in} and π_{out} are identifiable from the distribution of X if $n \geq 3$. They also state that it is necessary that $n \geq Q$ to identify the group proportion parameters and π_{in} and π_{out}. Finally, they state that if the group proportions are uniform ($\alpha_q = 1/Q$ for every q), the parameters π_{in}, π_{out} and the number of groups Q are identifiable from the distribution of X if $n \geq 4$.

We will discuss the estimation of these parameters in the binary SBM in Section 1.4.4.1.

Now, let us talk about identifiability in the weighted SBM. Allman et al. (2011) prove the identifiability of the parameters of the weighted SBM (up to label swapping) from the distribution of X if $n \geq 3$, as long as

- The $Q(Q+1)/2$ parameters $\{\eta_{ql}\}_{1 \leq q \leq l \leq Q}$ are distinct
- $G(\cdot; \eta)$ has no point mass at zero for any η
- The parameters of finite mixtures of measures in $\{G(\cdot; \eta)\}_\eta$ are identifiable, up to label swapping.

Note that most of the classical parametric distributions have been shown to satisfy the assumptions on G. In particular, the truncated Poisson and Gaussian distributions satisfy these assumptions.

The authors also give results for the weighted affiliation model and for a weighted SBM where the edges can take a finite number of values.

Celisse et al. (2012) obtained the identifiability of the SBM, in the case of directed or undirected graphs, as long as $n \geq 2Q$ and the coordinates of $\pi\alpha$ are distinct. They show that this applies to the affiliation model as long as the group proportions are different. They also obtain an identifiability with weaker assumptions for $n = 4$ and $Q = 2$.

1.3.5.3 Extensions of the SBM

Some extensions of the SBM have been introduced. We will briefly talk about three well-know extensions, that are the degree-corrected SBM, the mixed membership SBM and the inclusion of covariates in the SBM.

Degree-corrected SBM (DCSBM) The degree-corrected SBM (DCSBM) (Karrer and Newman, 2011) is a more flexible extension of the SBM, allowing nodes in the same groups to have different degree distribution, hence better fitting real world networks which often exhibit degree heterogeneity within a same group. The principle of the DCSBM is to introduce additional parameters for degree-correction $\zeta = (\zeta_1, \ldots, \zeta_n)$ that are integrated in the distribution of the adjacency matrix X given the latent variables. In Karrer and Newman (2011), in a weighted graph context (with edges taking integer values), the distribution of X given the latent variables is given by a Poisson distribution

$$X_{ij} \mid Z_i, Z_j \sim \mathcal{P}(\zeta_i \zeta_j \pi_{Z_i Z_j}).$$

A binary version of the degree-corrected SBM is also used (see for example Gao et al. (2018)), where the distribution of the adjacency matrix X given the latent variables is given by a Bernoulli distribution

$$X_{ij} \mid Z_i, Z_j \sim \mathcal{B}(\zeta_i \zeta_j \pi_{Z_i Z_j}).$$

Mixed membership SBM (MMSBM) The mixed membership SBM (MMSBM) (Airoldi et al., 2008) allows partial membership to different groups, in the sense that each node have a membership distribution over the classes instead of a membership to a single class. It is a model of interest when performing clustering of nodes with overlapping groups, i.e. in the context where a single node can play more than one role in the network,

depending on the node it is interacting (or not) with. Airoldi et al. (2008) define the MMSBM as follows. Each node i has a mixed membership vector $m_i = (m_{i,1}, \dots, m_{i,Q})$ following a Dirichlet distribution with parameter $\alpha = (\alpha_1, \dots, \alpha_Q)$, where $\alpha_q > 0$ for all $q \in [\![1, Q]\!]$. Then for each pair of nodes $i \neq j$, the roles $Z_{i\to j}$ and $Z_{j\to i}$ of these two nodes corresponding to that particular interaction are drawn from multinomial distributions with parameter respectively m_i and m_j. The value X_{ij} of the edge from i to j then follows a Bernoulli distribution

$$X_{ij} \mid Z_{i\to j}, Z_{j\to i} \sim \mathcal{B}(\pi_{Z_{i\to j} Z_{j\to i}}).$$

Note that other extensions of the SBM have been introduced for the analysis of graphs with overlapping classes, such as the overlapping SBM of Latouche et al. (2011), in which each node may belong to any number of groups at the same time.

SBM with covariates Another possibility to extend the SBM is to introduce covariates (Tallberg, 2004; Mariadassou et al., 2010; Choi et al., 2012), allowing to take into account the information we may have on the nodes or pair of nodes. The considered covariates we introduce in the SBM can then be node-specific covariates (for example age, gender, income) or edge-specific covariates (for example a distance between the two considered nodes). For example, in Tallberg (2004), the distribution of the group membership of a node depends on its covariate. In Mariadassou et al. (2010), the covariates are taken into account in the distribution of the edges X_{ij} via a regression model, in the context of weighted graphs.

1.3.6 Latent Block Model (LBM)

The Latent Block Model (LBM) (Govaert (2003), see also Govaert and Nadif (2013) or Brault and Mariadassou (2015)) is a co-clustering model, i.e. we consider an array data structure with n observations (n rows) of m variables (m columns) and it is assumed that there exists a partition of the rows and of the columns.

We assume that $Z = (Z_i)_{1\leq i\leq n}$ and $W = (W_j)_{1\leq j\leq m}$ are two independent latent variables defining a partition on the rows and columns respectively. The Z_i are i.i.d. random variables, following a multinomial distribution, and so do the W_j. We observe $X = (X_{ij})_{1\leq i\leq n, 1\leq j\leq m}$, these variables being independent conditional on Z and W and such that each X_{ij} conditional on Z, W follows the same parametric distribution, with a parameter $\pi_{Z_i W_j}$ depending on the column and row groups. Note that this formulation is quite similar to that of the SBM. The difference lies in the fact that there are two

distinct partitions for the rows and columns in the LBM, whereas it is the same for the SBM. In fact, Mariadassou and Matias (2015) proposed a unified framework for studying both these models.

In a graph context, it can be used to model bipartite graphs (where the nodes are of one of two types, and they can be connected only to nodes of the other type, as in Figure 1.2). We thus assume that we have a clustering of nodes for each of the two types.

1.3.7 Latent Position Model (LPM)

.

The latent position model (LPM) was introduced by Hoff et al. (2002) in the context of social networks. In this model, the latent variables on the nodes are i.i.d. random variables taking their values in a *social space* in $\mathbb{R}^d$ (and not in $[\![1, Q]\!]$) and the probability of connection between two nodes is determined by their distance in the latent space. Precisely, the edges are independent given the positions in the latent space, and the probability of an edge X_{ij} to be present is based on the logistic regression

$$\text{logit}(\mathbb{P}(X_{ij} = 1 \,|\, Z_i, Z_j, y_{ij})) \coloneqq \log\left(\frac{\mathbb{P}(X_{ij} = 1 \mid Z_i, Z_j, y_{ij})}{1 - \mathbb{P}(X_{ij} = 1 \mid Z_i, Z_j, y_{ij})}\right) = \alpha + {}^{\mathrm{t}}\beta y_{ij} - \|Z_i - Z_j\|_2$$

where y_{ij} is a vector of covariates on the pair of nodes (i, j) (if we have covariates at our disposal), (α, β) is the model parameter and $\|\cdot\|_2$ is the Euclidean norm[9].

Later, Handcock et al. (2007) introduced the latent position cluster model, extending the model of Hoff et al. (2002) for clustering purposes, in which the nodes positions in the latent social space come from a mixture of multivariate Gaussian distributions, the components of the mixture corresponding to the clusters.

1.3.8 Graphon and W-graph model

1.3.8.1 Definition

The W-graph model is a general binary graph model satisfying the *exchangeability* property. The exchangeability property is the fact that any permutation on the nodes labels (i.e. of the nodes and columns of the adjacency matrix) leads to the same distribution, i.e. $(X_{\sigma(i)\sigma(j)})_{1\leq i,j\leq n} \sim (X_{ij})_{1\leq i,j\leq n}$ with σ any permutation on $[\![1, n]\!]$. This means that the nodes labels have no relevance. This model is based on the definition of graphon. Formally introduced by Lovász and Szegedy (2006) in the context of graph

[9]Note that any other distance can be used.

limits, a graphon is defined as a symmetric measurable function $W : [0,1]^2 \to [0,1]$. It defines a graph as follows:

- First, assign independently a value u_i from the random variable $U_i \sim \mathcal{U}([0,1])$ to each node $i \in \{1, \dots, n\}$.
- Then the edges are independent given $(u_1, \dots, , u_n)$, each edge X_{ij} following a Bernoulli distribution of parameter $W(u_i, u_j)$.

Such a graph is called a W-random graph. If W is a constant function equal to p, the associated W-random graph model is an Erdős-Rényi graph model of parameter p. Note that a SBM is a particular case of a W-graph, where the graphon function is block-wise constant with rectangular blocks of size $\alpha_q \times \alpha_l$ and value π_{ql}, as represented in Figure 1.5. In that sense, the graphon is in fact a continuous extension of the SBM.

This model has been introduced in Lovász and Szegedy (2006) (see also Lovász (2012); Borgs et al. (2008)) in the study of large graphs, as the limit object of a sequence of dense graphs (see Borgs et al. (2019) for sparse graphs).

This is related to the Aldous-Hoover theorem (Aldous (1981); Hoover (1979), see also Kallenberg (2006)), that is a two-dimensional version of De Finetti's theorem (see for example Diaconis and Janson (2008)). It states (using a formulation similar to that in Orbanz and Roy (2014)) that an array $X = (X_{ij})_{1 \leq i,j \leq n}$ (with X_{ij} in some space $\mathcal{A}$) is exchangeable if and only if there is a random function $F : [0,1]^3 \to \mathcal{A}$ such that X_{ij} is equal in distribution to $F(U_i, U_j, U_{ij})$, where $(U_i)_{1 \leq i \leq n}$ and $(U_{ij})_{1 \leq i < j \leq n}$ are i.i.d. random variables following a uniform distribution on $[0,1]$, and $U_{ji} = U_{ij}$. In our context of binary undirected graphs, i.e. with $\mathcal{A} = \{0,1\}$, this can be expressed in terms of the graphon function, with X_{ij} following a $\mathcal{B}(W(U_i, U_j))$ (see Orbanz and Roy (2014) for more details). One can refer to Diaconis and Janson (2008) for more details on the connection between the work on exchangeable arrays and the work on graph limits. Note that this theorem has been used by Hoff (2008), Bickel and Chen (2009) and Bickel et al. (2011) in the context of graph modeling.

An issue with the graphon is that it is not identifiable. Indeed, if $\psi : [0,1] \to [0,1]$ is a measure preserving function, then the graphon $W(\psi(\cdot), \psi(\cdot))$ leads to the same probability distribution on the graph as $W(\cdot, \cdot)$. To tackle this issue, Bickel and Chen (2009) propose to impose monotonicity on the function g defined by $g(u_i) = \int W(u_i, u_i) du_j$ (see also Yang et al. (2014); Chan and Airoldi (2014)).

As far as estimation (of the underlying graphon function of an observed graph) is concerned, most of the techniques are based on the approximation of the W-graph by a

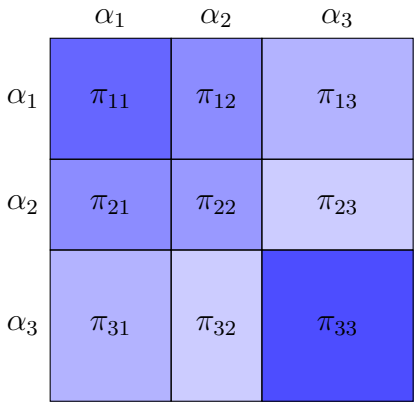


Fig. 1.5 Link between the graphon and SBM. Note that this is symmetric.

SBM (Airoldi et al., 2013; Gao et al., 2015; Latouche and Robin, 2016; Klopp et al., 2017), by computing group averages after finding groups of nodes, thus obtaining a block-wise constant graphon, and usually performing a smoothing step afterwards. For example, to obtain the grouping of nodes, Latouche and Robin (2016) use a Variational Bayes EM algorithm, and obtain an estimator of the graphon function by averaging stochastic block models with increasing number of blocks.

Some methods (Chatterjee, 2015; Yang et al., 2014) are based on a spectral method, the universal singular value thresholding (USVT) algorithm (Chatterjee, 2015).

Chan and Airoldi (2014) proposed an algorithm relying on the ordering of the nodes based on the observed degrees and on the smoothing of the histogram obtained from the sorted graph. Other methods have been introduced, see for example Zhang et al. (2017b) or Lloyd et al. (2012).

1.4 Node clustering techniques

In this section, we will review some techniques used to cluster vertices in a graph into groups of nodes with similar connection behaviour, and give some theoretical or empirical results for these techniques. Node clustering in networks has been intensively studied and we will not be able to be exhaustive, but will present some of the widely used methods. In particular, a lot of work has been done on community detection, i.e. when considering groups of vertices that have a high within-group connectivity and a low between-group connectivity, as on the left of Figure 1.4. Some methods can identify a more general structure with groups of nodes sharing a similar connection behaviour towards other nodes (in the same class or in any other class). For example, as in the middle of Figure 1.4, where we observe two classes, a group of two hubs and a class

of peripheral nodes, or as in the right of Figure 1.4 where we observe groups of nodes interacting mainly with nodes from other groups. Note that the term community is often used to talk about these more general classes, but we will stick to the strict definition of a community in this work. We will then refer to community detection when doing clustering for community structures and simply to clustering for the clustering of more general structures.

1.4.1 Spectral clustering

A commonly used technique in community detection is the spectral clustering, i.e. algorithms that cluster points using eigenvectors of matrices derived from the data. See Von Luxburg (2007) for more details on the spectral clustering.

The matrix we rely on in spectral clustering is a Laplacian matrix (or sometimes directly the adjacency matrix) (Chung and Graham, 1997). There are different definitions of such matrices, hence leading to different versions of the spectral clustering.

The use of graph Laplacians instead of the adjacency matrix is justified by its good properties that we will state right after, and enables a better detection of the clusters. In particular, using the normalised versions of the Laplacian allows to regularise the degree distribution and gives better results when the degree distribution is heterogeneous (Von Luxburg, 2007).

To describe this technique, let us first introduce a definition of the graph Laplacian matrix. We need to define the degree matrix D of a graph, that is the diagonal matrix with the weighted degrees $d_1, \dots, d_n$ on its diagonal, i.e. $d_i = \sum_{j=1}^n X_{ij} = \sum_{j=1}^n X_{ji}$[10]. We then define the (unnormalised) Laplacian matrix for a graph of adjacency matrix X and degree matrix D as

$$L = D - X. \tag{1.4.1}$$

This matrix is symmetric, and satisfies for any $u \in \mathbb{R}^n$

$$u^\top L u = u^\top D u - u^\top X u = \frac{1}{2} \sum_{1 \leq i,j \leq n} X_{ij}(u_i - u_j)^2.$$

It is then positive semi-definite, and its eigenvalues, denoted by $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, are nonnegative. Moreover, since $d_i = \sum_{j=1}^n X_{ij}$ for every $i \in [\![1, n]\!]$, we have that 0 is an eigenvalue with associated eigenvector $(1, \dots, 1)$. An important result about the Laplacian graph is that the multiplicity of this eigenvalue is equal to the number of

[10] The spectral clustering method is defined for undirected graphs.

connected components in the graph and the corresponding eigenspace is spanned by the indicator vectors of the components.

Note that in the case where we have k connected components in the graphs, the Laplacian matrix (as the adjacency matrix) has a block diagonal form (up to a reordering of the nodes). The principle of spectral clustering is that when we want to find communities, that we can see as "almost" connected components, we want to identify the "almost" block diagonal structure of the graph.

To partition the graph into Q clusters, the algorithm then consists of selecting the Q eigenvectors corresponding to the smallest nonzero eigenvalues[11] of the Laplacian matrix and performing a k-means clustering with Q groups on the lines of the $n \times Q$ matrix formed by these eigenvectors (see Algorithm 3).

Algorithm 3: Spectral clustering

input : An adjacency matrix X, a number of clusters Q
output : Q clusters $C_1, \dots, C_Q$ forming a partition of $\{1 \dots, n\}$

1 **Compute** the Laplacian matrix L as in (1.4.1);
2 **Compute** the Q eigenvectors $u_1, \dots, u_Q$ associated with the Q smallest eigenvalues of L;
3 **Let** $U \in \mathbb{R}^{n \times Q}$ be the matrix containing the vectors $u_1, \dots, u_Q$ as columns;
4 **Cluster** the n lines of U with the k-means algorithm into clusters $C_1, \dots, C_Q$;
5 **return** $C_1, \dots, C_Q$

Other common versions of the spectral clustering algorithm use normalised Laplacian matrices, that are defined as

$L_{\text{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} X D^{-1/2}$ or $L_{\text{rw}} = D^{-1} L = I - D^{-1} X$. It is mentioned in Von Luxburg (2007) that for regular graphs where most vertices have approximately the same degree, the different Laplacians are similar, and using any of them lead to similar clustering results, but that it is not the case if the degrees in the graph are very heterogeneous. Von Luxburg (2007) recommends using normalised rather than unnormalised spectral clustering, and in the normalised case to use the eigenvectors of L_{rw} rather than those of L_{sym} .

This technique is mainly used for community detection, but Rohe et al. (2011) introduced the absolute spectral clustering, based on a different definition of the Laplacian matrix that can have negative eigenvalues, and which then uses the absolute values of the eigenvalues, selecting the largest (in absolute value) positive and negative eigenvalues. This allows to recover structures more complex than communities, for example groups

[11] In practice, we apply spectral clustering on connected graphs, or on the connected components of a graph separately.

that are more likely to interact with each other than with themselves, or a mix of these two.

1.4.2 Modularity

An approach in community detection is the optimisation of a measure known as modularity (Newman and Girvan, 2004). This measure is proportional to the number of edges within groups minus the expected number in an equivalent network with random edges. It was first used in Newman and Girvan (2004) to evaluate the division obtained by a clustering algorithm, i.e. check that the considered clustering divides the nodes into communities by checking that the modularity is high enough. Considering a partition of the network in Q communities, the modularity is defined in Newman and Girvan (2004) as

$$\mathcal{Q} = \sum_{1 \leq q \leq Q} (e_{qq} - a_q^2),$$

where $E = (e_{qq'})_{1 \leq q,q' \leq Q}$ is the matrix of fractions of edges in the network that link vertices in community q to vertices in community q', and $a_q = \sum_{q'=1}^{Q} e_{qq'}$ is the sum of the row (or column) q of E, representing the fraction of edges that connect to vertices in community q. Large values of the modularity are then supposed to indicate that the network has a community structure. Newman (2004) then proposed to optimise the modularity to find communities. Unfortunately, it is too costly to optimise directly this quantity by computing it for every possible division of the network (an exhaustive search of all possible divisions would take at least an amount of time exponential in the number of nodes). So in order to obtain a result in reasonable time, one must use some approximate optimisation strategy. For example, Newman (2004) proposed a greedy algorithm running in $O((m+n)n)$ (m being the number of edges) that starts with a state in which each vertex form a community and repeatedly chooses communities to join together in pairs leading to the greatest increase (or smallest decrease) in $\mathcal{Q}$. They then obtain a dendrogram and can select the best cut by looking for the maximal value of $\mathcal{Q}$. Clauset et al. (2004) improved the running time by making use of suitable data structures to $O(n \log^2 n)$ for a sparse graph. A rather similar approach is the Louvain algorithm (Blondel et al., 2008), which starts by assigning a different community to each node of the network, and repeats the following two steps. The first step is the repetition of a node reassignment to another community (among the communities of the neighbours of this node), according to the greatest increase in the modularity, until there is no reassignment increasing the modularity left. The second step consists in building a new network whose nodes are the communities from the first step, by summing the weights of

the nodes between every two communities. The algorithm stops when there are no more changes. They obtain good results in terms of modularity value in shorter time. Traag (2015) proposed a faster version of the Louvain algorithm, in which in the first step of the algorithm, each node is moved to a random neighbour community, instead of the best neighbour community.

Simulated annealing can also be used (see Guimera and Amaral (2005b,a); Medus et al. (2005) among others) and gives good results, the main disadvantage of this approach being that it is slow. The optimisation of the modularity can also be based on the genetic algorithm (for example Li et al. (2010)) that is a heuristic inspired by the process of natural selection. Newman (2006) gives a reformulation of the modularity in terms of eigenvectors of a characteristic matrix for the network (the modularity matrix), leading to a spectral algorithm for community detection to divide the networks into two communities, and then repeat the algorithm in the case of more than two communities. A lot of other techniques have been introduced but we will not describe them all (for example see Duch and Arenas (2005),Wakita and Tsurumi (2007), He et al. (2016)). For more details see Danon et al. (2005) who performed a test comparing the performance of a large number of different community detection algorithms. Bickel and Chen (2009) defined a new modularity called the *likelihood modularity* and showed that modularities allow to recover the groups with probability tending to one, under some conditions on these modularities.

Even though this method has been widely used, it suffers from several limitations. It admits a resolution limit, i.e. the method fails to recover small communities in large networks (Fortunato and Barthélemy, 2007; Good et al., 2010), a degeneracy problem, i.e. there are at least an exponential (in n) number of distinct partitions whose modularity values are very close to the global maximum (Good et al., 2010). See for example Lancichinetti and Fortunato (2011) for some details about the limits of the method of modularity maximisation.

Research is still done to tackle this problem, see Chakraborty et al. (2017) for a review, or more recent articles Chen et al. (2018); Long (2019); Haq et al. (2019).

1.4.3 Other community detection methods

Another way to identify communities in a graph is to run dynamical processes on the graph, usually random walks. That method is based on the idea that if there are a lot of connections inside each community and only a few edges connect different communities together, a random walk is going be trapped in each community for some time before finding a way out and moving to another community. For example, Pons and Latapy (2005) obtain a hierarchical community structure by computing a measure of similarities

between nodes that is based on random walks on the graph (and that can be computed efficiently).

More algorithms for community detection have been introduced (see for example Nascimento and De Carvalho (2011); Yang et al. (2016); Hajek et al. (2016)).

Note that there also has been a lot of research on overlapping community detection, i.e. when one assumes that the groups may not form a partition of the nodes, the nodes being allowed to belong to more than one community (see for example Xie et al. (2013); Devi and Poovammal (2016)).

One can refer to Fortunato and Hric (2016) for a review of community detection in networks.

1.4.4 Model-based clustering

1.4.4.1 Maximum likelihood estimation in the SBM with a Variational EM algorithm

In the SBM, we can be interested in estimating the parameters of the model in order to describe our network. We can obtain them by using a modified version of the Expectation-Maximisation (EM) algorithm based on a variational approximation of the distribution of the latent variables given the observations. If we also want to obtain a clustering, we can then obtain the latent groups by using a maximum a posteriori estimation.

In this model, we cannot compute the Maximum Likelihood Estimator (MLE) except for very small values of n, because it involves a summation over all the Q^n possible latent configurations. Indeed, the log-likelihood is

$$\ell(\theta) \coloneqq \log \mathbb{P}_\theta(X) = \log\left(\sum_{z\in[\![1,Q]\!]^n} \mathbb{P}_\theta(X, Z=z)\right) = \log\left(\sum_{z\in[\![1,Q]\!]^n} e^{\log \mathbb{P}_\theta(X,Z=z)}\right)$$

where

$$\begin{aligned}\log \mathbb{P}_\theta(X, Z=z) &= \log \mathbb{P}_\theta(X \mid Z=z) + \log \mathbb{P}_\theta(Z=z) \\ &= \sum_{1\le q,l\le Q} \sum_{1\le i<j\le n} Z_{iq}Z_{jl}\left[X_{ij}\log \pi_{ql} + (1-X_{ij})\log(1-\pi_{ql})\right] \\ &\quad + \sum_{q=1}^{Q}\sum_{i=1}^{n} Z_{iq}\log \alpha_q, \end{aligned} \tag{1.4.2}$$

defining $Z_{iq} = \mathbb{1}_{Z_i=q}$ for every i and q. We neither can use the EM algorithm (often used in latent variables models) to approximate it because it involves the computation

of the conditional distribution of the latent variables given the observations which is not tractable. A common solution is to use the Variational Expectation-Maximisation (VEM) algorithm that optimises a lower bound of the log-likelihood (see for example Daudin et al. (2008)).

EM algorithm Let us first describe briefly the EM algorithm in a general case. This is an iterative algorithm introduced by Dempster et al. (1977), used to approximate maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. Assume that we observe the variable $X = (X_1, \ldots, X_n)$, that $Z = (Z_1, \ldots, Z_n)$ is a set of latent (i.e. unobserved) variables taking their values in a finite set and that θ is a vector of unknown parameters. We want to maximise the log-likelihood

$$\begin{aligned} \ell(\theta) &= \log \sum_z \mathbb{P}_\theta(X, Z = z) \\ &= \log \mathbb{P}_\theta(X, Z) - \log \mathbb{P}_\theta(Z \mid X). \end{aligned} \tag{1.4.3}$$

This quantity is often intractable due to the sum on all the possible configurations z. We start the algorithm with some initial value of the parameter $\theta^{(0)}$. Each iteration t of the EM algorithm is composed of two consecutive steps (E step and M step) and the algorithm stops when the relative difference between the estimates at two consecutive steps is small enough, or when a maximum number of iterations is reached.

E step: The E (Expectation) step consists of computing the quantity

$$\begin{aligned} Q(\theta|\theta^{(t-1)}) =& \mathbb{E}_{\theta^{(t-1)}} \left[\log \mathbb{P}_\theta \left(Z, X\right) \mid X\right] \\ =& \mathbb{E}_{\theta^{(t-1)}} \left[\log \mathbb{P}_\theta \left(X \mid Z\right) \mid X\right] + \mathbb{E}_{\theta^{(t-1)}} \left[\log \mathbb{P}_\theta \left(Z\right) \mid X\right]. \end{aligned} \tag{1.4.4}$$

M step: the M (Maximisation) step consists of maximising the quantity $Q(\theta|\theta^{(t-1)})$ with respect to θ to obtain $\theta^{(t)}$ the estimate at step t,

$$\theta^{(t)} = \arg\max_\theta Q(\theta|\theta^{(t-1)}).$$

It is proven that the log-likelihood increases at each iteration (Dempster et al., 1977). However, it can converge to a local maximum, and it is then common to run the algorithm multiple times with different initial values in order to obtain the global maximum.

Variational EM (VEM) algorithm In the SBM, the EM algorithm as described before is still intractable since we cannot compute the distribution of the latent variables given the observations because it is not factorised, so the computation of Q in (1.4.4) is out of reach. We will then rely on a variational approximation of this distribution. First, for any distribution $\mathbb{Q}$ on the latent variables Z, let us introduce two quantities, $\mathcal{H}(\mathbb{Q})$ the entropy of $\mathbb{Q}$, and $\mathrm{KL}(\mathbb{Q}, \mathbb{P}_\theta(\cdot \mid X))$ the Kullback-Leibler divergence from $\mathbb{P}_\theta(\cdot \mid X)$ to $\mathbb{Q}$

$$\mathcal{H}(\mathbb{Q}) = -\mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{Q}(Z)\right] = -\sum_z \mathbb{Q}(z) \log \mathbb{Q}(z) \tag{1.4.5}$$
$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}_\theta(\cdot \mid X)) = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathbb{Q}(Z)}{\mathbb{P}_\theta(Z \mid X)}\right] = \sum_z \mathbb{Q}(z) \log \frac{\mathbb{Q}(z)}{\mathbb{P}_\theta(z \mid X)}.$$

We can then rewrite the log-likelihood in (1.4.3) as follows, by taking the expectation with respect to $\mathbb{Q}$ on both sides of the equation,

$$\begin{aligned}\ell(\theta) &= \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X, Z)\right] - \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(Z \mid X)\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X, Z)\right] + \mathcal{H}(\mathbb{Q}) + \mathrm{KL}(\mathbb{Q}, \mathbb{P}_\theta(\cdot \mid X)).\end{aligned}$$

The idea of this algorithm is to find the distribution $\mathbb{Q}$ in a set of factorised distributions which minimises the quantity $\mathrm{KL}(\mathbb{Q}, \mathbb{P}_{\theta^{(t-1)}}(\cdot \mid X))$[12] and then to maximise $\mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_{\theta^{(t-1)}}(X, Z)\right] + \mathcal{H}(\mathbb{Q})$ in θ, this step now being tractable thanks to factorised form of $\mathbb{Q}$, approximating $\mathbb{P}_{\theta^{(t-1)}}(\cdot \mid X)$. We call the estimator obtained with this algorithm *variational estimator*. The quantity we optimise (in $\mathbb{Q}$ and θ) in the VEM algorithm is then

$$\mathcal{J}(\mathbb{Q}, \theta) = \ell(\theta) - \mathrm{KL}(\mathbb{Q}, \mathbb{P}_\theta(\cdot|X)) = \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X, Z)\right] + \mathcal{H}(\mathbb{Q}), \tag{1.4.6}$$

i.e. a lower bound of the log-likelihood, the Kullback-Leibler divergence being nonnegative. In the case of the SBM, we consider the factorised distributions of the form

$$\mathbb{Q}(Z) = \prod_{i=1}^{n} \mathbb{Q}(Z_i) = \prod_{i=1}^{n} \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}},$$

with $\tau_{iq} = \mathbb{Q}(Z_i = q) = \mathbb{E}_{\mathbb{Q}}[Z_{iq}]$ (with $Z_{iq} = \mathbb{1}_{Z_i = q}$) such that $\sum_{q=1}^{Q} \tau_{iq} = 1$ for any i. The algorithm is then tractable for the SBM, and simple expressions can be obtained

[12] That is equivalent to maximising $\mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_{\theta^{(t-1)}}(X, Z)\right] + \mathcal{H}(\mathbb{Q})$ with respect to $\mathbb{Q}$. Note that the distribution minimising this quantity over the set of all distributions is $\mathbb{P}_{\theta^{(t-1)}}(\cdot \mid X)$ leading to a Kullback-Leibler divergence equal to zero.

as follows. The quantity to optimise (1.4.6) is written (see Equation (1.4.2) and the definition of the entropy (1.4.5))

$$\begin{aligned}\mathcal{J}(\mathbb{Q},\theta) =& \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X,Z)\right] + \mathcal{H}(\mathbb{Q}) \\ =& \sum_{1\leq q,l\leq Q} \sum_{1\leq i<j\leq n} \tau_{iq}\tau_{jl}\left[X_{ij}\log \pi_{ql} + (1-X_{ij})\log(1-\pi_{ql})\right] + \sum_{q=1}^{Q}\sum_{i=1}^{n}\tau_{iq}\log\frac{\alpha_q}{\tau_{iq}}.\end{aligned}$$

Maximising this quantity in τ_{iq} gives that $\hat{\tau} = (\hat{\tau}_{iq})_{i,q}$ satisfies a fixed-point equation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{j=i+1}^{n}\prod_{l=1}^{Q}\left[\pi_{ql}^{X_{ij}}(1-\pi_{ql})^{1-X_{ij}}\right]^{\hat{\tau}_{jl}}.$$

Then maximising it with respect to $\theta = (\alpha, \pi)$ gives the closed-form expressions

$$\begin{aligned}\hat{\alpha}_q &= \frac{1}{n}\sum_{i=1}^{n}\tau_{iq} \\ \hat{\pi}_{ql} &= \frac{\sum_{1\leq i<j\leq n}\tau_{iq}\tau_{jl}X_{ij}}{\sum_{1\leq i<j\leq n}\tau_{iq}\tau_{jl}}.\end{aligned}$$

If we want to obtain an estimation of the class memberships, we can then take the maximum a posteriori, i.e.

$$\forall i \in [\![1,n]\!], \quad \hat{Z}_i = \arg\max_{q\in[\![1,Q]\!]} \tau_{iq}.$$

We recall that this method allows to recover a node clustering in a more general context than community structure. We described it for the binary SBM, but it can also be used for weighted SBM (Mariadassou et al., 2010).

Note that Gunawardana and Byrne (2005) obtained that in the general case, the VEM algorithm does not converge to local maxima of the likelihood, except for degenerate models but gives good results in practice for the SBM (Gazal et al., 2012), and we will also give some theoretical results later for the variational estimators in the SBM.

1.4.4.2 Other methods

Many other model-based clustering methods have been introduced, and we present briefly some of them.

A Bayesian version of the variational EM (Attias, 1999; Beal and Ghahramani, 2003) can also be used for the estimation of parameters in the SBM (Latouche et al., 2010, 2012; Aicher et al., 2013), based on an approximation of the joint probability distribution of

the latent variables Z and the parameters (α, π) given the observations X by a factorised distribution.

Still in the SBM, Channarond et al. (2012) proposed an algorithm to recover the clustering (and estimate the parameters), relying only on the degrees of the observed graph, called Largest Gaps algorithm. They obtained the consistency of this method.

Node clustering for bipartite graphs can be obtained based on the latent block model. In this model, the estimation can be based on variational EM, classification EM (inserting a classification step in which we find a partition) (see Govaert and Nadif (2008), Chapter 2 of Govaert and Nadif (2013)), stochastic EM with Gibbs sampling (Keribin et al., 2010, 2012), or Bayesian inference (Keribin et al., 2015; Wyse and Friel, 2012). Brault and Channarond (2016) introduce an adaptation of the Largest Gaps algorithm of Channarond et al. (2012) for the LBM. Concerning theoretical results, Mariadassou and Matias (2015) give sufficient conditions for the groups posterior distribution to converge to a Dirac mass located at the actual groups configuration, for every parameter in a neighborhood of the true one. Consistency and asymptotic normality results have been obtained for the maximum likelihood estimators and variational estimators in the LBM in Brault et al. (2020).

Clustering can also be performed based on the ERGM. For example, Salter-Townshend and Murphy (2015); Wang et al. (2019) introduce mixture models of ERGMs. The model of Salter-Townshend and Murphy (2015) is based on ego-networks, i.e. for each node, the network composed of the considered node and its neighbours, and all the edges between thoses nodes. The inference of these two models is based on the EM algorithm. Wang et al. (2019) use an online variant of the EM, alternatively assigning the nodes to groups and updating the parameter. Vu et al. (2013) propose a model for large networks. They assume dyad independence given the groups, and estimation is based on a VEM combined with a minorisation-maximisation algorithm (Hunter and Lange, 2004). See also Agarwal and Xue (2019) for clustering in weighted ERGMs.

Clustering can also be performed based on the latent position cluster model of Handcock et al. (2007), who propose two methods for estimation of parameters in this model. The first method is a two-stage maximum likelihood estimation, the first step consisting of assigning positions in the latent space based on a maximisation of the likelihood, and the second step consisting in finding a maximum likelihood estimator of the mixture model parameter conditionally on the latent positions from the first step, thanks to an EM procedure. The second method is a Bayesian approach using MCMC sampling.

1.4.5 Choice of the number of classes

In the models we considered in the previous section, we assume that the number of classes (i.e. the number of values taken by the latent variable representing the group membership) is known. However, in practice, we usually do not have this information. We then introduce briefly some of the existing criteria for the choice of a number of classes in a general context, and more generally for model selection, and we will then present two criteria for the special case of the SBM. For more details about information criteria, see Konishi and Kitagawa (2008).

1.4.5.1 General case

The AIC (Akaike Information Criterion) Introduced by Hirotugu Akaike (Akaike, 1973, 1974), the AIC relies on an asymptotic approximation of the Kullback-Leibler divergence of the estimated distribution from the unknown true distribution. It provides a tool for evaluating models in which the parameters are estimated by the maximum likelihood method. For this criterion, the Kullback-Leibler divergence is estimated using the empirical distribution of the observations, and the correction term for the bias of this estimator (induced by the fact that the same data is used both in the estimation of the model parameter and in the estimation of the expected log-likelihood in the Kullback-Leibler divergence) is approximated by the number of free parameters. For any considered model, the criterion is then defined as

$$\mathrm{AIC} = -2 \log L(\hat{\theta}) + 2k,$$

with k the number of free parameters in the considered model and $L(\hat{\theta})$ the maximum of the likelihood of the model. To choose the "best" model from a set of candidate models, we take the one minimising the AIC. It has been shown that the AIC tends to overestimate the number of groups in the case of mixture models.

The BIC (Bayesian Information Criterion) Introduced by Schwarz et al. (1978), the BIC is an evaluation criterion for models defined in terms of their posterior probability. As for the AIC, it is adapted for the evaluation of models estimated by using the maximum likelihood method. This criterion is based on an approximation of the marginal distribution of the observations x, that is given for any considered model with parameter θ by

$$p(x) = \int p(x \mid \theta)\pi(\theta)\, d\theta,$$

with π a prior distribution on θ. By using Laplace approximation method for the integral above, based on the idea that for a large enough number of observations, the integrand is concentrated in a neighborhood of the maximum likelihood estimator $\hat{\theta}$ and that the value of the integral depends on the behaviour of the function in this neighborhood, the criterion is then defined as the following approximation of $-2\log p(x)$

$$\text{BIC} = -2\log L(\hat{\theta}) + k\log(n),$$

where k is the number of free parameters of the model, $L(\hat{\theta})$ the maximum of the likelihood of the model and n the number of observations. As for the AIC, we choose the model minimising the BIC[13]. This criterion has some limitations. First, as for the AIC, n must be large enough for the approximation to be valid. Moreover, it has additional limitations in a model-based clustering context, as mentioned in Biernacki et al. (2000), when assessing the number of clusters. Indeed, for the approximation to be valid, the estimated vector parameter must be within the parameter space, which is not the case when the true model has a smaller number of components that the model we consider. They also raise the fact that, as for the AIC, this criterion does not take into account the clustering purpose, and that it tends to overestimate the number of components in the case of mixture models when the true distribution is not in the considered family of distributions.

The ICL (Integrated Classification Likelihood) The ICL is a criterion introduced by Biernacki et al. (1998) in order to circumvent the limitations of the BIC in a clustering context. Contrary to the previously introduced AIC and BIC, which were designed in a density estimation purpose, the ICL criterion has been derived in a clustering purpose. Indeed, this criterion is based on the complete (log-)likelihood, thus taking into account the clustering (i.e. the discrete latent variable z). They consider the integrated complete likelihood (also called integrated classification likelihood) that is given by the following expression for any considered model with parameter $\theta = (m, a)$ with m the mixing proportions parameter and a the parameter of the conditional distribution of the observations given the latent classes

$$p(x, z) = \int p(x, z \mid \theta)\pi(\theta)\, d\theta,$$

[13]Note that maximising the marginal distribution $p(x)$ of the observations (with respect to the model) is equivalent to maximising the posterior probability of the model (given the observations), assuming than the models are a priori equally probable.

with π a prior distribution on θ.

Assuming that $\pi(\theta) = \pi(a)\pi(m)$, they show that

$$\log p(x, z) = \log \left(\int p(x \mid z, a)\pi(a)\, da \right) + \log \left(\int p(z \mid m)\pi(m)\, dm \right)$$

and then apply the BIC approximation that is valid for the first term $\int p(x \mid z, a)\pi(a)\, da$, obtaining

$$\log \int p(x \mid z, a)\pi(a)\, da \approx \max_a \log p(x \mid z, a) - \frac{\lambda}{2} \log n$$

with λ the number of free components in a. Then, calculating the second term $\log(\int p(z \mid m)\pi(m)\, dm)$ by choosing a Jeffrey's noninformative prior for the proportions m, and replacing the latent data z with the maximum a posteriori (MAP) $\tilde{z}$, they propose the criterion

$$\begin{aligned} \text{ICL} = & \max_a \log p(x \mid \tilde{z}, a) - \frac{\lambda}{2} \log n \\ & + \log \Gamma\left(\frac{Q}{2}\right) + \sum_{q=1}^{Q} \log \Gamma\left(\tilde{n}_q + \frac{1}{2}\right) - Q \log \Gamma\left(\frac{1}{2}\right) - \Gamma\left(n + \frac{Q}{2}\right) \end{aligned}$$

with $\tilde{n}_q$ the number of $\tilde{z}_i$ equal to q and Γ the Gamma function. Moreover, when the $\tilde{n}_q$s are large, using the approximation of the Gamma function with the Stirling formula, they obtain

$$\begin{aligned} \text{ICL} &= \max_a \log p(x \mid \tilde{z}, a) + \max_m \log p(\tilde{z} \mid m) - \frac{\lambda}{2} \log n - \frac{Q-1}{2} \log n \\ &= \max_\theta \log p(x, \tilde{z} \mid \theta) - \frac{k}{2} \log n \end{aligned} \tag{1.4.7}$$

with $k = \lambda + Q - 1$ the number of free parameters in the model. This quantity can be maximised to select the most probable model. Note that considering the complete log-likelihood instead of the log-likelihood yields two penalisations in the first formulation of the ICL in (1.4.7), accounting for the distribution of the observations given the latent variables and for the distribution of the latent variables respectively. Even though it leads to a single penalisation $(k/2) \log n$ that is similar to that of BIC in this context, we will see why it is interesting when presenting an ICL for the SBM in the next section.

Note that in Biernacki et al. (2000), a second version of Biernacki et al. (1998), the authors derive a criterion in a slightly different way and obtain the following ICL that is

based on $\hat{\theta}$ the maximum likelihood estimator of θ

$$\text{ICL} = \log p(x, \tilde{z} \mid \hat{\theta}) - \frac{k}{2} \log n.$$

1.4.5.2 For the Stochastic Block Model

For the Stochastic Block Model, the AIC and BIC cannot be used because their computation involves the intractable log-likelihood of the observations. We present some criteria that were introduced in that context.

ICL criterion for the SBM An ICL criterion has been derived by Daudin et al. (2008) for model selection in the special case of the SBM, using the same technique as Biernacki et al. (1998) applied to the SBM, and based on the estimated $\tilde{z}$ (obtained from a VEM). They obtain

$$\text{ICL} = \max_{\theta} \log p(x, \tilde{z} \mid \theta) - \frac{1}{2} \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} - \frac{Q-1}{2} \log n,$$

where $Q(Q+1)/2$ is the number of free connectivity parameters (i.e. free parameters of the distribution of the observations given the latent variables), and $Q-1$ is the number of free proportion parameters (i.e. free parameters of the distribution of the latent variables), and $n(n-1)/2$ is the number of observations while n is the number of latent variables.

Note that the first part of the penalty comes from the approximation of the conditional log-likelihood of the observations given the latent variables (i.e. the part associated with the parameter π, whose dimension is $Q(Q+1)/2$, and the number of random variables is $n(n-1)/2$), and the second part of it comes from the approximation of the log-likelihood of the latent variables (i.e. the part associated with the parameter α, of dimension $Q-1$, and n group memberships random variables).

ILvb (Integrated Likelihood Variational Bayes) Later, Latouche et al. (2012) introduced another criterion, the ILvb, that is based on the log-likelihood, unlike the ICL. This quantity being intractable, they compute it using the variational approximation of the joint distribution of the latent variables and model parameters given the observations, the marginal log-likelihood then being approximated by the lower bound in the VEM algorithm. They then obtain a non asymptotic approximation (unlike the ICL again) through the variational Bayes EM algorithm (with Dirichlet and Beta prior on the group proportions and connectivity parameter). After convergence of the algorithm, the lower

bound of the log-likelihood is used to approximate this log-likelihood, and gives a criterion named ILvb, depending on the posterior probabilities and the normalising constants of the Dirichlet and beta distributions. Q is then chosen as the value maximising the ILvb.

Exact ICL Côme and Latouche (2015) introduced the exact ICL, obtaining an analytical expression of $\log p(x, z \mid Q)$, and proposed an algorithm for the maximisation of this quantity. Their algorithm allows to recover the clusters and number of classes at the same time, starting from an upper bound of Q, and allowing clusters to disappear. They prove that, setting respectively a Dirichlet and Beta priors for the distribution of Z and of X given Z, we obtain an exact ICL, depending only on X, Z, Q and the prior parameters. The best Z and Q could then be obtained by finding the maximum of this quantity. However, it cannot be maximised directly, as for each Q, the number of possible configurations is Q^n, hence the need for an algorithm to find the maximum. They introduce a greedy algorithm, starting with a SBM with the upper bound of number of classes. At each step, a single node is moved to another group if it increases the exact ICL (going to the group with maximal increase of the ICL). When a group is empty, it is removed. It stops when no swapping leads to an increase[14].

Criterion based on the empirical degrees As mentioned before, Channarond et al. (2012) proposed an algorithm to recover the clustering and estimate the parameters, and they propose in addition a selection criterion for the number of classes, relying only on the degrees of the observed graph. This criterion is based on the gaps between the mean degrees of the groups given by their algorithm.

1.4.6 Theoretical results

A lot of theoretical results have been obtained for the SBM, as for instance parameter estimator consistency or asymptotic normality results, or results on clustering error rate. We will mention a few of these results. Some theoretical results are presented in the introduction of Chapter 2 regarding parameter estimation, so we will focus here on other results on clustering.

We are interested in the *misclassification proportion*, i.e. the proportion of vertices classified in a wrong class by the considered algorithm (up to a permutation). If Z is the true configuration and $\hat{Z}$ the configuration obtained with the considered algorithm, the

[14]Note that a local maximum is obtained, and a common strategy is then to run the algorithm with multiple initialisations.

misclassification proportion is

$$r(Z,\hat{Z}) = \min_{\sigma\in\mathfrak{S}_Q} \frac{1}{n}\|Z-\sigma(\hat{Z})\|_0 := \min_{\sigma\in\mathfrak{S}_Q} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Z_i\neq\sigma(\hat{Z}_i)}$$

with $\mathfrak{S}_Q$ the set of permutations on $[\![1,Q]\!]$.

Minimax risk for the misclassification proportion Zhang and Zhou (2016) obtained a minimax risk decreasing exponentially with the number of nodes for the misclassification proportion for community detection (when the within-community probabilities are assumed to be larger than the between-communities probabilities). They work in a SBM framework but in which we assume that the true configuration z is a parameter of the model. Their result includes dense and sparse networks, equal and nonequal community orders and finite and growing number of communities. For example, to study sparse or dense networks, the connection probabilities can be as small as $O(1/n)$ and as large as a constant. Regarding the number of communities, it can be as large as $n/\log n$. The parameter space $\Theta(n,Q,a,b,\beta)$ contains the assignments z and the connection probabilities π such that the number of nodes in each of the communities is in $[n/(\beta Q),\beta n/Q]$ and $\pi_{qq}\geq a/n$ for any q and $\pi_{qq'}\leq b/n$ for any $q\neq q'$, with $\beta>1$ and bounded and with $0<b<a<(1-c_0)$ for some positive constant c_0. Defining I the Rényi divergence of order $1/2$ between two Bernoulli distributions of parameter a/n and b/n respectively, i.e.

$$I = D_{1/2}\left(\mathrm{Ber}\left(\frac{a}{n}\right)\middle\|\mathrm{Ber}\left(\frac{b}{n}\right)\right) = -2\log\left(\sqrt{\frac{a}{n}\frac{b}{n}}+\sqrt{1-\frac{a}{n}}\sqrt{1-\frac{b}{n}}\right),$$

their main result is as follows.

Theorem 1.4.1 (Zhang and Zhou (2016))**.** *Assume $nI/(Q\log Q)\to\infty$, then*

$$\inf_{\hat{z}}\sup_{\Theta(n,Q,a,b,\beta)}\mathbb{E}[r(z,\hat{z})] = \begin{cases}\exp\left(-(1+o(1))\frac{nI}{2}\right) & \text{if } Q=2\\ \exp\left(-(1+o(1))\frac{nI}{\beta Q}\right) & \text{if } Q\geq 3\end{cases}$$

where $1+\varepsilon_n\leq\beta\leq\sqrt{5/3}$ for some $\varepsilon_n=CQ/n$ with constant C large enough. In addition, if $nI/Q=O(1)$, there are at least a constant proportion of nodes mis-clustered, that is, $\inf_{\hat{z}}\sup_{\Theta(n,Q,a,b,\beta)}\mathbb{E}[r(z,\hat{z})]\geq c$, for some constant $c>0$.

They then derive thresholds on parameters to distinguish settings for which strong consistency (when the minimax rate is $o(1/n)$) or weak consistency (when the minimax

rate is $o(1)$) can be attained. They give a penalised likelihood procedure that achieves the minimax bound, which is however not tractable. Gao et al. (2017a) proposed an algorithm of community detection that achieves this optimal misclassification proportion under some regularity conditions, and that computes in polynomial time. This algorithm consists of applying an existing community detection algorithm that satisfies a certain weak consistency condition (for example spectral clustering) and then refining the result by optimising a local penalised likelihood for each node separately.

Resolution limit in the planted partition In the case of the affiliation model, it is obvious that the more the between-groups and the within-group connection probabilities are close, the more difficult it is to recover the clustering. Many results have been obtained on this, in particular in the case of planted partition, i.e. with two balanced groups, and with the within-group connection probability larger than the between-groups one. For example a conjecture by Decelle et al. (2011) which was proved later (see Mossel et al. (2012, 2015); Massoulié (2014); Mossel et al. (2018)) gives conditions for weak recovery (or detection, meaning that the obtained partition of the nodes is positively correlated with the true partition with probability converging to one as the number of nodes increases). When $\pi_{\text{in}} = a/n$ and $\pi_{\text{out}} = b/n$ with a and b constants (in a sparse case), it states that

- If $(a-b)^2 > 2(a+b)$, it is possible to cluster in a way correlated with the true partition.
- If $(a-b)^2 < 2(a+b)$, it is impossible to cluster in a way correlated with the true partition.

Abbe et al. (2016) obtained threshold for exact recovery (i.e. zero misclassification proportion with probability converging to one as the number of nodes increases). More precisely, if $\pi_{\text{in}} = a\log(n)/n$ and $\pi_{\text{out}} = b\log(n)/n$, then

- If $(a+b/2) - \sqrt{ab} > 1$, it is possible to recover the true partition with probability tending to one
- If $(a+b/2) - \sqrt{ab} < 1$, it is possible to recover the true partition with probability tending to one.

For more details, see for example Abbe (2018).

Some other results Regarding the spectral clustering method (and particularly the absolute spectral clustering), Rohe et al. (2011) give asymptotic results on the normalised graph Laplacian and its eigenvectors, allowing the number of clusters to grow with the number of nodes. They also provide bounds on the number of misclustered nodes, requiring an assumption on the degree distribution. Lei and Rinaldo (2015) prove consistency for the recovery of communities in the spectral clustering on the adjacency matrix, with milder conditions on the degrees, and also extend this result to degree corrected stochastic block models.

We recall that Bickel and Chen (2009) gave conditions on the modularity under which it allows to recover the communities with probability tending to one (exact recovery). Zhao et al. (2011) propose a method to "extract" communities one at a time and obtain the consistency (exact recovery) of their procedure under certain conditions. Zhao et al. (2012) introduce a framework for studying the consistency of community detection under the degree-corrected SBM based on different criteria (generalising Bickel and Chen (2009)).

1.5 Time-evolving networks

The first part of this work is devoted to dynamic (or time-evolving) networks, where the role or behaviour of the nodes in the network and the relationships between them are allowed to change over time. See Holme (2015) for an introduction to dynamic networks. These types of networks arise in many domains. Some obvious examples of dynamic networks are human contacts or proximity networks (obtained by recording when two people are close to each other) (Barrat and Cattuto, 2013), communication networks (such as e-mails or phone calls between people) (Saramäki and Moro, 2015; Ebel et al., 2002) or social networks (such as friendship networks). Another example is the transportation networks (for instance based on airlines connection or public transportation). Such networks also arise in neuroscience (for example networks representing the temporal correlations between brain regions based on functional magnetic resonance imaging (fMRI) data) (Sizemore and Bassett, 2018) or more generally in biology.

Such types of networks have been widely studied and can take many different forms. For example, we can consider discrete or continuous time, the interactions can be instantaneous or have a duration, we can assume that the nodes are present the entire observation time (and only the edges are evolving) or not, etc. We describe such networks in the following, distinguishing two main types, discrete-time (which will be our interest in this work) and continuous-time networks. We also present briefly some existing models

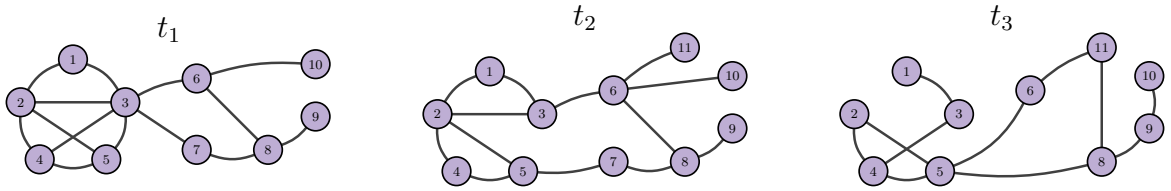


Fig. 1.6 Representation of a discrete-time dynamic network: in this network, the edges change over time and nodes can enter or exit the network

for dynamic networks. The model we are interested in, that is a discrete dynamic version of the SBM, will be described in Section 1.6, and some more discrete dynamic graph models based on the (static) SBM are presented in the introduction of Chapter 2. One can also refer to Kim et al. (2018) for a review of latent variables dynamic network models.

1.5.1 Discrete-time dynamic networks

A possibility when studying dynamic networks is to work in discrete time, i.e. to look at snapshots, or aggregated relational data over time ranges, in order to get a sequence of graphs. The data can be aggregated by splitting the observation period into T intervals and considering that an edge is present in the interval $t \in [\![1, T]\!]$ if it is present at any time in this interval to obtain a sequence of binary graphs. We can also consider a sequence of weighted graphs by doing as follows. In the case of instantaneous interactions between nodes, the edge between two nodes in the interval $t \in [\![1, T]\!]$ can represent the number of interactions observed in this interval. In the case of interactions of variable lengths, the edge between two nodes in the interval $t \in [\![1, T]\!]$ can represent the time of the interaction observed in this interval. Aggregating data obviously leads to a loss of information, and is not adapted to any kind of network.

An adapted representation of discrete-time dynamic networks is a sequence of graphs, as in Figure 1.6. See also Figure 1.7 for some representations of real world dynamic networks.

It is obviously important to take into account the evolutionary behaviour of the graphs, instead of just studying separate snapshots as unique graphs, the graphs being dependent over time.

A lot of models for discrete-time dynamic networks have been introduced. A common approach is to define dynamic extensions of existing (static) models. As mentioned before, discrete dynamic graph models based on the (static) SBM are introduced in Section 1.6 and in the introduction of Chapter 2. Note that some variants of the static SBM have also

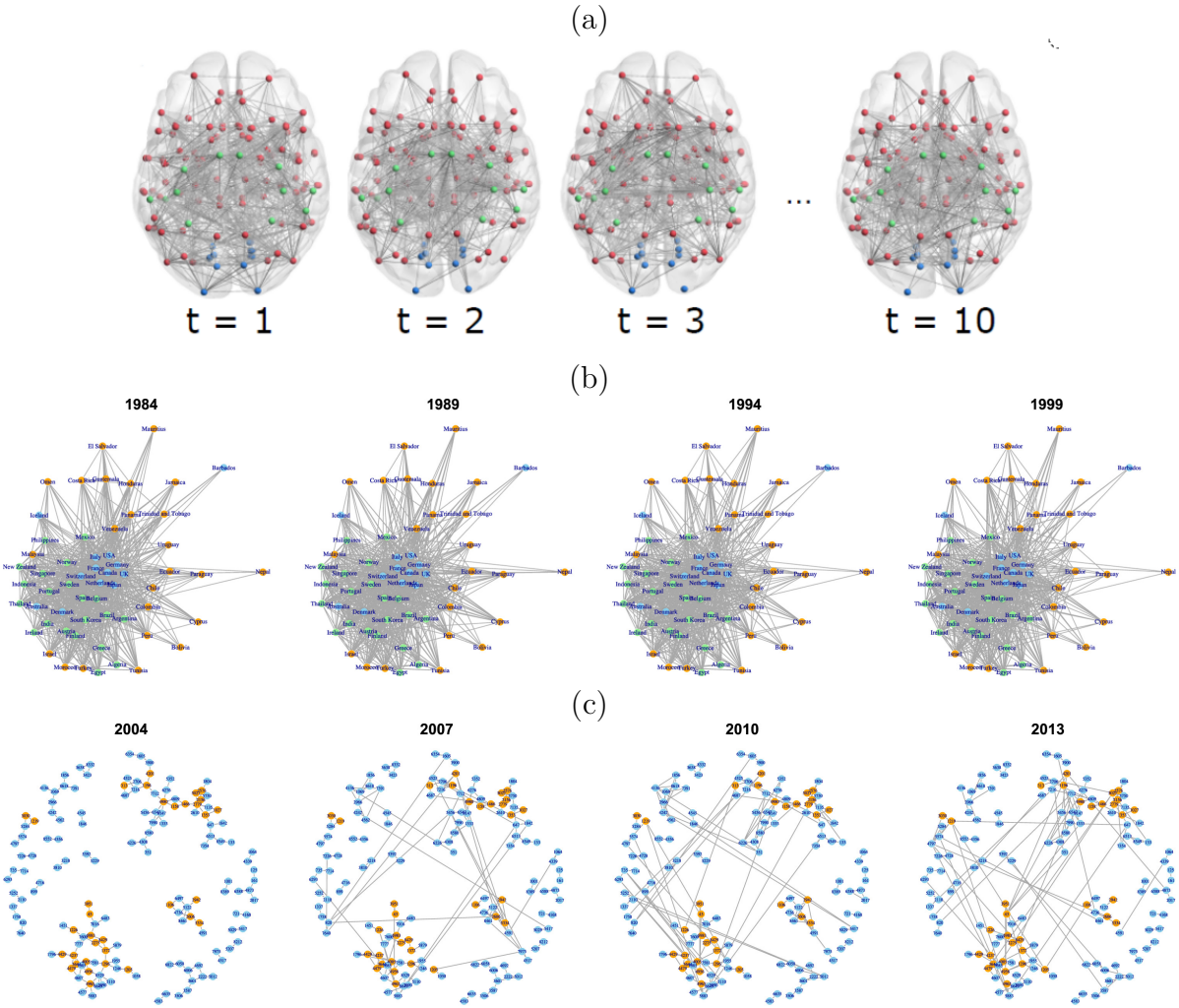


Fig. 1.7 Three discrete representations of dynamic networks: network (a) is from Sizemore and Bassett (2018) and networks (b) and (c) are from Lee et al. (2020). (a) represents functional connectivity between brain regions based on functional magnetic resonance imaging (fMRI) data (collected while the individual was learning to play a sequence of finger movements), (b) represents yearly international trade networks at four different years, and (c) represents yearly collaboration networks at a large research university at four different years. Note that the colors of the nodes in these graphs represent groups of nodes, which is not our interest for the moment.

been extended to the dynamic case, for example the MMSBM (see for instance Xing et al. (2010)). Such models can be used for node clustering, as their static versions. Extensions of the ERGM have also been introduced for dynamic networks. For example, Hanneke and Xing (2007); Hanneke et al. (2010) introduce the Temporal Exponential Random Graph Model (TERGM), for the purposes of studying social network evolution. In the TERGM, a Markov assumption is made on the evolution of the network, assuming that the graph at time t given the graph at time $t-1$ follows an ERGM, the statistic $S(X^t, X^{t-1})$ involving X^t and X^{t-1}, the adjacency matrices at times t and $t-1$. The statistics involving both X^t and X^{t-1} can include stability[15], reciprocity[16] or transitivity[17], these measures being particularly relevant in the context of social networks. See also Krivitsky and Handcock (2014), who introduced the Separable Temporal Exponential Random Graph Model (STERGM), adding an assumption of separability between the formation and duration of edges, obtaining a more convenient model. These models can be used for clustering (for example Lee et al. (2020) introduce a model-based clustering method for time-evolving networks based on a finite mixture of discrete time exponential-family random graph models).

Dynamic extensions of the LPM have also been introduced (Sarkar and Moore, 2006; Sewell and Chen, 2015; Friel et al., 2016; Sewell and Chen, 2016). For example, the model of Sarkar and Moore (2006) assumes that the nodes can move in the latent space between two time steps, with a Markov assumption on the movement of the nodes, and assuming that the observed graphs are independent given the locations of the nodes. Friel et al. (2016) introduce a model for the analysis of bipartite networks, extending the model of Sarkar and Moore (2006) by adding temporal evolution through Markovian dependence on the model parameters and on the edges (the edges are not conditionally independent anymore). Sewell and Chen (2016) extends the LPM for dynamic weighted graphs.

Note that the preferential attachment models (see Section 1.3.3) such as the Barabási–Albert model are generative discrete-time graph models, with a graph growing at each step of the algorithm.

[15] the tendency of an edge to stay present or absent between the two time steps

[16] the tendency of an edge from i to j to result in an edge from j to i at the next time step

[17] the tendency of an edge from i to j and from j to k to result in an edge from i to k at the next time step

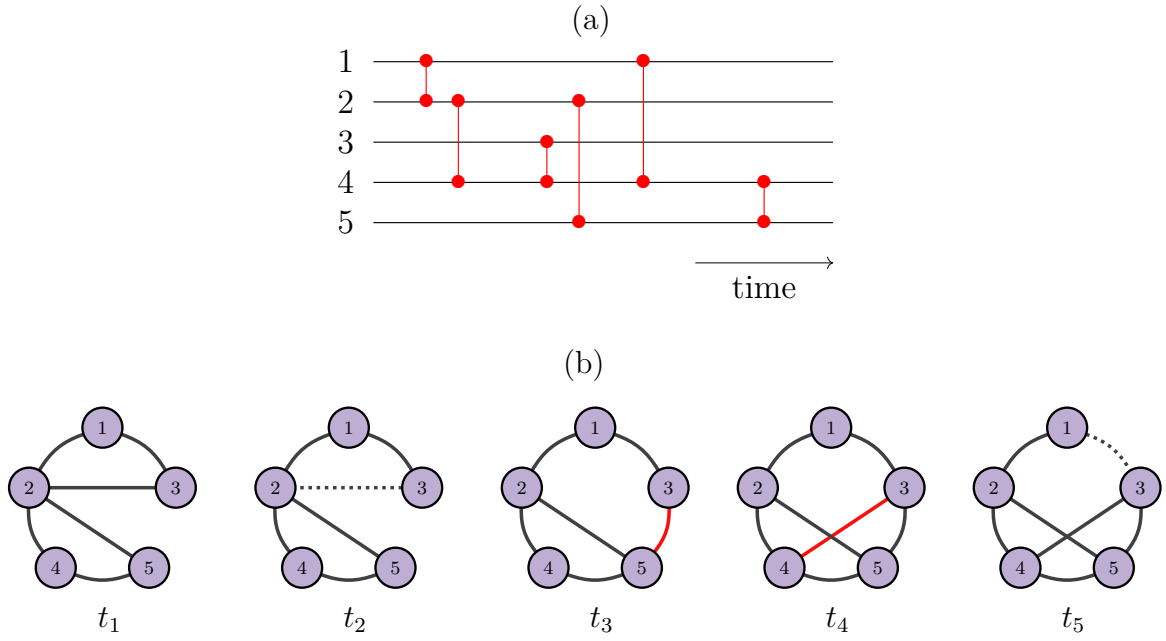


Fig. 1.8 Representation of two types of continuous time dynamic networks: (a) represents instantaneous contacts between the nodes and (b) represents a network in which edges are added (red lines) or suppressed (dotted lines) at time points $t_1, \ldots, t_5 \in \mathbb{R}$.

1.5.2 Continuous-time dynamic networks

We talked in the previous section about discrete-time graphs. They can be used to describe temporal evolution in graphs with low temporal resolution. However they may not be adapted to networks with high temporal resolution, such as e-mail networks.

There are different kinds of continuous-time dynamic graphs. For example, if we consider data of e-mails being sent between individuals, the interactions are instantaneous and can occur at any time. Another example would be phone calls between individuals, in which the interactions can occur at any time and have a continuous-time duration. The dynamics of the graph can also be defined by the addition or suppression of edges at time points. Different representations of such graphs can be used, depending on their form (see for example Figure 1.8).

A formalism has been introduced by Latapy et al. (2018) to represent and analyse continuous dynamic networks, defining the notions of stream graphs and link streams. A stream graph is defined by $S = (T, V, W, E)$ with V a finite set of nodes, T a set of time instants, $W \subseteq T \times V$ a set of temporal nodes (indicating the presence of the nodes over time) and $E \subseteq T \times V \otimes V$ a set of links (a link being allowed to be present only if the two involved nodes are present). In the case where all nodes are present all the time (i.e. there is no dynamics on nodes, but only on edges), then S is called a link stream and is

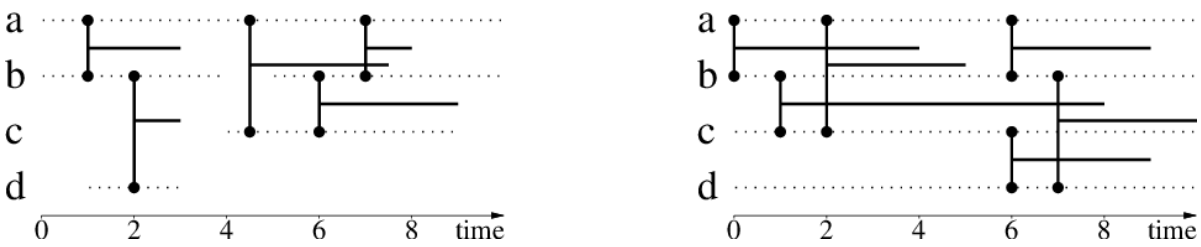


Fig. 1.9 A stream graph on the left, and a link stream on the right, both with 4 nodes $\{a, b, c, d\}$ (picture from Latapy et al. (2018))

denoted by $L = (T, V, E)$. See Figure 1.9 for a representation of these definitions (from Latapy et al. (2018)).

Number of models in that context are based on stochastic point processes. The use of such methods go way back in the analysis of dynamic networks, as Holland and Leinhardt (1977) introduced a continuous time Markov process to model changes in social relations (see also Wasserman (1980)). More recently, the stochastic actor-oriented model has been introduced by Snijders (1996, 2001), in which opportunities of a single edge modification happen, the time between two opportunities following an exponential distribution and one of the two nodes or the two nodes (called actors) involved have the "choice" (hence the name actor-oriented) of modifying or not the edge. Some approaches are based on doubly stochastic Poisson processes which are Poisson processes with random intensities (Butts, 2008). In particular, extensions of the SBM for interactions in continuous time based on such processes have been introduced (DuBois et al., 2013; Corneli et al., 2016; Matias et al., 2018). In such models, the interactions between two nodes given their latent groups follow an inhomogeneous Poisson process with intensity depending on the groups. Self-exciting Hawkes processes have been used (Blundell et al., 2012; Masuda et al., 2013; Junuthula et al., 2019), for which occurrence of events increases the probability of additional events in the future and then leading to some "concentration" of events in time, which is observed in some types of real-world networks.

See also Durante et al. (2016) who extend the LPM for continuous-time evolving networks.

1.6 Dynamic SBM with Markov membership evolution

In this work, we study the dynamic SBM as described in Yang et al. (2011) and Matias and Miele (2017), based on Markov chains modeling the temporal evolution of the group memberships over time, and on the (static) SBM.

More precisely, we consider a set of n vertices, split into Q latent classes, with Z_i^t the label of the i-th vertex at time t. Letting $Z_i = (Z_i^1, \dots, Z_i^T)$, the $\{Z_i\}_{1 \leq i \leq n}$ are independent and identically distributed and each Z_i is a homogeneous aperiodic and stationary Markov chain with transition matrix $\Gamma = (\gamma_{ql})_{1 \leq q,l \leq Q}$.

At each time t, we observe a binary graph of adjacency matrix $X^t = \{X_{ij}^t\}_{1 \leq i,j \leq n}$, following a stochastic block model so that, conditional on the latent groups $\{Z_i^t\}_{1 \leq i \leq n}$, the $\{X_{ij}^t\}_{1 \leq i,j \leq n}$ are independent Bernoulli random variables , i.e.

$$X_{ij}^t \mid Z_i^t = q, Z_j^t = l \sim \mathcal{B}(\pi_{ql}^t)$$

where $(\pi_{ql}^t)_{1 \leq q,l \leq Q, 1 \leq t \leq T} \in [0,1]^{Q^2 T}$ are the connectivity parameters. The model is thus parameterised by $\theta = (\Gamma, \pi)$, with $\Gamma = (\gamma_{ql})_{1 \leq q,l \leq Q}$ and $\pi = (\pi^t)_{1 \leq t \leq T}$ with $\pi^t = (\pi_{ql}^t)_{1 \leq q,l \leq Q}$. Note that we will consider both this model and the one where the connection probability parameter is fixed over time, i.e. where for every $t, t' \in [\![1, T]\!]$, we have $\pi^t = \pi^{t'} \coloneqq \pi$ (this is the model in Yang et al. (2011)) See Figure 1.10 for a representation of the temporal evolution in the model.

Note that we will assume that each Markov chain starts from an initial distribution $\alpha = (\alpha_1, \dots, \alpha_Q)$, that is its stationary distribution.

1.6.1 Label switching and identifiability in the dynamic SBM

An important issue arising when considering multiple graphs separately is label switching. Indeed, for the classical (static) SBM, identifiability can only be obtained up to label switching. This is not surprising, as this just means that the name or number attributed to each group is arbitrary and have no relevance. Then, when considering multiple graphs without any assumptions on their dependency, one may identify groups and/or connection and class membership parameters for each graph, but there is no simple way to identify the different groups across time or space. Some results have been obtained for the previously introduced dynamic SBM with Markov membership evolution (Matias and Miele, 2017; Becker and Holzmann, 2018). First, note that as mentioned in Matias and Miele (2017), imposing a Markov structure on the nodes labels is not sufficient to be able

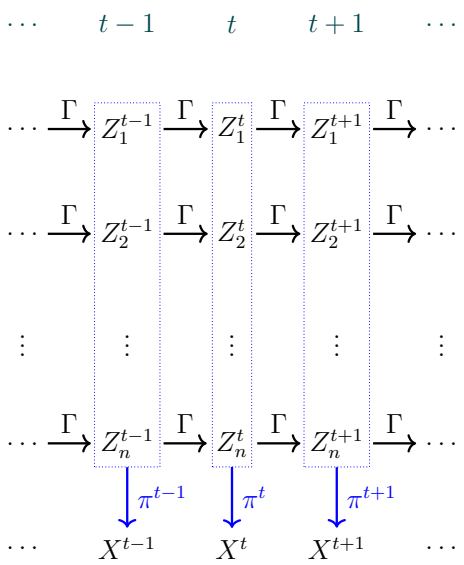


Fig. 1.10 Temporal evolution of the dynamic SBM. Γ is the transition matrix of the Markov chains (each Markov chain models the evolution of the group membership of a single node) and π is the connection probability matrix of the SBM.

to identify the parameters and some constraints have to be imposed on the parameters too. To give some intuition about this, see Figure 1.11 inspired from Matias and Miele (2017). This is a toy example, representing a stochastic block model at two time steps and composed of three groups, that are are a hub (node 6), peripheral nodes (nodes 7 to 10 at time t and nodes 1 to 5 at time $t+1$), and a community (nodes 1 to 5 at time t and nodes 7 to 10 at time $t+1$). Recall that we do not observe the groups, represented by the different colours in the figure. Two interpretations of this evolution could be done in the context of clustering with $Q = 3$ groups, leading to very different parameters for the dynamic SBM. The first one, at the top of Figure 1.11, is to consider that the groups are stable (i.e. that the nodes stayed in their original group), but the connection behaviour of these groups changed between the two time steps, the purple group going from a community behaviour to a peripheral one, and the green one, inversely, going from a peripheral behaviour to a community one. The second interpretation, at the bottom of Figure 1.11, is to consider that the three groups have a stable connectivity behaviour (the purple one being a community, the yellow one a hub, and the green one peripheral nodes), and that the nodes 1 to 5 changed groups from the purple one to the green one, and inversely the nodes 7 to 10 changed groups from the green one to the purple one (while node 6 stayed in the yellow group). The first interpretation corresponds to the case where the connectivity parameter π is very different between the two time steps (leading to groups behaving differently between these two time steps) and the transition

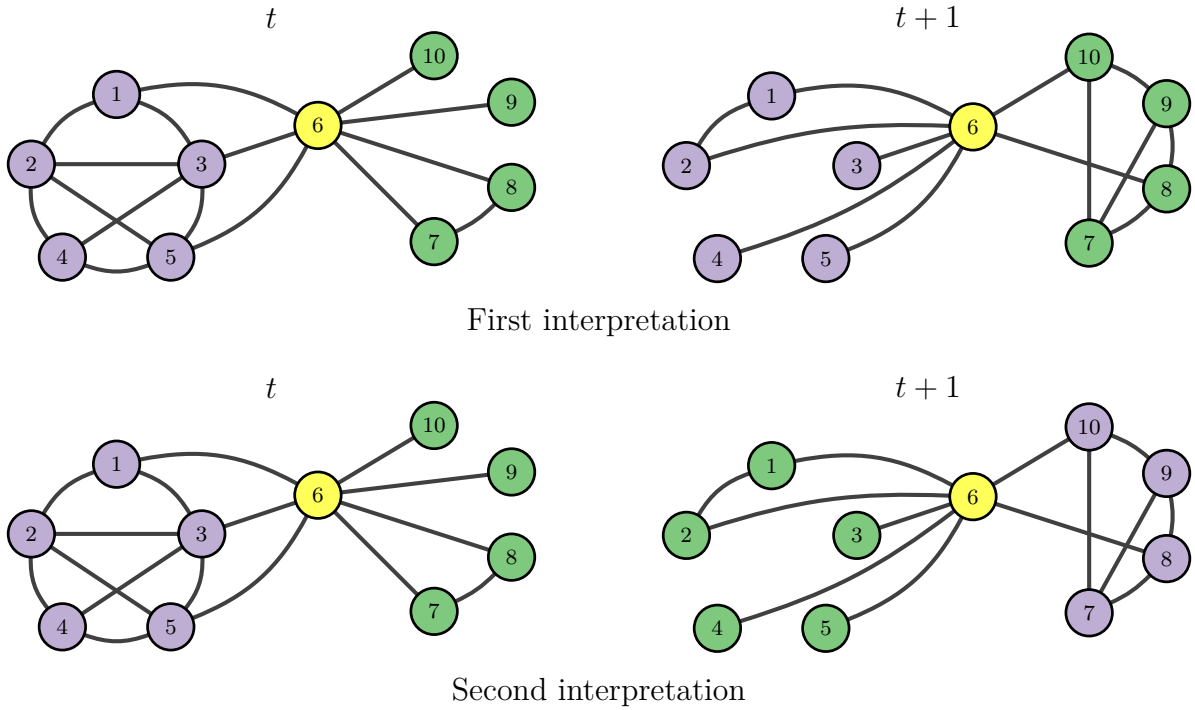


Fig. 1.11 Illustration of the label switching, graphs at time t (left) and $t+1$ (right)

matrix has large diagonal elements (leading to nodes staying in the same group with high probability). The second one corresponds to the case where the connectivity parameter is similar between two time steps (leading to a similar connection behaviour for each groups between these two time steps) and the transition matrix has high values for the transition from the purple to the green group and inversely (leading to nodes going from the purple to green group, and inversely). For the estimation to be feasible, Matias and Miele (2017) choose the second interpretation by imposing some constraint of stability over time to the connectivity parameter (this is also the choice made in Becker and Holzmann (2018)), and point out that this is a suitable choice for the analysis of social networks or contact data.

Identifiability in this dynamic SBM has been studied both for the binary and weighted graphs, but we focus here on the binary case. In Matias and Miele (2017), they obtain that the model is generically identifiable (up to global label switching, i.e. with the same permutation acting on the groups for every time step) from the distribution of the observed graphs, for n large enough, under the assumption that the within-group connectivity parameters are stable over time, i.e. that for every $t, t' \in [\![1, T]\!]$ and every $q \in [\![1, Q]\!]$, we have $\pi^t_{qq} = \pi^{t'}_{qq} \coloneqq \pi_{qq}$. Their identifiability result is generic, and the generic part concerns only the connectivity parameter π. This generic constraint arises in the proof of Theorem 2 in Allman et al. (2011) which requires among others in the

static case that the connection probabilities $(\pi_{qq'})_{1\leq q\leq q'\leq Q}$ are distinct. This assumption may not be necessary, however there are supplementary necessary assumptions to impose as without further assumptions, some cases may lead to non identifiability. We can refer to Matias and Miele (2017) for some intuition on the non identifiability in the affiliation case, and empirical evidence for the label switching between time steps.

We formulate and prove in the following two results, of non identifiability in a particular case, and of a necessary condition for the identifiability of the parameter.

Proposition 1.6.1. *If the parameter $(\Gamma, \pi^{1:T})$ of a dynamic SBM is such that there exists $1 \leq q_1 \neq q_2 \leq Q$ such that*

- $\gamma_{q_1q_1} = \gamma_{q_2q_2}$ *and* $\gamma_{q_1q_2} = \gamma_{q_2q_1}$,
- $(\gamma_{q_1q})_{q\in[\![1,Q]\!]\setminus\{q_1,q_2\}} = (\gamma_{q_2q})_{q\in[\![1,Q]\!]\setminus\{q_1,q_2\}}$ *and* $(\gamma_{qq_1})_{q\in[\![1,Q]\!]\setminus\{q_1,q_2\}} = (\gamma_{qq_2})_{q\in[\![1,Q]\!]\setminus\{q_1,q_2\}}$,
- $\pi_{q_1q_1} = \pi_{q_2q_2}$,

then it cannot be identified (up to global label switching).

Proof. To prove that the parameter cannot be identified, we exhibit a different parameter (different up to global label switching) leading to the same distribution for $X^{1:L}$. For any parameter $(\Gamma, \pi^{1:T})$, we define a parameter $(\tilde{\Gamma}, \tilde{\pi}^{1:T})$ such that $\tilde{\pi}^t = \pi^t$ if t is odd and $\tilde{\pi}^t = (\pi^t_{\sigma(q)\sigma(q')})_{1\leq q,q'\leq Q}$ if t is even with σ the permutation on $[\![1, Q]\!]$ permuting q_1 and q_2 only[18], and such that $\tilde{\gamma}_{q_1q_1} = \tilde{\gamma}_{q_2q_2} = \gamma_{q_1q_2}$, $\tilde{\gamma}_{q_1q_2} = \tilde{\gamma}_{q_2q_1} = \gamma_{q_1q_1}$, and $\tilde{\gamma}_{qq'} = \gamma_{qq'}$ if $q \notin \{q_1, q_2\}$ or $q' \notin \{q_1, q_2\}$. Let us also denote $\tilde{z}^{1:T}$ the transformation of any configuration $z^{1:T}$ such that $\tilde{z}^t = z^t$ at odd t values and $\tilde{z}^t := (\tilde{z}^t_1, \dots, \tilde{z}^t_n) = (\sigma(z^t_1), \dots, \sigma(z^t_n))$ at even t values. The proof that $(\Gamma, \pi^{1:T})$ and $(\tilde{\Gamma}, \tilde{\pi}^{1:T})$ induce the same distribution on $X^{1:T}$ then relies on the fact that for any $t \in [\![1, T]\!]$ and $i \in [\![1, n]\!]$, $\gamma_{z^t_i z^{t+1}_i} = \tilde{\gamma}_{\tilde{z}^t_i \tilde{z}^{t+1}_i}$, and $\alpha_{\tilde{z}^1_i} = \alpha_{z^1_i}$. We should also note that α is also the stationary distribution of the Markov chain of transition matrix $\tilde{\Gamma}$. Indeed, the assumptions on Γ implies that $\alpha_{q_1} = \alpha_{q_2}$. To see this, notice that for any transition matrix P with stationary distribution $p = (p_1, \dots, p_Q)$, permuting the rows q_1 and q_2 and the columns q_1 and q_2 of P is equivalent to a relabeling of the states (by permuting states q_1 and q_2) and then the corresponding stationary distribution is the same distribution p with permuted coordinates q_1 and q_2, i.e. $(p_{\sigma(1),\dots,\sigma(Q)})$[19]. Then noticing that with our assumptions, $\Gamma_\sigma := (\gamma_{\sigma(q)\sigma(q')})_{1\leq q,q'\leq Q} = \Gamma$, we have the equality of the stationary distributions $(\alpha_1, \dots, \alpha_Q) = (\alpha_{\sigma(1)}, \dots, \alpha_{\sigma(Q)})$, i.e. $\alpha_{q_1} = \alpha_{q_2}$. Then it is easily seen that α is also the stationary distribution of $\tilde{\Gamma}$[20].

[18] i.e. $\sigma(q_1) = q_2$, $\sigma(q_2) = q_1$ and $\sigma(q) = q$ for every $q \in [\![1, Q]\!] \setminus \{q_1, q_2\}$

[19] This can easily be written formally with the property $pP = p$ of the stationary distribution.

[20] by noticing that $\alpha\Gamma = \alpha$ implies $\alpha\tilde{\Gamma} = \alpha$

We can then write the probability distribution of $X^{1:T}$ under the parameter (Γ, π) as follows

$$\begin{aligned}
\mathbb{P}_\theta(X^{1:T}) &= \sum_{z^{1:T}\in[\![1,Q]\!]^{nT}} \prod_{t=1}^{T} \prod_{1\le i<j\le n} (\pi^t_{z_i^t z_j^t})^{X_{ij}^t}(1-\pi^t_{z_i^t z_j^t})^{1-X_{ij}^t} \prod_{i=1}^{n} \alpha_{z_i^1} \prod_{t=1}^{T-1} \gamma_{z_i^t z_i^{t+1}} \\
&= \sum_{z^{1:T}\in[\![1,Q]\!]^{nT}} \prod_{t=1}^{T} \prod_{1\le i<j\le n} (\tilde{\pi}^t_{\tilde{z}_i^t \tilde{z}_j^t})^{X_{ij}^t}(1-\tilde{\pi}^t_{\tilde{z}_i^t \tilde{z}_j^t})^{1-X_{ij}^t} \prod_{i=1}^{n} \alpha_{\tilde{z}_i^1} \prod_{t=1}^{T-1} \tilde{\gamma}_{\tilde{z}_i^t \tilde{z}_i^{t+1}} \\
&= \sum_{z^{1:T}\in[\![1,Q]\!]^{nT}} \prod_{t=1}^{T} \prod_{1\le i<j\le n} (\tilde{\pi}^t_{z_i^t z_j^t})^{X_{ij}^t}(1-\tilde{\pi}^t_{z_i^t z_j^t})^{1-X_{ij}^t} \prod_{i=1}^{n} \alpha_{z_i^1} \prod_{t=1}^{T-1} \tilde{\gamma}_{z_i^t z_i^{t+1}},
\end{aligned}$$

the last inequality being true because it is equivalent to sum over the $z^{1:T}$ in $[\![1,Q]\!]^{nT}$ or over the $\tilde{z}^{1:T}$ in $[\![1,Q]\!]^{nT}$. This proves that the parameter $(\tilde{\Gamma}, \tilde{\pi})$ leads to the same distribution of $X^{1:T}$, even though it is not equal to (Γ, π) up to global label switching of the groups.

□

For the case where the connectivity parameter is fixed over time ($\pi^t = \pi$), it is necessary (as it is well known in the static case) that there are no two equal rows in the connectivity matrix in order to distinguish the groups. Indeed, if it is not satisfied, i.e. $\exists q_1 \neq q_2 \in [\![1,Q]\!]$ such that $\pi_{q_1\bullet} = \pi_{q_2\bullet}$ (where $\pi_{q\bullet}$ denotes the q^{th} row of π for any $q \in [\![1,Q]\!]$), we can prove that for example, the parameter (Γ_σ, π) with $\Gamma_\sigma := (\gamma_{\sigma(q)\sigma(q')})_{1\le q,q'\le Q}$ (where σ is the permutation on $[\![1,Q]\!]$ permuting q_1 and q_2 only) leads to the same distribution for $X^{1:T}$ than (Γ, π)[21].

In the case where the connectivity parameter varies over time, however, it may not be necessary that the matrix π^t has all its rows distinct for every t. Indeed, it may be possible to identify the different groups even if some groups have the same behaviour at some (but not all) time steps, thanks to the transition matrix that is homogeneous over time. We formulate nonetheless a necessary assumption on this parameter.

Proposition 1.6.2. *In the dynamic SBM with varying connectivity parameter, it is necessary that* $\forall q \neq q' \in [\![1,Q]\!]$, $\exists t \in [\![1,T]\!]$ *such that* $\pi^t_{q\bullet} \neq \pi^t_{q'\bullet}$ *for the parameters* (Γ, π) *to be identifiable, where* $\pi^t_{q\bullet} = (\pi^t_{qq'})_{1\le q'\le Q}$.

Proof. If the assumption is not satisfied, i.e. $\exists q_1 \neq q_2 \in [\![1,Q]\!]$ such that $\forall t \in [\![1,T]\!]$, $\pi^t_{q_1\bullet} = \pi^t_{q_2\bullet}$ (implying that $\pi^t_{\bullet q_1} = \pi^t_{\bullet q_2}$, where $\pi^t_{\bullet q}$ denotes the q^{th} column of π^t for any $q \in [\![1,Q]\!]$), we cannot differentiate groups q_1 and q_2, these groups having the same connection behaviour. We can exhibit a specific parameter different from the true

[21] The proof is similar to that of Proposition 1.6.2 below.

parameter (up to global label switching) that leads to the same distribution of $X^{1:T}$. Indeed, we have, denoting by σ the permutation on $[\![1,Q]\!]$ permuting only q_1 and q_2, that $\pi_\sigma \coloneqq (\pi_{\sigma(q)\sigma(q')}) = \pi$. Then, the distribution of $X^{1:T}$ under the true parameter (Γ, π) can be written as

$$\begin{aligned}\mathbb{P}_\theta(X^{1:T}) &= \sum_{z^{1:T}\in[\![1,Q]\!]^{nT}} \prod_{t=1}^{T} \prod_{1\le i<j\le n} (\pi^t_{\sigma(z_i^t)\sigma(z_j^t)})^{X_{ij}^t}(1-\pi^t_{\sigma(z_i^t)\sigma(z_j^t)})^{1-X_{ij}^t} \prod_{i=1}^{n} \alpha_{z_i^1} \prod_{t=1}^{T-1} \gamma_{z_i^t z_i^{t+1}} \\ &= \sum_{z^{1:T}\in[\![1,Q]\!]^{nT}} \prod_{t=1}^{T} \prod_{1\le i<j\le n} (\pi^t_{z_i^t z_j^t})^{X_{ij}^t}(1-\pi^t_{z_i^t z_j^t})^{1-X_{ij}^t} \prod_{i=1}^{n} \alpha_{\sigma(z_i^1)} \prod_{t=1}^{T-1} \gamma_{\sigma(z_i^t)\sigma(z_i^{t+1})},\end{aligned}$$

noticing that it is equivalent to sum over the $z^{1:T}$ in $[\![1,Q]\!]^{nT}$ or over the $\sigma(z) = (\sigma(z_i^t))_{1\le i\le n, 1\le t\le T}$ in $[\![1,Q]\!]^{nT}$. Noticing that $(\alpha_{\sigma(1)},\dots,\alpha_{\sigma(Q)})$ is the stationary distribution associated with the transition matrix $\Gamma_\sigma \coloneqq (\gamma_{\sigma(q)\sigma(q')})_{1\le q,q'\le Q}$, this proves that the parameter (Γ_σ, π) leads to the same distribution of $X^{1:T}$. □

1.6.2 Estimation

As for the static SBM, we cannot compute the MLE except for very small values of n and T, and we neither can use the Expectation-Maximisation (EM) algorithm because it involves the computation of the intractable conditional distribution of the latent variables given the observations. Estimation can then be performed thanks to a VEM algorithm in that case too. The distribution of the latent variables given the observations is then replaced by a distribution that is factorised over the nodes (but not over time) (see Matias and Miele (2017)).

Let us denote $Z_{iq}^t = \mathbb{1}_{Z_i^t=q}$ for every t, i and q. Using the same approach as in Matias and Miele (2017) for the VEM algorithm in the dynamic SBM, we consider a variational approximation of the conditional distribution of the latent variable $Z^{1:T}$ given the observed variable $X^{1:T}$ in the class of probability distributions parameterised by $\chi = (\tau, \eta) = \left(\{\tau_{iq}^t\}_{t,i,q}, \{\eta_{iql}^t\}_{t,i,q,l}\right)$ of the form

$$\mathbb{Q}_\chi(Z^{1:T}) = \prod_{i=1}^{n} \mathbb{Q}_\chi(Z_i^1) \prod_{t=2}^{T} \mathbb{Q}_\chi(Z_i^t \mid Z_i^{t-1}) = \prod_{i=1}^{n} \left\{ \left[\prod_{q=1}^{Q} (\tau_{iq}^1)^{Z_{iq}^1}\right] \prod_{t=1}^{T-1} \prod_{1\le q,l\le Q} \left(\frac{\eta_{iql}^t}{\tau_{iq}^t}\right)^{Z_{iq}^t Z_{il}^{t+1}} \right\},$$

i.e. with $\mathbb{Q}_\chi$ such that $\mathbb{E}_{\mathbb{Q}_\chi}\left[Z_{iq}^t Z_{il}^{t+1}\right] = \eta_{iql}^t$ and $\mathbb{E}_{\mathbb{Q}_\chi}\left[Z_{iq}^t\right] = \tau_{iq}^t$. The quantity to optimise in the VEM algorithm is then

$$\mathcal{J}(\chi,\theta) = \ell(\theta) - \mathrm{KL}(\mathbb{Q}_\chi, \mathbb{P}_\theta(\cdot|X^{1:T})) = \mathbb{E}_{\mathbb{Q}_\chi}\left[\log \mathbb{P}_\theta(X^{1:T}, Z^{1:T})\right] + \mathcal{H}(\mathbb{Q}_\chi),$$

with $\ell(\theta)$ the log-likelihood.

1.6.3 Contributions in the dynamic SBM

In Chapter 2, we study the consistency of the maximum likelihood and variational estimators in the model described above. We prove the consistency (as the number of nodes and time steps increase) of the maximum likelihood and variational estimators of the model parameters, and obtain upper bounds on the rates of convergence of these estimators. We also explore the particular case where the number of time steps is fixed and connectivity parameters are allowed to vary. The assumption we make are that there are no two equal rows in the connection probability matrix[22] (that there are no two equal rows in any connection probability matrices for the finite time case), that the transition and connection probabilities are bounded away from 0 and 1 (excluding the sparse case), and for the finite time case, that the diagonal of the connection probability matrix is fixed over time and that its values are distinct.

The consistency of the transition matrix estimators requires an additional assumption that the connection parameter estimator converges at a rate that is $o(\sqrt{\log(nT)}/n)$, which we did not prove. Indeed, we only proved that the rate is faster than $r_{n,T}/n^{1/4}$ with $\{r_{n,T}\}_{n,T\geq 1}$ any sequence increasing to infinity. This is however a reasonable assumption, based on the convergence rates obtained in Bickel et al. (2013) for the estimators in the case of the static SBM, as they obtain a rate of n^{-1} in a non sparse case (when $\rho := \mathbb{P}(X_{ij} = 1)$ is constant) for the connection parameter estimator.

Chapter 2 is the reproduction of the article "Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model" published in Electronic Journal of Statistics (Longepierre and Matias, 2019).

1.7 Markov Random Fields (MRF)

The second part of this work, considering space-evolving networks, relies on Markov random fields. We will then give some definitions and results in this section. For a detailed introduction to Markov random fields (MRF), see for example Brémaud (2013) or Stoehr (2017). A MRF is a set of random variables having a Markov property described by an undirected graph. This is used in various fields like ferromagnetism, image analysis, epidemiology or geography.

[22] which is a necessary assumption for the classical SBM

1.7.1 Definition

Consider an undirected graph $\mathcal{G} = (V, E)$ and a set of random variables $Z = (Z^l)_{l \in V} := Z^{1:L}$ indexed by $V = \{1, \dots, L\}$ (the elements of V being called *locations* or *sites*) and each Z^l taking their values in a finite space Λ_l. Two locations forming an edge in $\mathcal{G}$ are said to be neighbours. Let us introduce some notation. For a subset A of V, let us denote by Z^A the set of random variables on A, and identically for any configuration z (realisation of the random variable Z) by z^A the restriction of this configuration to A. Let us denote by $-A$ the complement of A in V. Then, for any location l, we denote by Z^{-l} the set of random variables at every location but l, i.e. $Z^{-l} = \{Z^1, \dots, Z^{l-1}, Z^{l+1}, \dots, Z^L\}$, and identically $z^{-l} = \{z^1, \dots, z^{l-1}, z^{l+1}, \dots, z^L\}$. We define $\mathcal{N}(l)$ the neighbourhood of a location l as the set of locations that are adjacent to l in the graph $\mathcal{G}$, i.e. the set of neighbours of l in $\mathcal{G}$.

The set Z of random variables forms a Markov random field with respect to $\mathcal{G}$ if it satisfies the following (local) Markov property:

For any configuration $z = (z^l)_{l \in V}$ and any location l,

$$\mathbb{P}(Z^l = z^l \mid Z^{-l} = z^{-l}) = \mathbb{P}(Z^l = z^l \mid Z^{\mathcal{N}(l)} = z^{\mathcal{N}(l)}).$$

This means that the distribution of the random variable at a location l is independent of other locations conditional on the random variables at the neighbouring locations. Markov random fields have been especially studied with regular graphs $\mathcal{G}$, such as lattices, for example for image analysis or in ferromagnetism with the well known Ising model (modeling interacting spins).

At this point, a Markov random field is then described only by the different conditional probabilities at each location given their neighbours, which is not convenient to manipulate, and may not even imply the existence of a joint probability distribution. A fundamental result for MRF is the Hammersley-Clifford theorem, which states that the joint distribution of a MRF is factorised over the set of cliques of $\mathcal{G}$, when $\mathbb{P}(Z = z) > 0$ for every configuration z (see for example Besag (1974) or Clifford (1990)). More precisely, it states that under this positivity condition (i.e. if $\mathbb{P}(Z = z) > 0$ for every z), the Markov random field Z introduced before follows a Gibbs distribution of the form

$$\mathbb{P}_{\psi,\mathcal{G}}(Z = z) = \frac{1}{S(\psi, \mathcal{G})} \exp\left[-H(z, \psi, \mathcal{G})\right] \tag{1.7.1}$$

for some parameter ψ and some energy function H which decomposes into potential functions V_c associated to each clique $c \in \mathcal{C}$ (defining $\mathcal{C}$ the set of cliques of $\mathcal{G}$)

$$H(z, \psi, \mathcal{G}) = \sum_{c \in \mathcal{C}} V_c(z^c, \psi)$$

and with $S(\psi, \mathcal{G})$ a normalising constant (also called *partition function*). In the following, for simplicity of notation, we will drop $\mathcal{G}$ from the notation, since there will be no ambiguity.

Note that reciprocally, if $Z^{1:L}$ follows a Gibbs distribution with potential functions $\{V_c\}_{c \in \mathcal{C}}$ relative to a neighbourhood system $\mathcal{G}$[23], then $Z^{1:L}$ is a Markov random field with the neighbourhood graph $\mathcal{G}$. Moreover, its local specification is given by the formula

$$\mathbb{P}_{\psi,\mathcal{G}}(Z^l = z^l \mid Z^{\mathcal{N}(l)} = z^{\mathcal{N}(l)}) = \frac{\exp\left(-\sum_{c \in \mathcal{C}; l \in c} V_c(z^c, \psi)\right)}{\sum_{\lambda \in \Lambda_l} \exp\left(-\sum_{c \in \mathcal{C}; l \in c} V_c(z^c_{(l,\lambda)}, \psi)\right)}$$

where $z_{(l,\lambda)} = (z^{l'}_{(l,\lambda)})_{l' \in V}$ is defined as the configuration such that $z^l_{(l,\lambda)} = \lambda$ and $z^{l'}_{(l,\lambda)} = z^{l'}$ for $l' \neq l$, and with $\sum_{c \in \mathcal{C}; l \in c}$ the summation over the cliques of $\mathcal{G}$ containing l. Unlike the Hammersley-Clifford theorem, this result is quite straightforward (see for example Theorem 2.1 in Brémaud (2013)).

1.7.2 Autologistic model and Potts model

We present two classical MRF models that are the autologistic model (for binary data) and the Potts model (with variables taking their values in a finite set). We will focus in this work on the Potts model. These models have been used for instance in image analysis, solid-state physics and ferromagnetism.

We saw earlier with the Hammersley-Clifford theorem that under the positivity condition, the joint distribution of a MRF is a Gibbs distribution which factorises over cliques, decomposing into potential functions. We may assume that it is not necessary to consider the potentials on large cliques to correctly model the spatial dependency. Then, in the *autologistic model* introduced by Besag (1972) for binary data, we only consider the cliques of one and two nodes, i.e. potentials at one location, and interactions between two locations, and the potential functions associated with larger cliques are set to zero.

[23] i.e. such that $\mathcal{C}$ is the set of cliques of $\mathcal{G}$

The energy function then writes

$$H(z, \psi, \mathcal{G}) = \sum_{l \in V} V_l(z^l, \alpha) + \sum_{(l,l') \in E} V_{ll'}(z^l, z^{l'}, \beta),$$

with α the parameter on locations and β the parameter on pairs of neighbours, the parameter of this model being $\psi = (\alpha, \beta)$. Note that $\sum_{(l,l') \in E}$ is the sum over the edges of the undirected graph $\mathcal{G}$, and we consider each edge only once, i.e. as (l', l) is the same edge as (l, l') for any $l, l' \in [\![1, L]\!]$, we do not consider both (l, l') and (l', l).

More precisely, the energy function in the autologistic model can be written as

$$H(z, \psi, \mathcal{G}) = \alpha \sum_{l \in V} z^l + \sum_{(l,l') \in E} \beta_{ll'} z^l z^{l'}.$$

where originally $z^l \in \{0, 1\}$ for every $l \in V$. However, nowadays we preferably use this model with z^l taking their values in $\{-1, 1\}$ instead of $\{0, 1\}$. As mentioned in Pettitt et al. (2003), this parameterisation has the advantage of avoiding problems of non-invariance when states 0 and 1 are interchanged. In particular, the widely used Ising model (Ising, 1925) is an autologistic model with the random variables z^l taking their values in $\{-1, 1\}$. With this formulation, the autologistic model is a particular case (when $Q = 2$) of the Potts model introduced right after.

Potts model In the Potts model (Potts, 1952), contrary to the autologistic model, the variables of the random field take their values in a finite set $\{1, \ldots, Q\}$, and the energy function can be written as follows

$$H(z, \psi, \mathcal{G}) = \sum_{q=1}^{Q} \alpha_q \sum_{l \in V} \mathbb{1}_{z^l = q} + \beta \sum_{(l,l') \in E} \mathbb{1}_{z^l = z^{l'}}. \tag{1.7.2}$$

Note that a more general version of this model exists, when the parameter α (respectively β) can have different values at different locations (respectively at different pairs of neighbour locations), i.e.

$$H(z, \psi, \mathcal{G}) = \sum_{q=1}^{Q} \sum_{l \in V} \alpha_q^l \mathbb{1}_{z^l = q} + \sum_{(l,l') \in E} \beta^{ll'} \mathbb{1}_{z^l = z^{l'}}.$$

However, in this version of the model, the number of parameters is at least of the order of the number of random variables L. In this work, we will then consider the definition

of the Potts model in (1.7.2), as we will be interested in parameter estimation (based on a single realisation of a Potts model).

The parameter $\alpha = (\alpha_q)_{1 \leq q \leq Q}$ is the parameter of the external field, i.e. the latent variables are more likely to take values associated with large values of the parameter α. In particular, if $\alpha_q = 0$ for all $q \in [\![1, Q]\!]$, all the Q states are equally probable (a priori). The parameter β determines the strength of interaction between two neighbour locations. If β is positive, the model encourages latent variables at neighbour locations to have the same value, and on the contrary, a negative β encourages the random variables at neighbour locations to have different values.

A constraint can be imposed on the external field parameter for identifiability purposes, since adding the same constant to each component of α leads to the same distribution. Then, one could for example impose that $\sum_{q=1}^{Q} \alpha_q = 0$. We will give more details about identifiability of the Potts model in Section 3.3.

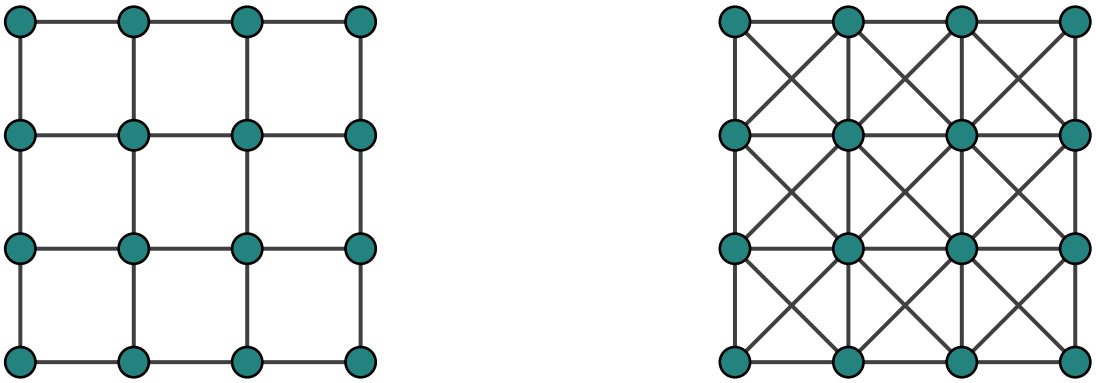
Note that for any l, the conditional probability of Z^l given the value $\tilde{z}^{-l}$ (or identically given the values $\tilde{z}^{\mathcal{N}(l)}$ of its neighbours) simply writes as follow

$$\mathbb{P}_\psi(Z^l \mid Z^{-l} = \tilde{z}^{-l}) = \mathbb{P}_\psi(Z^l \mid Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)}) = \frac{\exp\left(\alpha_{Z^l} + \beta \sum_{l' \in \mathcal{N}(l)} \mathbb{1}_{Z^l = \tilde{z}^{l'}}\right)}{\sum_{q=1}^{Q} \exp(\alpha_q + \beta \sum_{l' \in \mathcal{N}(l)} \mathbb{1}_{q = \tilde{z}^{l'}})}. \quad (1.7.3)$$

This quantity can be computed easily.

1.7.3 Strength of interaction and phase transition

The parameter β of the Potts model controls the strength of association between neighbour locations. If β is equal to 0, the locations are independent, and the bigger the β, the higher the probability that the variables at neighbour locations are equal. A particularity of the Potts model is that it exhibits a phenomenon known as phase transition (between a disordered phase and an ordered phase), if β becomes "too" large (exceeds a critical value), large parts of the random field (if not the whole field) are equal to the same value. This phenomenon has been widely studied (in physics for example). In particular, critical values for β have been obtained on lattices for Ising and Potts models. Note that these critical values depend on the graph $\mathcal{G}$ and tend to increase with the number of neighbours (degrees of the nodes). For example, the critical value is higher for a second



First order lattice Second order lattice

Fig. 1.12 Two different location graphs, that are first order and second order lattices

order lattice[24] than for a first order lattice[25] (see Figure 1.12). See Figure 1.13 illustrating the phenomenon for the Ising model on a first order lattice.

We do not give more details on phase transition here but one can refer to Georgii (2011); Duminil-Copin (2015).

In practice, when working on synthetic data in Chapter 3 we will visually check that we are not in a case where the strength of interaction is so strong that the configuration is "frozen".

1.7.4 Simulation with a Gibbs sampler

An important question is the sampling from a Gibbs distribution. This is not straightforward to draw realisations of this joint distribution because of its complexity. For this task, we can use a Gibbs sampler (Geman and Geman, 1984), which is an algorithm used to generate realisations of a joint distribution by starting from an initial configuration and updating the components one at a time, using the conditional probability distributions (see Algorithm 4).

Note that this Gibbs sampler is sometimes called systematic scan (or deterministic or sequential scan) Gibbs sampler, meaning that a fixed order is selected (in this case $1, \dots, L$) and the components z^l are updated in that specific order. A random scan Gibbs sampler can also be used, in which the component to update is selected randomly (from a uniform distribution) at each iteration.

It has been established that the Gibbs sampler converges to the wanted distribution (Geman and Geman, 1984). However, this convergence can be slow, especially for large β

[24] where each location (except those on the boundary) has for neighbours the 8 closest locations
[25] where each location (except those on the boundary) has for neighbours the 4 closest locations

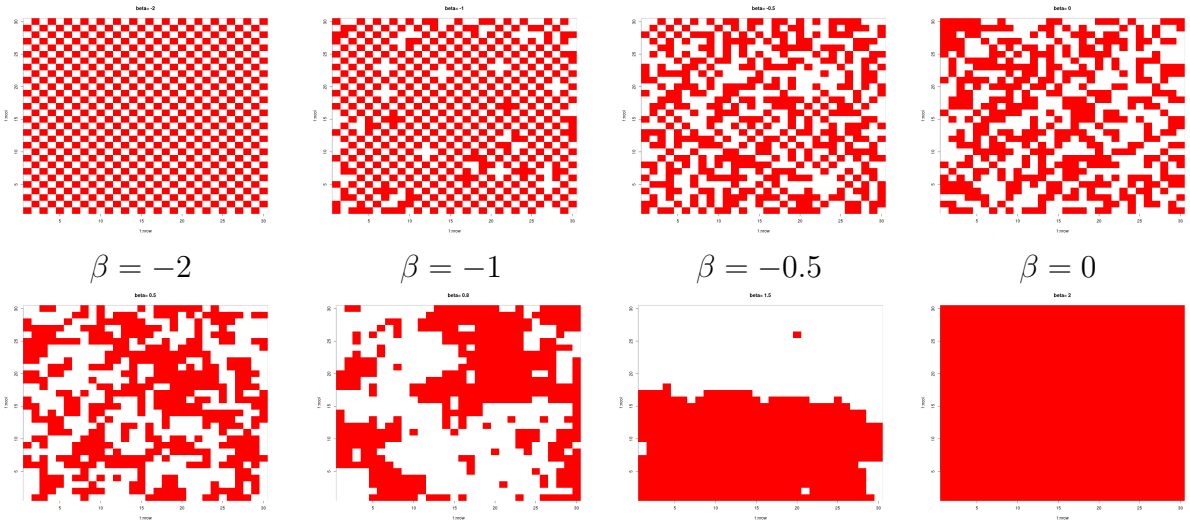


Fig. 1.13 Ising model for different values of β, on a first order lattice, with no external field (i.e. $\alpha_{iq} = 0$ for every i, q). The phase transition phenomenon can be observed for large values of β, leading to large parts of the lattices with the same value. Inversely, for negative values of β with large absolute value, neighbours tend to have different values, leading to a check pattern.

(for example when β is above the critical value of the phase transition). Other methods have been introduced, that can speed up the simulation, such as the modified random scan Gibbs sampler of Liu (1996), or the Swendsen-Wang algorithm introduced by Swendsen and Wang (1987), allowing updates of large parts of the field simultaneously, by incorporating auxiliary "bond" variables on pair of neighbours in $\mathcal{G}$.

Note that Friel and Rue (2007) proposed an exact sampling method for Markov random fields on small enough lattices. This is based on the recursive algorithm of Reeves and Pettitt (2004) which was introduced for the computation of the normalising constant, and is based on an appropriate factorisation of the unnormalised joint probability (due to the Markov property) reducing the complexity of the summation and giving an exact computation of the normalising constant for lattices up to about 20 rows.

1.7.5 Likelihood estimation and approximations

When studying the likelihood of such a model, for example for computing maximum likelihood estimates, one faces the problem of intractability of the normalising constant, this constant requiring a summation over Q^L configuration. Some methods have been introduced in order to circumvent this problem, mainly approximations neglecting some

Algorithm 4: Gibbs sampler

input : A number of iterations M, a parameter ψ
output : A realisation of the random variable $Z \sim \mathbb{P}_\psi$

Initialise an arbitrary configuration $z^{(0)} = (z^{(0)1}, \dots, z^{(0)L})$;
for $m = 1$ **to** M **do**
  for $l = 1$ **to** L **do**
    Draw $z^{(m)l}$ from the conditional distribution $\mathbb{P}_\psi(Z^l \mid z^{(m)1}, \dots, z^{(m)l-1}, z^{(m-1)l+1}, \dots, z^{(m-1)L})$ (see (1.7.3));
  end
end
return $z^{(M)}$

dependencies between the random variables. We will especially be interested in mean field and mean field like approaches.

1.7.5.1 Mean field approximation

The mean field approximation originated in statistical mechanics (see for example Chandler (1987)) where it was used to approximate the mean of a MRF. It consists of approximating the intractable Gibbs distribution by a simpler distribution, that can be factorised over locations, which will resolve the intractability of the normalising constant. This approximation consists of neglecting the fluctuations of the neighbours of each location by setting their values to their mean values. Before writing this approximation, let us introduce a new (equivalent) notation for the MRF $Z^{1:L}$. Let us denote $\mathbf{Z}^l = (Z^l_q)_{1\le q\le Q} := (\mathbb{1}_{Z^l=q})_{1\le q\le Q} \in [0,1]^Q$ for every $l \in [\![1, L]\!]$ and $\mathbf{Z}^{1:L} = (\mathbf{Z}^l)_{1\le l\le L}$. Let us also introduce an equivalent definition of the energy function in the Potts model

$$H(\mathbf{z}^{1:L}, \psi, \mathcal{G}) = \sum_{q=1}^{Q} \alpha_q \sum_{l=1}^{L} z^l_q + \beta \sum_{(l,l')\in E} \sum_{q=1}^{Q} z^l_q z^{l'}_q.$$

The mean-field approximation of the Gibbs distribution of $Z^{1:L}$ then writes as follows

$$\mathbb{P}^{\mathrm{MF}}_\psi(\mathbf{Z}^{1:L} = \mathbf{z}^{1:L}) = \prod_{l=1}^{L} \mathbb{P}_\psi(\mathbf{Z}^l = \mathbf{z}^l \mid \mathbf{Z}^{-l} = \mathbf{m}^{-l}) = \prod_{l=1}^{L} \mathbb{P}_\psi(\mathbf{Z}^l = \mathbf{z}^l \mid \mathbf{Z}^{\mathcal{N}(l)} = \mathbf{m}^{\mathcal{N}(l)}), \tag{1.7.4}$$

where $\mathbf{m}^{1:L} = (\mathbf{m}^1, \dots, \mathbf{m}^L)$[26] are the mean values of the variables $\mathbf{Z}^{1:L}$, and where

$$\mathbb{P}_\psi(\mathbf{Z}^l = \mathbf{z}^l \mid \mathbf{Z}^{\mathcal{N}(l)} = \mathbf{m}^{\mathcal{N}(l)}) = \frac{\exp(\sum_{q=1}^Q \alpha_q z_q^l + \beta \sum_{l' \in \mathcal{N}(l)} \sum_{q=1}^Q z_q^l m_q^{l'})}{\sum_{q=1}^Q \exp(\alpha_q + \beta \sum_{l' \in \mathcal{N}(l)} m_q^{l'})}. \tag{1.7.5}$$

Now remains the question of the computation of these unknown means, which was originally the purpose of this method. To obtain an approximation of these means that we will denote by $\bar{\mathbf{z}}^{1:L} = (\bar{\mathbf{z}}^1, \dots, \bar{\mathbf{z}}^L)$, as mentioned in Celeux et al. (2003), we rely on the self consistency condition, stating that the mean obtained based on the mean field approximation must be equal to the mean used to define this approximation. This means that for any location l, when fixing the other values at their mean $\bar{\mathbf{z}}^{-l} = (\bar{\mathbf{z}}^1, \dots, \bar{\mathbf{z}}^{l-1}, \bar{\mathbf{z}}^{l+1}, \dots, \bar{\mathbf{z}}^L)$[27], the expectation of $\mathbf{Z}^l = (Z_q^l)_{1 \leq q \leq Q}$ must be equal to $\bar{\mathbf{z}}^l$, i.e. $\bar{z}_q^l = \mathbb{E}_\psi^{\text{MF}}[Z_q^l] = \mathbb{E}_\psi[Z_q^l \mid \mathbf{Z}^{\mathcal{N}(l)} = \bar{\mathbf{z}}^{\mathcal{N}(l)}]$ for every q and l, i.e. $\bar{\mathbf{z}}$ must satisfy the fixed point equation

$$\bar{\mathbf{z}} \coloneqq (\bar{\mathbf{z}}^1, \dots, \bar{\mathbf{z}}^L) = \begin{cases} \mathbb{E}_\psi[\mathbf{Z}^1 \mid \mathbf{Z}^{\mathcal{N}(1)} = \bar{\mathbf{z}}^{\mathcal{N}(1)}] \\ \vdots \\ \mathbb{E}_\psi[\mathbf{Z}^L \mid \mathbf{Z}^{\mathcal{N}(L)} = \bar{\mathbf{z}}^{\mathcal{N}(L)}]. \end{cases}$$

Conditions for the existence have been discussed for example in Wu and Doerschuk (1995). When the solution exists, it can be computed iteratively.

Note that the mean-field approximation is equivalently the minimiser of the Kullback-Leibler divergence from the true distribution over the set of probability distributions that factorise over locations (see for example Chandler (1987); Peyrard (2001); Vignes (2007); Blei et al. (2017)), i.e. $\bar{\mathbf{z}}$ is the minimiser of $\text{KL}(\mathbb{Q}_\tau, \mathbb{P}_\theta(\cdot))$ with respect to τ, where $\mathbb{Q}_\tau$ is defined as the factorised (over the locations) distribution for $Z^{1:L}$ such that $\mathbb{Q}_\tau(Z^l = q) \coloneqq \mathbb{E}_{\mathbb{Q}_\tau}[Z_q^l] = \tau_q^l$. Indeed, we can see that computing the derivatives of the Kullback-Leibler divergence with respect to the components of τ and setting these derivatives equal to zero leads to the same fixed point equation as above. The mean-field approximation of the Gibbs distribution is then a variational approximation as described above in the VEM algorithm.

1.7.5.2 Mean field like approximation

The mean field approximation method has then been extended to mean field like approximations, in which we still neglect the fluctuations of the neighbours, their values

[26] with $\mathbf{m}^l = (m_1^l, \dots, m_Q^l)$ for every $l \in [\![1, L]\!]$

[27] with $\bar{\mathbf{z}}^l = (\bar{z}_q^l)_{1 \leq q \leq Q}$ for every l

being fixed to a value, as in (1.7.4), but not necessarily to the mean. These values can be the mode or a realisation of the random variable (Celeux et al., 2003). This will be our interest in Chapter 3 where it will be described more thoroughly.

1.7.5.3 Other methods

Pseudolikelihood The first method (apart from the coding technique (Besag, 1974, 1975), that have not been used much) is the pseudolikelihood, introduced by Besag (1975), which approximates the joint probability distribution by the product of the conditional distributions at each location, given all the other locations (thus given the neighbours). The pseudolikelihood is written

$$\mathrm{PL}(Z^{1:L}) = \prod_{l=1}^{L} \mathbb{P}_\psi(Z^l \mid Z^{-l}) = \prod_{l=1}^{L} \mathbb{P}_\psi(Z^l \mid Z^{\mathcal{N}(l)}),$$

where each term of the product can be computed easily (see Equation (1.7.3)). An estimator of the distribution parameter can then be obtained by maximising the pseudolikelihood. Note that this is not a valid probability distribution (unless the Z^ls are independent). Contrarily to the mean field like approaches, the neighbours are not fixed to constants, but are still random. This method is convenient because the estimator is easy to compute, and consistency and asymptotic results have been obtained (Gidas, 1988; Comets, 1992; Guyon and Künsch, 1992), but it has been shown that it does not always lead to good estimates, namely when the interaction is strong (see for example Geyer (1991) and Friel and Pettitt (2004)).

Some variations of the pseudolikelihood have been introduced to tackle its issues, such as the generalised pseudolikelihood (Huang and Ogata, 2002) or second order pseudolikelihood (Mase, 1995). See also Huang and Ogata (1999), who carry out a single Newton-Raphson step starting from the maximum pseudo-likehood.

Composite likelihood The composite likelihood (Lindsay, 1988) extends the pseudolikelihood by approximating the joint distribution by the product of tractable joint distributions of variables of a small number of locations. See Varin et al. (2011) for a review of composite likelihood methods in a general context. The composite likelihood can be defined as (Asuncion et al., 2010)

$$\mathrm{CL}(Z^{1:L}) = \prod_{m=1}^{M} \mathbb{P}_\psi(Z^{A_m} \mid Z^{B_m}),$$

where M is an integer smaller than L and $\{A_m\}_{1\leq m\leq M}$ and $\{B_m\}_{1\leq m\leq M}$ are sets of subsets of $[\![1, L]\!]$, such that $A_m \neq \emptyset$ and $A_m \cap B_m = \emptyset$. In particular, when $M = 1$, $A_1 = [\![1, L]\!]$ and $B_1 = \emptyset$, this gives the likelihood and when $M = L$, $A_m = \{m\}$ and $B_m = [\![1, L]\!] \setminus \{m\}$ (or identically $B_m = \mathcal{N}(m)$) for every $m \in [\![1, L]\!]$, the composite likelihood is the pseudolikelihood. When $B_m = \emptyset$ for every $m \in [\![1, M]\!]$, the composite likelihood is a product of marginal distributions and is usually called *marginal composite likelihood*. On the contrary, when for every $m \in [\![1, M]\!]$, $B_m = [\![1, L]\!] \setminus A_m$, the composite likelihood is called *conditional composite likelihood*. Note that in the case of spatial lattice processes (and a fortiori on a general spatial graph) such as the Potts model we consider, no marginal distributions can be computed, but conditional composite likelihood can be used (Okabayashi et al., 2011; Friel, 2012; Stoehr and Friel, 2015).

Okabayashi et al. (2011) show that composite likelihood gives better results than the pseudolikelihood approach, but in certain situations gives less satisfying results than maximum likelihood approximated using Markov chain Monte Carlo (MCMC) (see next paragraph).

Markov chains Monte Carlo methods Rather than replacing the likelihood with another tractable criterion as before, some methods have focused on the approximation of the maximum likelihood using Markov Chain Monte Carlo methods.

For example, Younes (1988) proposes to compute an approximate maximum likelihood estimator using a stochastic gradient algorithm. At each iteration, a small step is taken in the direction of the approximated gradient (based on a Gibbs sampler). Geyer and Thompson (1992) propose an algorithm to approximate the maximum likelihood based on a direct approximation of the likelihood from a MCMC sample (using a Metropolis algorithm or a Gibbs sampler) and its maximisation. Their procedure is iterative, and at each step, the approximated likelihood (based on a MCMC sample from the distribution with the current parameter) is maximised in a fixed neighbourhood of the current parameter. This is because the approximation of the likelihood is not good for parameters far from the one used for sampling (for a sample of reasonable size), so the maximisation cannot be done on the whole parameter space from a sample based on a single parameter. Descombes et al. (1999) introduce a MCMC algorithm based on the conjugate gradient principle, introducing a heuristic criterion to define a neighbourhood of the current parameter in which the MCMC approximation is robust.

Note that such methods may require a lot of computation time.

Other methods Other methods have been introduced to circumvent the difficulties caused by the intractability of the distribution. See the introduction of Chapter 3 for some existing methods. We also talk briefly about methods based on posterior distributions computations and not maximum likelihood, which is also problematic for the same reasons.

1.7.6 Hidden Markov random field

In this work, we will consider hidden Markov random field, in the sense that we do not observe the value of the Markov field, but that of a random variable $X^{1:L} = (X^1, \dots, X^L)$, the $\{X^l\}_{1 \leq l \leq L}$ being independent given $Z^{1:L}$, and $X^l \mid Z^l$ following a distribution of a given form, with a (usually) unknown parameter π. The conditional distribution of $X^{1:L}$ given $Z^{1:L}$ then factorises as follows

$$\mathbb{P}_\pi(X^{1:L} \mid Z^{1:L}) = \prod_{l=1}^{L} \mathbb{P}_\pi(X^l \mid Z^l).$$

The usual stakes in this situation are the recovery of the latent variables, and the estimation of the parameters of the Gibbs distribution, denoted by ψ, and/or of π the parameter of the emission distribution, i.e. the distribution of $X^{1:L}$ given $Z^{1:L}$. We will denote by θ the whole parameter, i.e. $\theta = (\psi, \pi)$, where $\psi = (\alpha, \beta)$. In the case of a hidden Markov random field, the problem of parameter estimation is even more complicated than in the case of an observed field. We will talk about some methods, and particularly the mean field and mean field like EM algorithm.

1.7.7 EM with mean field or mean field like approximation

In this work, the method we are interested in is the EM algorithm combined with a mean field (Zhang, 1992) or a mean field like approximation (Celeux et al., 2003) in order to approximate the MLE in a hidden MRF. First of all, we should point out the fact that we cannot maximise the likelihood directly because the computation of this quantity involves a summation over all the Q^L possible latent configurations, in addition to the normalising constant of the Gibbs distribution of the latent variables being intractable. We neither can use the EM algorithm to approximate it because the quantity to optimise

in this algorithm

$$\begin{aligned} Q(\theta|\theta^{(t-1)}) =& \mathbb{E}_{\theta^{(t-1)}}\left[\log \mathbb{P}_\pi\left(X^{1:L} \mid Z^{1:L}\right)\middle| X^{1:L}\right] \\ &+ \mathbb{E}_{\theta^{(t-1)}}\left[\log \mathbb{P}_\psi\left(Z^{1:L}\right)\middle| X^{1:L}\right] \\ :=& Q_1(\pi|\theta^{(t-1)}) + Q_2(\alpha,\beta|\theta^{(t-1)}). \end{aligned} \tag{1.7.6}$$

involves the computation of the conditional distribution of the latent variables $Z^{1:L}$ given the observations $X^{1:L}$ (appearing both in Q_1 and Q_2 in (1.7.6)) which is not tractable, and of the intractable normalising constant in the distribution of $Z^{1:L}$ (appearing in Q_2 in (1.7.6)).

The quantities $Q_1(\pi|\theta^{(t-1)})$ and $Q_2(\alpha,\beta|\theta^{(t-1)})$ can be written respectively as

$$Q_1(\pi|\theta^{(t-1)}) = \sum_{l=1}^{L}\sum_{z^l} \log \mathbb{P}_\pi\left(X^l \mid Z^l = z^l\right)\mathbb{P}_{\theta^{(t-1)}}\left(Z^l = z^l|X^{1:L}\right) \tag{1.7.7}$$

and

$$\begin{aligned} Q_2(\alpha,\beta|\theta^{(t-1)}) =& -\log S(\psi) + \sum_{q=1}^{Q}\alpha_q \sum_{l=1}^{L}\mathbb{P}_{\theta^{(t-1)}}\left(Z^l = q\middle| X^{1:L}\right) \\ &+ \beta \sum_{(l,l')\in E}\mathbb{P}_{\theta^{(t-1)}}\left(Z^l = Z^{l'}\middle| X^{1:L}\right). \end{aligned} \tag{1.7.8}$$

It is important to remember that $S(\psi)$ depends on α and β and we cannot ignore it when maximising $Q_2(\alpha,\beta|\theta^{(t-1)})$ with respect to α and β.

A variation of the EM algorithm has then been introduced to circumvent the problems of intractable distributions, relying on a mean field or mean field like approximation (described in Section 1.7.5) for both the distribution of the latent variables and of the latent variables given the observations. In such methods, at each step of the EM algorithm, we compute the mean (or mode or we simulate a configuration) of the conditional distribution of the latent variables given the observations. Then, in criterion (1.7.6), both intractable distributions (of the latent variables given the observations, and of the latent variables) are replaced by their mean field (resp. mean field like) approximations based on this expectation (resp. mode or simulated configuration). This allows to compute the conditional expectation and to get rid of the intractable normalising constant.

As pointed out in Celeux et al. (2003), if we want the approximate distributions (of the latent variables, and of the latent variables given the observations) to satisfy the Bayes rule, these two approximations must be based on the same simulated configuration (or mean or mode). The authors mention that it is more reasonable to base the approximation on

the conditional distribution (of the latent variables given the observations) $\mathbb{P}_\theta(Z^{1:L} \mid X^{1:L})$ rather than on the distribution of the latent variables $\mathbb{P}_\psi(Z^{1:L})$, as it takes the observations directly into account. See also the appendix of Celeux et al. (2001) for reasons dissuading from using the mean field approximation based on the distribution of $Z^{1:L}$.

We present here the algorithm when using the approximation based on a simulated configuration (that is called simulated field EM or simulated EM), but it can be defined similarly using the mean or mode. At each step t of the algorithm, we simulate a configuration $\tilde{z}^{1:L}$ from the conditional distribution of the latent variables given the observations $\mathbb{P}_{\theta^{(t-1)}}(Z^{1:L} \mid X^{1:L})$, under the parameter $\theta^{(t-1)} = (\psi^{(t-1)}, \pi^{(t-1)})$ obtained at step $t-1$. This simulation can be obtained using a Gibbs sampler (Algorithm 4), such that at each iteration $m \in [\![1, M]\!]$, for each $l \in [\![1, L]\!]$, $z^{(m)l}$ is simulated from the distribution

$$\begin{aligned}&\mathbb{P}_{\theta^{(t-1)}}(Z^l \mid X^l, z^{(m)1}, \dots, z^{(m)l-1}, z^{(m-1)l+1}, \dots, z^{(m-1)L})\\ \propto&\mathbb{P}_{\pi^{(t-1)}}(X^l \mid Z^l)\mathbb{P}_{\psi^{(t-1)}}(Z^l \mid z^{(m)1}, \dots, z^{(m)l-1}, z^{(m-1)l+1}, \dots, z^{(m-1)L}).\end{aligned}$$

Then an EM step is performed, with approximate distributions (of the latent variables, and of the latent variables given the observations). More precisely, the distribution of $Z^{1:L}$ appearing in Q_2 is approximated by the following distribution

$$\mathbb{P}^{\tilde{z}}_\psi(Z^{1:L} = z^{1:L}) \coloneqq \prod_{l=1}^{L} \mathbb{P}_\psi(Z^l = z^l | Z^{-l} = \tilde{z}^{-l}) = \prod_{l=1}^{L} \mathbb{P}_\psi(Z^l = z^l | Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)}) \quad (1.7.9)$$

where $\mathbb{P}_\psi(Z^l = z^l \mid Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)})$ is given by the formula in (1.7.3), and the distribution of the latent variable $Z^{1:L}$ given the observations $X^{1:L}$ (that appears in the expectation in both Q_1 and Q_2) is approximated by (using the Bayes formula and (1.7.9))

$$\begin{aligned}\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^{1:L} = z^{1:L} \mid X^{1:L}) =& \frac{\mathbb{P}_{\pi^{(t-1)}}(X^{1:L} \mid Z^{1:L} = z^{1:L})\mathbb{P}^{\tilde{z}}_{\psi^{(t-1)}}(Z^{1:L} = z^{1:L})}{\sum_{z^{1:L}} \mathbb{P}_{\pi^{(t-1)}}(X^{1:L} \mid Z^{1:L} = z^{1:L})\mathbb{P}^{\tilde{z}}_{\psi^{(t-1)}}(Z^{1:L} = z^{1:L})}\\ =& \prod_{l=1}^{L} \frac{\mathbb{P}_{\pi^{(t-1)}}(X^l \mid Z^l = z^l)\mathbb{P}_{\psi^{(t-1)}}(Z^l = z^l \mid Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)})}{\sum_{z^l} \mathbb{P}_{\pi^{(t-1)}}(X^l \mid Z^l = z^l)\mathbb{P}_{\psi^{(t-1)}}(Z^l = z^l \mid Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)})}\\ =& \prod_{l=1}^{L} \mathbb{P}_{\theta^{(t-1)}}(Z^l = z^l \mid Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)}, X^l)\\ =& \prod_{l=1}^{L} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l = z^l \mid X^l).\end{aligned}$$

Then, we perform an EM iteration using these approximate distributions. The quantities we want to maximise are now $\tilde{Q}_1(\pi|\theta^{(t-1)})$ and $\tilde{Q}_2(\alpha,\beta|\theta^{(t-1)})$, that are the approximations (using the approximate distributions) of $Q_1(\pi|\theta^{(t-1)})$ and $Q_2(\alpha,\beta|\theta^{(t-1)})$ respectively, that are given by

$$\tilde{Q}_1(\pi|\theta^{(t-1)}) = \sum_{l=1}^{L}\sum_{z^l} \log \mathbb{P}_\pi\left(X^l \mid Z^l = z^l\right) \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}\left(Z^l = z^l|X^l\right)$$

and

$$\tilde{Q}_2(\alpha,\beta|\theta^{(t-1)}) = \sum_{l=1}^{L}\sum_{z^l} \log \mathbb{P}_\psi\left(Z^l = z^l \mid Z^{\mathcal{N}(l)} = \tilde{z}^{\mathcal{N}(l)}\right) \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}\left(Z^l = z^l|X^l\right).$$

These quantities are tractable, and we can then update the parameter by maximising them thanks to numerical approximations if needed with respect to π and to α and β

$$\theta^{(t)} = (\arg\max_{\alpha,\beta} \tilde{Q}_2(\alpha,\beta|\theta^{(t-1)}), \arg\max_{\pi} \tilde{Q}_1(\pi|\theta^{(t-1)})).$$

1.7.8 Other methods

Other methods have been introduced to estimate the parameters and/or the classification in hidden Markov random fields, that can be based for example on Monte Carlo techniques, EM algorithm, pseudolikelihood, composite likelihood... See the introduction of Chapter 3 for the introduction to some of these methods.

1.7.9 Choice of the number of classes for hidden MRF

In the context of HMRF, the classical criteria for the choice of the number of classes (i.e. the number of states of the latent variable Z) such as the BIC or ICL introduced in Section 1.4.5 cannot be computed as it relies on intractable quantities (the maximised log likelihood or maximised complete log likelihood). Some methods have been introduced for choosing the number of classes in this context, for example by approximating these criteria.

We will focus here on approximations of the BIC, based either on mean field like approximations (Forbes and Peyrard, 2003) or on an approximation of the intractable likelihood by a distribution factorising on blocks (Stoehr et al., 2016). One can also see Stanford and Raftery (2002); Cucala and Marin (2013) and Stoehr et al. (2015).

Approximation of the BIC using a mean field-like approximation Forbes and Peyrard (2003) proposed to estimate the BIC of Schwarz et al. (1978) using the mean field like approximation.

Let us denote by $\tilde{z}^{1:L}$ and $\tilde{\theta}$ respectively the outputted configuration and estimator of θ (i.e. the approximation of the MLE) obtained from the mean-field like (more specifically simulated) EM algorithm. An approximation of the BIC under the mean field-like approximation based on these quantities is

$$\mathrm{BIC}^{\tilde{z}^{1:L}}(\tilde{\theta}) = 2 \log \mathbb{P}_{\tilde{z},\tilde{\theta}}(X^{1:L}) - d \log n$$

with d the number of free parameters in the model, and with the distribution of the observations under the mean-field like approximation and the approximation of the MLE

$$\mathbb{P}_{\tilde{z},\tilde{\theta}}(X^{1:L}) = \sum_{z^{1:L}} \mathbb{P}_{\tilde{z},\tilde{\theta}}(X^{1:L} \mid z^{1:L}) \mathbb{P}_{\tilde{z},\tilde{\theta}}(z^{1:L}) = \prod_{l=1}^{L} \sum_{z^l} \mathbb{P}_{\tilde{\theta}}(X^l \mid z^l) \mathbb{P}_{\tilde{\theta}}(z^l \mid z^{-l}).$$

The authors obtained unstable results for the choice of the number of classes on simulations for Potts model.

Other choices for $\tilde{z}^{1:L}$ and $\tilde{\theta}$ can be made. For example, Forbes and Peyrard (2003) mention that using for $\tilde{z}^{1:L}$ and $\tilde{\theta}$ the values obtained by the use of the Iterated Conditional Modes of Besag (1986) leads to the Pseudo-Likelihood Information Criterion (PLIC)[28] (Stanford, 1999; Stanford and Raftery, 2002).

Approximation of the BIC based on the partition functions Forbes and Peyrard (2003) also propose a selection criterion based on an approximation of the partition function (using the mean-field approximation). They express the BIC in terms of the partition functions of the conditional and marginal field (i.e. of the distribution of $Z^{1:L}$ and of the distribution of $Z^{1:L}$ given the observations $X^{1:L}$). Recall that $S(\psi)$ is the partition function for the Gibbs distribution of $Z^{1:L}$, and let us denote by $S(X^{1:L}, \theta)$ the partition function for the conditional field (i.e. distribution of $Z^{1:L}$ given $X^{1:L}$), i.e.

$$S(X^{1:L}, \theta) = \sum_{z^{1:L}} \exp\left[-H(z^{1:L} \mid X^{1:L}, \theta)\right]$$

[28] also denoted $\mathrm{BIC}_{\mathrm{PL}}$

where

$$H(z^{1:L} \mid X^{1:L}, \theta) = H(z^{1:L}, \psi) - \sum_{l=1}^{L} \log \mathbb{P}_\theta(X^l \mid z^l). \tag{1.7.10}$$

Then the likelihood can be written as follows in terms of the partition functions

$$\begin{aligned}\mathbb{P}_\theta(X^{1:L}) =& \frac{\mathbb{P}_\theta(X^{1:L} \mid Z^{1:L})\mathbb{P}_\theta(Z^{1:L})}{\mathbb{P}_\theta(Z^{1:L} \mid X^{1:L})} = \frac{\mathbb{P}_\theta(X^{1:L} \mid Z^{1:L}) \exp(-H(Z^{1:L}, \psi))}{\exp(-H(Z^{1:L} \mid X^{1:L}, \theta))} \frac{S(X^{1:L}, \theta)}{S(\psi)} \\ =& \frac{S(X^{1:L}, \theta)}{S(\psi)}\end{aligned}$$

using the expression in (1.7.10). The BIC can then be expressed as

$$\text{BIC} = 2 \log S(X^{1:L}, \hat{\theta}) - 2 \log S(\hat{\psi}) - d \log n,$$

where $\hat{\theta}$ and $\hat{\psi}$ are the maximum likelihood estimators of θ and ψ respectively. The partition functions could then be approximated using Monte Carlo techniques (see for example Potamianos and Goutsias (1997) for the approximation of the partition function, in a different context), but these methods can be slow. Forbes and Peyrard (2003) propose to use an approximation of the partition function based on the mean field approximation, and moreover to replace the unknown maximum likelihood estimators by the estimation $\tilde{\theta}$ of θ outputted by the simulated EM algorithm. Their approximation of the partition functions uses both facts that the mean field approximation is the minimiser of the Kullback-Leibler divergence from the true distribution, and the nonnegativity of this divergence. Indeed, the mean field approximation of the distribution of $Z^{1:L}$ can be written as (see (1.7.4) and (1.7.5))

$$\mathbb{P}_\psi^{\text{MF}}(Z^{1:L}) = S^{\text{MF}}(\psi)^{-1} \exp(-H^{\text{MF}}(Z^{1:L}, \psi))$$

with $S^{\text{MF}}(\psi)$ and $H^{\text{MF}}(Z^{1:L}, \psi)$ denoting respectively the partition function and energy function of the mean field approximation of the distribution of $Z^{1:L}$[29]. Then, the positivity of the Kullback-Leibler divergence (from the true Gibbs distribution of $Z^{1:L}$ to its mean field approximation) gives that

$$S(\psi) \geq S^{\text{MF}}(\psi) \exp(\mathbb{E}^{\text{MF}}[H(Z^{1:L}, \psi) - H^{\text{MF}}(Z^{1:L}, \psi)]). \tag{1.7.11}$$

[29]Note that the normalising constant $S^{\text{MF}}(\psi)$ can be computed thanks to the factorised form of the mean field approximation. Indeed, it is the product (over $l \in [\![1, L]\!]$) of the tractable normalising constants appearing in (1.7.5).

Note that this inequality still holds for any other approximation of the distribution of the form $\mathbb{P}_{\psi}^{\text{approx}}(Z^{1:L}) = S^{\text{approx}}(\psi)^{-1}\exp(-H^{\text{approx}}(Z^{1:L},\psi))$ instead of the mean field one. However, as the mean field approximation is optimal in the sense of the Kullback-Leibler minimisation among the distributions factorising over locations, the lower bound in (1.7.11) is optimal among such approximations. They then define an approximation of the BIC, using this lower bound as an approximation for the partition function, both for the marginal and conditional fields (defining $S^{\text{MF}}(X^{1:L},\theta)$ and $H^{\text{MF}}(Z^{1:L}\mid X^{1:L},\theta)$ similarly as for the marginal field). This leads to the following BIC^{GBF} criterion[30]

$$\begin{aligned}\text{BIC}^{\text{GBF}} =& 2\log S^{\text{MF}}(X^{1:L},\tilde{\theta}) - 2\mathbb{E}^{\text{MF}}\left[H(Z^{1:L}\mid X^{1:L},\tilde{\theta}) - H^{\text{MF}}(Z^{1:L}\mid X^{1:L},\tilde{\theta})\mid X^{1:L}\right]\\ &- 2\log S^{\text{MF}}(\tilde{\psi}) + 2\mathbb{E}^{\text{MF}}\left[H(Z^{1:L},\tilde{\psi}) - H^{\text{MF}}(Z^{1:L},\tilde{\psi})\right]\\ &- d\log n.\end{aligned}$$

The authors show that this criterion is more satisfactory than the previously introduced approximation of the criterion $\text{BIC}^{\tilde{z}^{1:L}}(\tilde{\theta})$, both from a theoretical and empirical point of view.

Block Likelihood Information Criterion (BLIC) Stoehr et al. (2016) introduced a selection criterion, the Block Likelihood Information Criterion (BLIC), approximating the BIC by replacing the intractable likelihood with a product distribution on independent blocks of the lattice, and using the method of Reeves and Pettitt (2004) to obtain an exact computation of the normalising constant for small enough blocks. The approximation $\text{BIC}^{\tilde{z}^{1:L}}(\tilde{\theta})$ of the BIC and the PLIC introduced above can be seen as particular cases of the BLIC with locations as blocks.

This criterion shows good results (compared to other criteria) for the choice of the number of classes on simulated data.

1.8 Space-evolving networks

We will focus in Chapter 3 on space-evolving networks, i.e. we assume that we observe a network at different locations and we want a global statistical analysis of these networks rather than considering them separately. This is motivated by the observation of ecological networks, i.e. the observation of the interactions between species in their environment. Edges in ecological networks can be of several types, such as predation,

[30] The bound in (1.7.11) is known as the Gibbs-Bogoliubov-Feynman (GBF) bound.

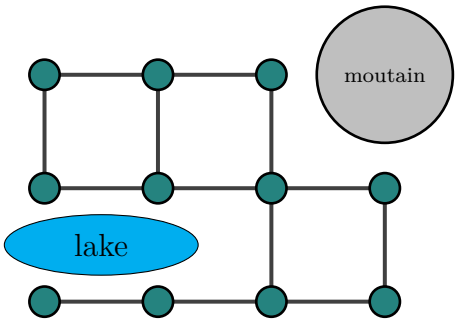


Fig. 1.14 Location graph for a species. This graph depends on the geography and environment. We assume here that the species is not present at high altitude (on the mountain) and cannot cross the lake.

parasitism, mutualism[31] or commensalism[32]. See for example Delmas et al. (2019) for more information on the analysis of ecological networks. Nonetheless, our model can be applied to different types of space-evolving networks, for example to the study of relations between different socio-economic classes at different geographical regions, or the collaboration of researchers from different domains in different universities.

In the model we consider, we model the correlation over space using a known graph over the locations (that we will call *location graph*). In such a graph, an edge exists when two locations are linked, meaning that the status of an individual (towards the interaction graph) at these two locations are correlated. For example, for the ecological application, this graph is based on the geography and environment, i.e. an edge exists between two locations if they are nearby and if there are no natural barrier between them that cannot be crossed by the species, such as mountains or rivers (see Figure 1.14). Note that we talk about space-evolving network, but this could apply to other types of dependencies between graphs, for example a dynamic graph including a seasonality effect, as in Figure 1.15. The field of application is then larger than space-evolving networks.

1.8.1 Contributions in the spatial SBM

In this work, we choose to consider a dependency based on Markov random fields for graphs observed at different locations. In short, we will assume that we observe the interaction graphs of n species at L locations. These n species are divided into Q (unobserved) classes, and for each species, its class membership will be distributed according to a Markov random field. At each location, we observe a binary interaction

[31] ecological interaction between two or more species where each species has a benefit, such as flowering plants being pollinated by animals

[32] interaction which is beneficial for a species and neutral for the other

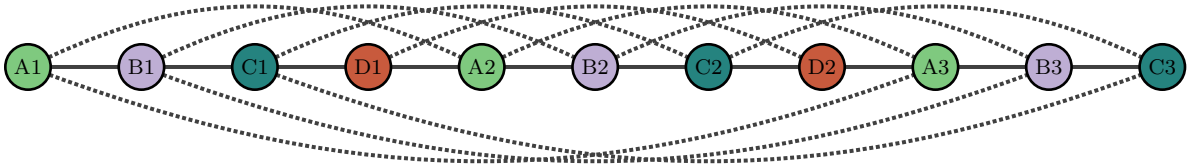


Fig. 1.15 A location graph modeling seasonality effects. Nodes A, B, C, D represent the four seasons, indexed by year of measurement (three years in total). The solid lines represent the dependency between two consecutive seasons, and the dotted lines represent the dependency between same seasons of different years.

graph between the species, which is assumed to follow a SBM so that, conditional on the latent classes, the connections between the species are independent Bernoulli random variables with parameter depending on the classes of the two considered species. See Figure 1.16 for a representation of our model. The model is described in details in Chapter 3. In this chapter, we propose an algorithm based on the simulated EM of Celeux et al. (2003) (see Section 1.7.7) to estimate the parameters of our model (which are the parameters of the MRF and the connectivity parameters of the SBM), and prove the generic identifiability of these parameters under certain conditions. This algorithm involves an additional approximation to solve intractability issues due to the dependencies in the conditional distribution of the latent variables given the observations induced by the SBM. We illustrate our results through synthetic datasets.

Let us recall that generic identifiability means that the nonidentifiable parameters form a set of Lebesgue measure zero. We do not specify the form of the subspace of non identifiable parameters in this work, and it is important to keep in mind that when we impose a constraint on the parameter reducing the parameter space to a subspace of smaller dimension[33], parameter identifiability is no longer guaranteed.

[33]for example by setting one component of the parameter to a fixed value or adding polynomial constraints on this parameter

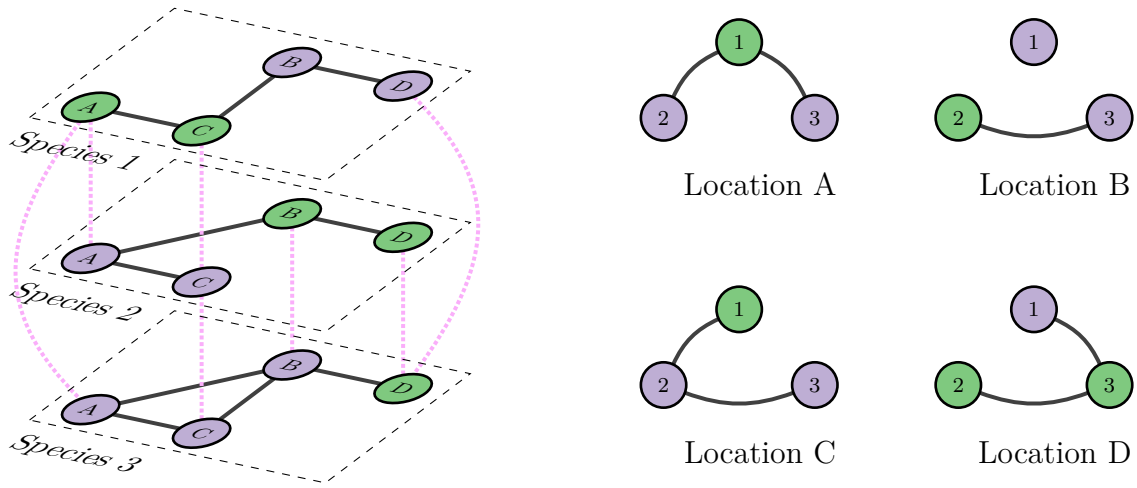


Fig. 1.16 Example of representation of the model for $Q = 2$, $L = 4$ and $n = 3$. On the left, the three layers represent the graphs on locations of the three species (i.e. the three location graphs), and for each location (A, B, C and D), the dotted edges between layers represent the connection between species at that location. On the right are the observed interaction graphs between the 3 species at each of the four locations. Note that the representation on the right only contains the interactions graphs at each location, and not the space dependency between these locations. The Q classes are represented by colors on the nodes (green and purple).

Chapter 2

Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model

2.1 Introduction

Random graphs are a suitable tool to model and describe interactions in many kinds of datasets such as biological, ecological, social or transport networks. Here we are interested in time-evolving networks, which is a powerful tool for modeling real-world phenomena, where the role or behaviour of the nodes in the network and the relationships between them are allowed to change over time. Indeed, it is important to take into account the evolutionary behaviour of the graphs, instead of just studying separate snapshots as static graphs. We focus on graphs evolving in discrete time and refer to Holme (2015) for an introduction to dynamic networks.

A myriad of dynamic graph models has been introduced in the past few years, see for instance Zhang et al. (2017a). We focus here on those which are based on the (static) stochastic block model (SBM, Holland et al., 1983) in which the nodes are partitioned into classes. In the SBM, class memberships of the nodes are represented by latent variables and the connection between two nodes is drawn from a distribution depending on the classes of these two nodes (a Bernoulli distribution in the case of binary graphs). A first dynamic version of the SBM with discrete time is proposed in Yang et al. (2011). There, the nodes are partitioned into Q classes and the graphs are binary or weighted. The nodes are allowed to change membership over time, and these changes are governed by independent Markov chains with values in the Q classes, while the connection

probabilities are constant over time. Xu and Hero (2014) introduce a state-space model on the logit of the connection probabilities for dynamic (binary) networks with connection probabilities and group memberships varying over time. Unfortunately, their model presents parameter identifiability issues (Matias and Miele, 2017). Xu (2015) proposes a stochastic block transition model in which the presence or absence of an edge between two nodes at a particular time affects the presence or absence of such an edge at a future time. There, the nodes can change classes over time, new nodes can enter the network, and the connection probabilities are allowed to vary over time. The model in Matias and Miele (2017) and in Becker and Holzmann (2018) is quite similar to that of Yang et al. (2011) except that it allows the connection probabilities to vary and the latter is moreover nonparametric. Bartolucci et al. (2018) extend the model of Yang et al. (2011) to deal with different forms of reciprocity in directed graphs, by directly modeling dyadic relations and with the assumption that the dyads are conditionally independent given the latent variables. Paul and Chen (2016) and Han et al. (2015) study multi-graph SBM, arising in settings including dynamic networks and multi-layer networks where each layer corresponds to a type of edge. In these two models, the nodes memberships stay constant over the layers. Pensky (2019); Pensky et al. (2019) study a dynamic SBM for undirected and binary edges where both connection probabilities and group memberships vary over time, assuming that the connection probabilities between groups are a smooth function of time. Xing et al. (2010) and Ho et al. (2011) introduce dynamic versions of the mixed-membership stochastic block model, allowing each actor to carry out different roles when interacting with different peers. Zreik et al. (2016) introduce the dynamic random subgraph model, given a known decomposition of the graph into subgraphs, in which the latent class membership depends on the subgraph membership and the edges are categorical variables, their types being sampled from a distribution depending on the latent classes of the two nodes. There, a state-space model is used to characterize the temporal evolution of the latent classes proportions.

As far as estimation is concerned, different methods of inference are proposed to estimate groups and model parameters. The maximum likelihood estimator (MLE) is not tractable in the SBM, thus neither in its dynamic versions. Variational methods are rather popular to approximate that MLE (Xing et al., 2010; Ho et al., 2011; Han et al., 2015; Paul and Chen, 2016; Zreik et al., 2016; Matias and Miele, 2017; Bartolucci et al., 2018). Yang et al. (2011) rely on Gibbs sampling and simulated annealing. Pensky et al. (2019) propose an estimator of the connection probabilities matrix at each time step by a discrete kernel-type method and obtain a clustering of the nodes thanks to spectral clustering on this estimated matrix. They also give an estimator for the number

of clusters. Spectral clustering algorithms are also used by Han et al. (2015) on the mean graph over time and by Liu et al. (2018) who use eigenvector smoothing to get some similarity across time periods (and allow the number of classes to be unknown and possibly varying over time).

Some theoretical results on the convergence of the procedures have been proven, mainly for static graphs. In the static SBM, Celisse et al. (2012) prove the consistency of the MLE and variational estimates as the number of nodes increases[1], and Bickel et al. (2013) establish their asymptotic normality. Mariadassou and Matias (2015) have a different approach and give sufficient conditions for the groups posterior distribution to converge to a Dirac mass located at the actual groups configuration, for every parameter in a neighborhood of the true one. Rohe et al. (2011) give asymptotic results on the normalized graph Laplacian and its eigenvectors for the spectral clustering algorithm, allowing the number of clusters to grow with the number of nodes. They also provide bounds on the number of misclustered nodes, requiring an assumption on the degree distribution. Lei and Rinaldo (2015) prove consistency for the recovery of communities in the spectral clustering on the adjacency matrix, with milder conditions on the degrees, and also extend this result to degree corrected stochastic block models. Klopp et al. (2017) derive oracle inequalities for the connection probabilities estimator and obtain minimax estimation rates, including the sparse case where the density of edges converges to zero as the number of nodes increase thus extending previous results of Gao et al. (2015). Gaucher and Klopp (2019) propose a bound on the risk of the maximum likelihood estimator of network connection probabilities, and show that it is minimax optimal in the sparse graphon model.

In the dynamic setting, fewer theoretical results have been established. Pensky (2019) derives a penalized least squares estimator of the connection probabilities adaptive to the number of blocks and which does not require knowledge of the number of classes Q. She shows that it satisfies an oracle inequality. Under the additional assumption that at most n_0 nodes change groups between two time steps, this estimator attains minimax lower bounds for the risk. She also introduces a dynamic graphon model and shows that the estimators (that do not require knowledge of a degree of smoothness of the graphon function) are minimax optimal within a logarithmic factor of the number of time steps. Based on the same dynamic SBM with at most n_0 nodes changing groups between two time steps, Pensky et al. (2019) give an upper bound for the (non asymptotic) error of their estimators of the connection probabilities matrix and group memberships (and also

[1] the consistency results for the estimators of the parameter of the latent groups distribution requiring an assumption on the rate of convergence of the connection parameter estimators

an estimator for the number of clusters). Han et al. (2015) show consistency (as the number of time steps increases but the number of nodes is fixed) of two estimators of the class memberships for dynamic SBM (and more generally multi-graph SBM) in which the nodes memberships are constant over time but the connection probabilities are allowed to vary and the considered graphs are binary and symmetric. They show that the spectral clustering (on the mean graph over time) estimator of the class memberships is consistent under some stationarity and ergodicity conditions on the connection probabilities. They also prove that the MLE of the class memberships is consistent (i.e. that the fraction of misclustered nodes converges to 0) in the general case (without any structure on the connection probabilities), provided certain sufficient conditions are satisfied. In their multi-layer model, Paul and Chen (2016) give minimax rates of misclassification under certain conditions on the growth of the types of relations, number of nodes and number of classes, extending the result of Han et al. (2015).

Here, we consider a dynamic version of the binary SBM as in Yang et al. (2011), where each node is allowed to change group membership at each time step according to a Markov chain, independently of other nodes. We prove the consistency of the connectivity parameter MLE and, under some additional conditions, of the transition matrix MLE, when the number of nodes and of time steps are increasing. We also give upper bounds on the rates of convergence of these estimators. While these upper bounds are known to be non optimal in the static case where asymptotic normality is obtained with classical parametric rates of convergence (Bickel et al., 2013), these are the first to be established in a dynamic setting for the MLE. As already mentioned, the log-likelihood is intractable (except for very small values of the number of nodes n and the number of time steps T), as it requires to sum over Q^{nT} terms. Thus, while its consistency remains an important result, the estimator cannot be computed. A possible alternative is to rely on a variational estimator to approximate the MLE (see for instance Matias and Miele, 2017). We also establish the consistency of the variational estimator of the connectivity parameter and under some additional assumptions, that of the variational estimator of the transition matrix and obtain the same upper bounds on the rates of convergence as for the MLE. In the particular case where the number of time steps T is fixed, we also consider the model of Matias and Miele (2017), in which the connection probabilities are allowed to vary over time and generalise these results with only the number of nodes increasing. When $T = 1$, we not only recover the results of Celisse et al. (2012) but extend these by giving rates of convergence. Unlike the model studied in Han et al. (2015) and Paul and Chen (2016), the node memberships in our model evolve over time. Our context is different from Pensky (2019) that focuses on least squares estimate.

This article is organized as follows. Section 2.2 introduces our model and notation. More precisely, Section 2.2.1 describes the dynamic stochastic block model as introduced in Yang et al. (2011), Section 2.2.2 gives the assumptions we make on the model parameters, Section 2.2.3 describes the dynamic stochastic block model as in Matias and Miele (2017) for the finite time case and Section 2.2.4 states the expression of the likelihood of this model to define the MLE. Section 2.3 establishes the consistency and upper bounds of the rates of convergence for the MLE of the connection probabilities in Section 2.3.1 and of the transition matrix in Section 2.3.2. Section 2.4 is dedicated to variational estimators: Section 2.4.1 and 2.4.2 establish the consistency of the variational estimators of the connection probabilities and transition matrix, respectively, along with upper bounds of the associated rates of convergence. All the proofs of the main results are postponed to Section 2.5, except those for the fixed T case that are in Appendix A.1, while the more technical proofs are deferred to Appendix A.2.

2.2 Model and notation

2.2.1 Dynamic stochastic block model

We consider a set of n vertices, forming a sequence of binary undirected graphs with no self-loops at each time $t = 1, \dots, T$. The case of a set of directed graphs, with or without self-loops, may be handled similarly. These vertices are assumed to be split into Q latent classes, and we denote by Z_i^t the label of the i-th vertex at time t. Letting $Z_i = (Z_i^1, \dots, Z_i^T)$, we assume that the $\{Z_i\}_{1 \leq i \leq n}$ are independent and identically distributed (iid) and each Z_i is a homogeneous and stationary Markov chain with transition probabilities

$$\mathbb{P}(Z_i^{t+1} = l \mid Z_i^t = q) = \gamma_{ql}, \quad \forall 1 \leq q, l \leq Q$$

where $\Gamma = (\gamma_{ql})_{1 \leq q,l \leq Q}$ is a stochastic matrix, i.e. with nonnegative coefficients and with each row summing to one. We let $\alpha = (\alpha_1, \dots, \alpha_Q)$ the stationary distribution of the Markov chain. For any $i \in [\![1, n]\!]$, the probability distribution of Z_i is then

$$\mathbb{P}_\theta(Z_i) = \alpha_{Z_i^1} \prod_{t=1}^{T-1} \gamma_{Z_i^t Z_i^{t+1}}.$$

We will also denote $Z^t = (Z_1^t, \dots, Z_n^t)$ and $Z^{1:T} = (Z^1, \dots, Z^T) = (Z_i^t)_{1 \leq t \leq T, 1 \leq i \leq n}$.

Consider $X^t = \{X_{ij}^t\}_{1\leq i,j\leq n}$ the symmetric binary adjacency matrix of the graph at time t such that for every nodes $1 \leq i, j \leq n$, we have $X_{ii}^t = 0$ and $X_{ij}^t = X_{ji}^t$. Each X^t follows a stochastic block model so that, conditional on the latent groups $\{Z_i^t\}_{1\leq i\leq n}$, the $\{X_{ij}^t\}_{1\leq i,j\leq n}$ are independent Bernoulli random variables

$$X_{ij}^t \mid Z_i^t = q, Z_j^t = l \sim \mathcal{B}(\pi_{ql})$$

where $(\pi_{ql})_{1\leq q,l\leq Q} \in [0,1]^{Q^2}$ are the connectivity parameters. More precisely, conditional on the whole sequence of latent groups $\{Z_i^t\}_{1\leq t\leq T, 1\leq i\leq n}$, the graphs $X^{1:T} = X^1, \ldots, X^T$ are assumed to be independent, each X^t having a distribution depending only on $\{Z_i^t\}_{1\leq i\leq n}$. The model is thus parameterized by $\theta = (\Gamma, \pi)$, with $\Gamma = (\gamma_{ql})_{1\leq q,l\leq Q}$ and $\pi = (\pi_{ql})_{1\leq q,l\leq Q}$. Note that π is a symmetric matrix in the undirected setup. We denote by $\mathbb{P}_\theta$ (resp. $\mathbb{E}_\theta$) the probability distribution (resp. expectation) of all the random variables $\{Z_i^t, X_{ij}^t\}_{t\geq 1; i,j\geq 1}$, under the parameter value θ. In the following, we assume that we observe $\{X_{ij}^t\}_{1\leq i,j,\leq n,\, 1\leq t\leq T}$ and we denote by $\theta^* = (\Gamma^*, \pi^*) = ((\gamma_{ql}^*)_{1\leq q,l\leq Q}, (\pi_{ql}^*)_{1\leq q,l\leq Q})$ the true parameter value, with corresponding probability distribution $\mathbb{P}_{\theta^*}$ and expectation $\mathbb{E}_{\theta^*}$, and by $\alpha^* = (\alpha_q^*)_{1\leq q\leq Q}$ the (true) stationary distribution corresponding to the transition matrix Γ^*. We also let $\mathbb{1}_A$ denote the indicator function of the set A and A^c the complementary set of A in the ambient set. For any integer $M \geq 1$, the set $[\![1, M]\!]$ is the set of integers between 1 and M. For any finite set A, let $|A|$ denote its cardinality. For any configuration $z^{1:T}$, we denote $N_q(z^t)$ (resp. $N_{ql}(z^{1:T})$) the number of nodes assigned to class q by the configuration z^t (resp. the number of transitions from class q to class l in configuration $z^{1:T}$), that is

$$N_q(z^t) = |\{i \in [\![1,n]\!]; z_i^t = q\}| \quad \text{and} \quad N_{ql}(z^{1:T}) = \sum_{t=1}^{T-1}\sum_{i=1}^{n} \mathbb{1}_{z_i^t=q, z_i^{t+1}=l}. \tag{2.2.1}$$

We also define for any two parameters $\theta = (\Gamma, \pi)$ and $\theta' = (\Gamma', \pi')$ the following distances

$$\|\pi - \pi'\|_\infty = \max_{1\leq q,l\leq Q} |\pi_{ql} - \pi'_{ql}| \quad \text{and} \quad \|\Gamma - \Gamma'\|_\infty = \max_{1\leq q,l\leq Q} |\gamma_{ql} - \gamma'_{ql}|.$$

2.2.2 Assumptions

The assumptions we make on the model parameters are the following.

1. For every $1 \leq q \neq q' \leq Q$, there exists some $l \in [\![1, Q]\!]$ such that $\pi_{ql} \neq \pi_{q'l}$.
2. There exists some $0 < \delta < 1/Q$ such that for any $(q, l) \in [\![1, Q]\!]^2$, we have $\gamma_{ql} \in [\delta, 1 - \delta]$.

3. There exists some $\zeta > 0$ such that for any $(q, l) \in [\![1, Q]\!]^2$, we have $\pi_{ql} \in [\zeta, 1 - \zeta]$.

Assumption 1 is necessary for identifiability of the model. Indeed, if it does not hold, we cannot distinguish between classes q and q'. Assumption 2 ensures that each Markov chain Z_i is irreducible, aperiodic and recurrent. This assumption could be weakened at the cost of technicalities. In particular, it implies that the stationary distribution α exists. Moreover, Assumption 2 also implies that for any $q \in [\![1, Q]\!]$, we have $\alpha_q \in [\delta, 1 - \delta]$. Note that this can be seen as an equivalent of Assumption 2 in Celisse et al. (2012) (on the probability distribution of the class memberships) in the dynamic case. Celisse et al. (2012) however also have an additional assumption that is an empirical version of this assumption (which states that the observed class proportions are bounded away from 0) that is true with high probability. We do not make such an assumption and use the fact that the probability of this event converges to 1. Assumption 3 is technical and could also be weakened with additional technicalities. For example, Celisse et al. (2012) also consider the case $\pi_{ql} \in \{0, 1\}$ (i.e. $\pi_{ql} \in \{0, 1\} \cup [\zeta, 1 - \zeta]$) whereas we do not. The whole parameter set defined by these constraints is denoted by Θ. In the following, we assume that $\theta^* \in \Theta$.

In what follows, we work up to label permutation on the groups. Indeed, as in any latent group model, the parameters can only be recovered up to label switching on the latent groups. We then define the following notation for any permutation $\sigma \in \mathfrak{S}_Q$ with $\mathfrak{S}_Q$ the set of permutations on $[\![1, Q]\!]$

$$\theta_\sigma = (\Gamma_\sigma, \pi_\sigma) = \left((\gamma_{\sigma(q)\sigma(l)})_{1 \leq q,l \leq Q}, (\pi_{\sigma(q)\sigma(l)})_{1 \leq q,l \leq Q} \right).$$

2.2.3 Finite time case

If the number of time steps T is fixed, it is possible to let the connection probabilities vary over time. We then consider this case, the connection parameter now being $\pi^{1:T} = (\pi^1, \dots, \pi^T)$ with $\pi^t = (\pi^t_{ql})_{1 \leq q,l \leq Q}$ for every $t \in [\![1, T]\!]$ and $\pi^t_{ql} = \mathbb{P}_\theta(X^t_{ij} = 1 \,|\, Z^t_i = q, Z^t_j = l)$ for any $(t, q, l) \in [\![1, T]\!] \times [\![1, Q]\!]^2$. Note that this is the more general model of Matias and Miele (2017), in which the model parameter is $\theta = (\Gamma, \pi^{1:T})$. Moreover, we introduce the following Assumptions 1' and 3' that are alternate versions of Assumptions 1 and 3 respectively for the finite time case.

1'. For every $t \in [\![1, T]\!]$, for every $1 \leq q \neq q' \leq Q$, there exists some $l \in [\![1, Q]\!]$ such that $\pi^t_{ql} \neq \pi^t_{q'l}$.

3'. There exists some $\zeta > 0$ such that for every $t \in [\![1, T]\!]$, for any $(q, l) \in [\![1, Q]\!]^2$, we have $\pi^t_{ql} \in [\zeta, 1 - \zeta]$.

Assumption 1' (resp. Assumption 3') expresses that for every $t \in [\![1, T]\!]$, π^t satisfies Assumption 1 (resp. Assumption 3). We also introduce the following additional assumption, which ensures (together with Assumption 1') that the model is identifiable (up to a label permutation). See Matias and Miele (2017).

4. For every $q \in [\![1, Q]\!]$, for every $t_1, t_2 \in [\![1, T]\!]$, $\pi_{qq}^{t_1} = \pi_{qq}^{t_2} \coloneqq \pi_{qq}$ and $\{\pi_{qq}; q \in [\![1, Q]\!]\}$ are Q distinct values.

Assumption 4 states that the diagonal of π does not change over time, and that its values are distinct. We denote by Θ^T the set of parameters satisfying Assumptions 1', 2, 3' and 4. As before, we assume in the following that $\theta^* \in \Theta^T$ in the fixed T case. We also define as before for any $\pi^{1:T}$ and $\pi'^{1:T}$ the distance

$$\|\pi^{1:T} - \pi'^{1:T}\|_\infty = \max_{(q,l,t)\in[\![1,Q]\!]^2\times[\![1,T]\!]} |\pi_{ql}^t - \pi_{ql}'^t|.$$

2.2.4 Likelihood

The conditional log-likelihood and the log-likelihood write

$$\begin{aligned}
\ell_c(\theta; Z^{1:T}) &= \log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T}) = \sum_{t=1}^{T} \log \mathbb{P}_\theta(X^t \mid Z^t) \\
&= \sum_{t=1}^{T} \sum_{1\le i<j\le n} X_{ij}^t \log \pi_{Z_i^t Z_j^t} + (1 - X_{ij}^t) \log(1 - \pi_{Z_i^t Z_j^t}) \\
\text{and} \quad \ell(\theta) &= \log \mathbb{P}_\theta(X^{1:T}) = \log \left(\sum_{z^{1:T}\in[\![1,Q]\!]^{nT}} e^{\ell_c(\theta; z^{1:T})} \mathbb{P}_\theta(Z^{1:T} = z^{1:T}) \right), \qquad (2.2.2)
\end{aligned}$$

respectively. We then denote the maximum likelihood estimator (MLE) by

$$\hat{\theta} = (\hat{\Gamma}, \hat{\pi}) = \arg\max_{\theta\in\Theta} \ell(\theta).$$

In the next section, we study separately the consistency of the connectivity parameter estimator $\hat{\pi}$ and that of the transition matrix estimator $\hat{\Gamma}$.

2.3 Consistency of the maximum likelihood estimator

2.3.1 Connectivity parameter

We first prove the consistency of the maximum likelihood estimator of the connectivity parameter $\pi = (\pi_{ql})_{1\leq q,l\leq Q}$ when the number of nodes and time steps increase. We denote the normalized log-likelihood by

$$M_{n,T}(\Gamma,\pi) = \frac{2}{n(n-1)T}\ell(\theta) = \frac{2}{n(n-1)T}\log \mathbb{P}_\theta(X^{1:T})$$

and introduce the quantities, for any $A = (a_{ql})_{1\leq q,l\leq Q} \in \mathcal{A}$ the set of $Q\times Q$ stochastic matrices,

$$\begin{aligned}\mathbb{M}(\pi,A) &= \sum_{1\leq q,l\leq Q}\alpha_q^*\alpha_l^* \sum_{1\leq q',l'\leq Q} a_{qq'}a_{ll'}[\pi^*_{ql}\log\pi_{q'l'} + (1-\pi^*_{ql})\log(1-\pi_{q'l'})]\\ \text{and}\quad \mathbb{M}(\pi) &= \sup_{A\in\mathcal{A}}\mathbb{M}(\pi,A) = \mathbb{M}(\pi,\bar{A}_\pi),\end{aligned} \tag{2.3.1}$$

where $\bar{A}_\pi = \arg\max_{A\in\mathcal{A}}\mathbb{M}(\pi,A)$. It is worth noticing that $\mathbb{M}(\pi)$, which will be the limiting value for $M_{n,T}(\Gamma,\pi)$ when n and T increase (see below), does not depend on Γ.

Theorem 2.3.1. *For any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity, if $\log(T) = o(n)$, we have for all $\epsilon > 0$*

$$\mathbb{P}_{\theta^*}\left(\sup_{(\Gamma,\pi)\in\Theta}|M_{n,T}(\Gamma,\pi) - \mathbb{M}(\pi)| > \frac{\epsilon r_{n,T}}{\sqrt{n}}\right) \xrightarrow[n,T\to+\infty]{} 0.$$

We then conclude on the consistency of the maximum likelihood estimator of the connection probabilities with the following corollary. Note that we also obtain an upper bound of the rate of convergence of this estimator.

Corollary 2.3.1. *For any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity such that $r_{n,T} = o(n^{1/4})$ and if $\log(T) = o(n)$, we have for every $\epsilon > 0$*

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\pi^* - \hat{\pi}_\sigma\|_\infty > \frac{\epsilon r_{n,T}}{n^{1/4}}\right) \xrightarrow[n,T\to\infty]{} 0.$$

We want to get equivalent consistency results if the number of time steps T is fixed and only the number of nodes n increases. In that case, denoting by $\hat{\theta} = (\hat{\Gamma}, \hat{\pi}^{1:T})$ the MLE of θ, we have the following Corollary that is the equivalent of Corollary 2.3.1.

Corollary 2.3.2. *If the number of time steps T is fixed, we have for every $\epsilon > 0$ and for any sequence $\{r_n\}_{n\geq 1}$ increasing to infinity such that $r_n = o(n^{1/4})$*

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q} \|\pi^{*1:T} - \hat{\pi}_\sigma^{1:T}\|_\infty > \frac{\epsilon r_n}{n^{1/4}}\right) \xrightarrow[n\to\infty]{} 0,$$

denoting $\hat{\pi}_\sigma^{1:T} = (\hat{\pi}_\sigma^t)_{t\in[\![1,T]\!]}$.

This result states that $\min_{\sigma\in\mathfrak{S}_Q} \|\pi^{*1:T} - \hat{\pi}_\sigma^{1:T}\|_\infty$ converges to 0 in $\mathbb{P}_{\theta^*}$-probability as n increases, i.e. the MLE of the connection probabilities is consistent up to label switching, and gives an upper bound of the rate of convergence of the MLE of the connection probabilities. The particular case when $T = 1$ is then a stronger result than that of Celisse et al. (2012) where no rate of convergence is given.

Remark 2.3.1. Note that in Corollaries 2.3.1 and 2.3.2, the results still hold for any sequences $r_{n,T}$ and r_n increasing to infinity, respectively. However, we are interested in sequences increasing slowly to infinity, giving the strongest results, namely the smallest lower bounds. Indeed, whenever these assumptions are not satisfied, the lower bounds appearing in the inequalities are larger, and the results may even become trivial.

2.3.2 Latent transition matrix

We now prove that the MLE for the transition matrix Γ is consistent when the number of nodes and time steps increase.

Lemma 2.3.1. *Any critical point $\breve{\theta} = (\breve{\Gamma}, \breve{\pi})$ of the likelihood function $\ell(\cdot)$ is such that $\breve{\Gamma}$ satisfies the fixed point equation*

$$\forall (q,l) \in [\![1,Q]\!]^2, \quad \breve{\gamma}_{ql} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}\left(Z_i^t = q, Z_i^{t+1} = l \mid X^{1:T}\right)}{\sum_{t=1}^{T-1}\sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}\left(Z_i^t = q \mid X^{1:T}\right)}. \tag{2.3.2}$$

There are two different possible cases for the MLE $\hat{\theta}$

- Either $\hat{\theta}$ is a critical point of the likelihood function. Then $\hat{\Gamma}$ satisfies equation (2.3.2).
- Or $\hat{\theta}$ is not a critical point (this can happen if it belongs to the boundary of Θ) and we assume that there exists $\breve{\Gamma}$ such that $(\breve{\Gamma}, \hat{\pi}) \in \Theta$ and $(\breve{\Gamma}, \hat{\pi})$ satisfies equation (2.3.2) (at least for n and T large enough). We then choose as our estimator $(\breve{\Gamma}, \hat{\pi})$. By an abuse of notation, we will denote this estimator $\hat{\theta} = (\hat{\Gamma}, \hat{\pi})$ and call it MLE in the following.

In what follows, for any fixed configuration $z^{1:T}$, any $\theta \in \Theta$ and any $\epsilon > 0$, we consider the event

$$\mathcal{E}(z^{1:T}, \theta, \epsilon) := \left\{ \frac{\mathbb{P}_\theta(Z^{1:T} \neq z^{1:T} \mid X^{1:T})}{\mathbb{P}_\theta(Z^{1:T} = z^{1:T} \mid X^{1:T})} > \epsilon \right\}.$$

The following result establishes that asymptotically, any estimator that correctly estimates the transition probability matrix π also recovers the group memberships. This result is similar to Theorem 1 in Mariadassou and Matias (2015).

Theorem 2.3.2. *For any estimator $\breve{\theta} \in \Theta$ (at least for n and T large enough), if $\log(T) = o(n)$, there exist some positive constants C, C_1, C_2, C_3, C_4 such that for any $\epsilon > 0$, for any positive sequence $\{y_{n,T}\}_{n,T\geq 1}$ such that $\log(1/y_{n,T}) = o(n)$, any $\eta \in (0, \delta)$ and for n and T large enough, we have*

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(\mathcal{E}(Z^{1:T}, \breve{\theta}, \epsilon y_{n,T})\right) \leq & QT\exp(-2\eta^2 n) + \mathbb{P}_{\theta^*}\left(\|\breve{\pi} - \pi^*\|_\infty > v_{n,T}\right) \\ & + CnT\Bigg\{\exp\left[-(\delta-\eta)^2 C_1 n + C_2\log(nT) - C_4\log(\epsilon y_{n,T})\right] \\ & + \exp\left[-C_3\frac{(\log(nT))^2}{nv_{n,T}^2} + 3n\log(nT)\right]\Bigg\},\end{aligned}$$

whenever $\{v_{n,T}\}_{n,T\geq 1}$ is a sequence decreasing to 0 such that $v_{n,T} = o(\sqrt{\log(nT)}/n)$.

Theorem 2.3.3. *If $\log(T) = o(n)$, for any $\epsilon > 0$ and $\{r_{n,T}\}_{n,T\geq 1}$ any sequence increasing to infinity such that $r_{n,T} = o\left(\sqrt{nT/\log n}\right)$, we have for any $\sigma \in \mathfrak{S}_Q$*

$$\mathbb{P}_{\theta^*}\left(\|\hat{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq Q^2(3Q+1)\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1)$$

with $\{v_{n,T}\}_{n,T\geq 1}$ a sequence decreasing to 0 such that $v_{n,T} = o(\sqrt{\log(nT)}/n)$.

Corollary 2.3.3. *Assume that $\log(T) = o(n)$ and $\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma - \pi^*\|_\infty = o_{\mathbb{P}_{\theta^*}}(v_{n,T})$ with $\{v_{n,T}\}_{n,T\geq 1}$ a sequence decreasing to 0 such that $v_{n,T} = o(\sqrt{\log(nT)}/n)$. Then for any $\epsilon > 0$ and $\{r_{n,T}\}_{n,T\geq 1}$ any sequence increasing to infinity such that $r_{n,T} = o\left(\sqrt{nT/\log n}\right)$, we have the convergence*

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \xrightarrow[n,T\to\infty]{} 0.$$

Remark 2.3.2. Note that the upper bound obtained in Corollary 2.3.1 on the rate of convergence in probability of $\hat{\pi}$ does not ensure that $\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma - \pi^*\|_\infty = o_{\mathbb{P}_{\theta^*}}(v_{n,T})$

holds. While the latter has never been established (to our knowledge), it is a reasonable assumption[2].

We want an equivalent result than that of Corollary 2.3.3 when the number of time steps T is fixed, and the connection probabilities are varying over time (the connection parameter being $\pi = \pi^{1:T} = (\pi^1, \dots, \pi^T)$ with $\pi^t = (\pi^t_{ql})_{q,l}$). For that, we are going to need an equivalent of Theorem 2.3.2 in that case.

Theorem 2.3.4. *For any fixed $T \geq 2$, for any estimator $\breve{\theta} \in \Theta^T$ (at least for n large enough), there exist some positive constants C, C_1, C_2, C_3, C_4 such that for any $\epsilon > 0$, for any positive sequence $\{y_n\}_{n\geq 1}$ such that $\log(1/y_n) = o(n)$, any $\eta \in (0, \delta)$ and for n large enough, we have*

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(\mathcal{E}(Z^{1:T}, \breve{\theta}, \epsilon y_n)\right) \leq & QT\exp(-2\eta^2 n) + \mathbb{P}_{\theta^*}\left(\|\breve{\pi}^{1:T} - \pi^{*1:T}\|_\infty > v_n\right) \\ & + CnT\Bigg\{\exp\left[-(\delta-\eta)^2 C_1 n + C_2\log(nT) - C_4\log(\epsilon y_n)\right] \\ & \quad + \exp\left[-C_3\frac{(\log(nT))^2}{nv_n^2} + 5n\log(nT)\right]\Bigg\},\end{aligned}$$

whenever $\{v_n\}_{n\geq 1}$ is a sequence decreasing to 0 such that $v_n = o(\sqrt{\log(n)}/n)$.

The following corollary gives the expected result.

Corollary 2.3.4. *Let the number of time steps $T \geq 2$ be fixed. We assume that $\min_{\sigma\in\mathfrak{S}_Q} \|\hat{\pi}^{1:T}_\sigma - \pi^{*1:T}\|_\infty = o_{\mathbb{P}_{\theta^*}}(v_n)$ with $\{v_n\}_{n\geq 1}$ a sequence decreasing to 0 such that $v_n = o(\sqrt{\log(n)}/n)$. Then for any $\epsilon > 0$ and $\{r_n\}_{n\geq 1}$ any sequence increasing to infinity such that $r_n = o\left(\sqrt{n/\log n}\right)$, we have the convergence*

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q} \|\hat{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_n\frac{\sqrt{\log n}}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 0.$$

The proof of Corollary 2.3.4 is the same as that of Corollary 2.3.3, but relying on Theorem 2.3.4 instead of Theorem 2.3.2 and is therefore omitted.

Remark 2.3.3. As in Remark 2.3.1 for Corollaries 2.3.1 and 2.3.2, the results of Corollaries 2.3.3 and 2.3.4 still hold for sequences $r_{n,T}$ and r_n increasing to infinity at any rate.

[2]In particular, in the static case, Bickel et al. (2013) obtained a rate of n^{-1} for the connectivity parameter in a non sparse setup (like ours)

2.4 Variational estimators

In practice, we cannot compute the MLE except for very small values of n and T, because it involves a summation over all the Q^{nT} possible latent configurations. We cannot either use the Expectation-Maximization (EM) algorithm to approximate it because it involves the computation of the conditional distribution of the latent variables given the observations which is not tractable. A common solution is to use the Variational Expectation-Maximization (VEM) algorithm that optimizes a lower bound of the log-likelihood (see for example Daudin et al. (2008)). Let us denote $Z_{iq}^t = \mathbb{1}_{Z_i^t=q}$ for every t, i and q. Using the same approach as in Matias and Miele (2017) for the VEM algorithm in the dynamic SBM, we consider a variational approximation of the conditional distribution of the latent variable $Z^{1:T}$ given the observed variable $X^{1:T}$ in the class of probability distributions parameterized by $\chi = (\tau, \eta) = \left(\{\tau_{iq}^t\}_{t,i,q}, \{\eta_{iql}^t\}_{t,i,q,l}\right)$ of the form

$$\mathbb{Q}_\chi(Z^{1:T}) = \prod_{i=1}^n \mathbb{Q}_\chi(Z_i^1) \prod_{t=2}^T \mathbb{Q}_\chi(Z_i^t \mid Z_i^{t-1}) = \prod_{i=1}^n \left\{ \left[\prod_{q=1}^Q (\tau_{iq}^1)^{Z_{iq}^1} \right] \prod_{t=1}^{T-1} \prod_{1\leq q,l\leq Q} \left(\frac{\eta_{iql}^t}{\tau_{iq}^t} \right)^{Z_{iq}^t Z_{il}^{t+1}} \right\},$$

i.e. with $\mathbb{Q}_\chi$ such that $\mathbb{E}_{\mathbb{Q}_\chi}\left[Z_{iq}^t Z_{il}^{t+1}\right] = \eta_{iql}^t$ and $\mathbb{E}_{\mathbb{Q}_\chi}\left[Z_{iq}^t\right] = \tau_{iq}^t$. Notice that $\mathbb{Q}_\chi(Z_i^{t+1} = l \mid Z_i^t = q) = \eta_{iql}^t/\tau_{iq}^t = \eta_{iql}^t/\sum_{q'=1}^Q \eta_{iqq'}^t$. The quantity to optimize in the VEM algorithm is then

$$\mathcal{J}(\chi, \theta) = \ell(\theta) - \mathrm{KL}(\mathbb{Q}_\chi, \mathbb{P}_\theta(\cdot|X^{1:T})) = \mathbb{E}_{\mathbb{Q}_\chi}\left[\log \mathbb{P}_\theta(X^{1:T}, Z^{1:T})\right] + \mathcal{H}(\mathbb{Q}_\chi)$$

with $\mathrm{KL}(\cdot,\cdot)$ denoting the Kullback-Leibler divergence and $\mathcal{H}(\cdot)$ denoting the entropy. Define

$$\hat{\chi}(\theta) = (\hat{\tau}(\theta), \hat{\eta}(\theta)) = \underset{\chi\in[0,1]^{T^2n^2Q^3}}{\arg\max}\ \mathcal{J}(\chi, \theta),$$

and the variational estimator of θ

$$\tilde{\theta} = (\tilde{\Gamma}, \tilde{\pi}) = \underset{\theta\in\Theta}{\arg\max}\, \mathcal{J}(\hat{\chi}(\theta), \theta).$$

Moreover, we denote $\tilde{\chi} = (\tilde{\tau}, \tilde{\eta}) = \hat{\chi}(\tilde{\theta}) = (\hat{\tau}(\tilde{\theta}), \hat{\eta}(\tilde{\theta}))$. In practice, the VEM algorithm is an iterative algorithm that maximizes the function $\mathcal{J}$ alternatively with respect to χ and θ in order to find $\tilde{\theta}$.

2.4.1 Connectivity parameter

Theorem 2.4.1. *For any sequence* $\{r_{n,T}\}_{n,T\geq 1}$ *increasing to infinity, if* $\log(T) = o(n)$, *we have for all* $\epsilon > 0$

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\mathcal{J}(\hat{\chi}(\theta),\theta) - \mathbb{M}(\pi)\right| > \frac{\epsilon r_{n,T}}{\sqrt{n}}\right) \underset{n,T\to+\infty}{\longrightarrow} 0.$$

We conclude on the consistency of the connection probabilities variational estimators as n and T increase thanks to the following corollary.

Corollary 2.4.1. *For any sequence* $\{r_{n,T}\}_{n,T\geq 1}$ *increasing to infinity such that* $r_{n,T} = o(n^{1/4})$, *we have for any* $\epsilon > 0$

$$\frac{1}{2}\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\tilde{\pi}_\sigma - \pi^*\|_\infty > \frac{\epsilon r_{n,T}}{n^{1/4}}\right) \xrightarrow[n,T\to\infty]{} 0.$$

We have the equivalent following corollary for a fixed number of time steps.

Corollary 2.4.2. *If the number of time steps* T *is fixed, we have for every* $\epsilon > 0$ *and for any sequence* $\{r_n\}_{n\geq 1}$ *increasing to infinity such that* $r_n = o(n^{1/4})$

$$\frac{1}{2}\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\tilde{\pi}_\sigma^{1:T} - \pi^{*1:T}\|_\infty > \frac{\epsilon r_n}{n^{1/4}}\right) \xrightarrow[n\to\infty]{} 0.$$

Remark 2.4.1. As for Corollaries 2.3.1 to 2.3.4, the results of Corollaries 2.4.1 and 2.4.2 still hold for any sequences $r_{n,T}$ and r_n increasing to infinity.

2.4.2 Latent transition matrix

We now prove that $\tilde{\Gamma}$ is consistent when the number of nodes and time steps increase.

Lemma 2.4.1. *Any critical point* $(\breve{\chi}, \breve{\theta})$ *of the function* $\mathcal{J}(\cdot,\cdot)$ *is such that* $\breve{\Gamma}$ *satisfies the fixed-point equation*

$$\forall (q,l) \in [\![1,Q]\!]^2, \quad \breve{\gamma}_{ql} = \frac{\sum_{i=1}^n \sum_{t=1}^{T-1} \breve{\eta}_{iql}^t}{\sum_{i=1}^n \sum_{t=1}^{T-1} \breve{\tau}_{iq}^t}. \tag{2.4.1}$$

We assume that $(\tilde{\chi}, \tilde{\theta})$ is a critical point of $\mathcal{J}(\cdot,\cdot)$. Then we have the fixed-point equation

$$\forall (q,l) \in [\![1,Q]\!]^2, \quad \tilde{\gamma}_{ql} = \frac{\sum_{i=1}^n \sum_{t=1}^{T-1} \hat{\eta}_{iql}^t(\tilde{\theta})}{\sum_{i=1}^n \sum_{t=1}^{T-1} \hat{\tau}_{iq}^t(\tilde{\theta})}. \tag{2.4.2}$$

The following theorem gives the consistency and a rate of convergence of this estimator, under an assumption on the rate of convergence of $\tilde{\pi}$.

Theorem 2.4.2. *If* $\log(T) = o(n)$*, for any* $\epsilon > 0$ *and* $\{r_{n,T}\}_{n,T\geq 1}$ *any sequence increasing to infinity such that* $r_{n,T} = o\left(\sqrt{nT/\log n}\right)$ *and for any* $\sigma \in \mathfrak{S}_Q$

$$\mathbb{P}_{\theta^*}\left(\|\tilde{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq 2Q^2(3Q+1)\mathbb{P}_{\theta^*}\left(\|\tilde{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1)$$

with $\{v_{n,T}\}_{n,T\geq 1}$ *a sequence decreasing to* 0 *such that* $v_{n,T} = o(\sqrt{\log(nT)}/n)$*.*

Corollary 2.4.3. *Assume that* $\log(T) = o(n)$ *and* $\min_{\sigma\in\mathfrak{S}_Q}\|\tilde{\pi}_\sigma - \pi^*\|_\infty = o_{\mathbb{P}_{\theta^*}}(v_{n,T})$ *with* $\{v_{n,T}\}_{n,T\geq 1}$ *a sequence decreasing to* 0 *such that* $v_{n,T} = o(\sqrt{\log(nT)}/n)$*. Then for any* $\epsilon > 0$ *and* $\{r_{n,T}\}_{n,T\geq 1}$ *any sequence increasing to infinity such that* $r_{n,T} = o\left(\sqrt{nT/\log n}\right)$*, we have the convergence*

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\tilde{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \xrightarrow[n,T\to\infty]{} 0.$$

The proof of Corollary 2.4.3 is the same as that of Corollary 2.3.3, using Theorem 2.4.2 instead of Theorem 2.3.3 and is therefore omitted.

When the number of time steps T is fixed and the connection probabilities can vary over time, we have the following Corollary that is the equivalent of Corollary 2.4.3.

Corollary 2.4.4. *Let the number of time steps* $T \geq 2$ *be fixed. We assume that* $\min_{\sigma\in\mathfrak{S}_Q}\|\tilde{\pi}_\sigma^{1:T} - \pi^{*1:T}\|_\infty = o_{\mathbb{P}_{\theta^*}}(v_n)$ *with* $\{v_n\}_{n\geq 1}$ *a sequence decreasing to* 0 *such that* $v_n = o(\sqrt{\log(n)}/n)$*. Then for any* $\epsilon > 0$ *and* $\{r_n\}_{n\geq 1}$ *any sequence increasing to infinity such that* $r_n = o\left(\sqrt{n/\log n}\right)$*, we have the convergence*

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\tilde{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_n\frac{\sqrt{\log n}}{\sqrt{n}}\right) \xrightarrow[n\to\infty]{} 0.$$

The proof of Corollary 2.4.4 is the same as that of Corollary 2.4.3, but relying on Theorem 2.3.4 instead of Theorem 2.3.2 and is therefore omitted.

Remark 2.4.2. As for Corollaries 2.3.1 to 2.4.2, the results of Corollaries 2.4.3 and 2.4.4 still hold for any sequences $r_{n,T}$ and r_n increasing to infinity.

2.5 Proofs of main results

2.5.1 Proof of Theorem 2.3.1

The proof follows the lines of the proof of Theorem 3.6 in Celisse et al. (2012). Nonetheless, our result is sharper as we establish an upper bound of the rate of convergence (in probability) of the normalised likelihood. We fix some $\theta \in \Theta$ and introduce the quantities

$$\hat{z}^{1:T} = \underset{z^{1:T}\in[\![1,Q]\!]^{nT}}{\arg\max} \log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T}), \tag{2.5.1}$$

$$\tilde{Z}^{1:T} = \underset{z^{1:T}\in[\![1,Q]\!]^{nT}}{\arg\max} \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T}) \,\middle|\, Z^{1:T}\right]. \tag{2.5.2}$$

Note that $\tilde{Z}^{1:T}$ is a random variable that depends on $Z^{1:T}$ and that

$$\begin{aligned}\hat{z}^{1:T} &= \underset{z^{1:T}\in[\![1,Q]\!]^{nT}}{\arg\max} \sum_{t=1}^{T} \log \mathbb{P}_\theta(X^t \mid Z^t = z^t) \\ &= \left(\underset{z\in[\![1,Q]\!]^{n}}{\arg\max} \log \mathbb{P}_\theta(X^1 \mid Z^1 = z), \ldots, \underset{z\in[\![1,Q]\!]^{n}}{\arg\max} \log \mathbb{P}_\theta(X^T \mid Z^T = z)\right).\end{aligned} \tag{2.5.3}$$

Similarly, for any $t \in [\![1,T]\!]$, we have $\tilde{Z}^t = \arg\max_{z\in[\![1,Q]\!]^n} \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^t \mid Z^t = z) \mid Z^t\right]$.

We bound the difference between $M_{n,T}(\Gamma,\pi)$ and $\mathbb{M}(\pi)$ by introducing three intermediate terms so that we can write, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ and any $\epsilon > 0$

$$\begin{aligned}&\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta} |M_{n,T}(\Gamma,\pi) - \mathbb{M}(\pi)| > \frac{\epsilon r_{n,T}}{\sqrt{n}}\right) \\ \leq &\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\log \mathbb{P}_\theta(X^{1:T}) - \frac{2}{n(n-1)T}\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T})\right| > \frac{\epsilon r_{n,T}}{3\sqrt{n}}\right) \\ &+ \mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T})\right.\right. \\ &\qquad\left.\left. - \frac{2}{n(n-1)T}\mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) \,\middle|\, Z^{1:T}\right]\right| > \frac{\epsilon r_{n,T}}{3\sqrt{n}}\right) \\ &+ \mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) \,\middle|\, Z^{1:T}\right] - \mathbb{M}(\pi)\right| > \frac{\epsilon r_{n,T}}{3\sqrt{n}}\right).\end{aligned} \tag{2.5.4}$$

In the following, we prove separately the convergence (in $\mathbb{P}_{\theta^*}$-probability) to zero of the three terms of this sum (while controlling for the rate of these convergences). Before

starting, let us remark that we have

$$\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T}) = \sum_{t=1}^{T} \sum_{1\le i<j\le n} X_{ij}^t \log \pi_{z_i^t z_j^t} + (1 - X_{ij}^t) \log(1 - \pi_{z_i^t z_j^t}) \tag{2.5.5}$$

$$\begin{aligned} \text{and} \quad & \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T}) \,\middle|\, Z^{1:T}\right] \\ &= \sum_{t=1}^{T} \sum_{1\le i<j\le n} \pi^*_{Z_i^t Z_j^t} \log \pi_{z_i^t z_j^t} + (1 - \pi^*_{Z_i^t Z_j^t}) \log(1 - \pi_{z_i^t z_j^t}). \end{aligned} \tag{2.5.6}$$

In particular, for every $t \in [\![1, T]\!]$, we have

$$\begin{aligned} \hat{z}^t &= \operatorname*{arg\,max}_{z=(z_1,\dots,z_n)\in[\![1,Q]\!]^n} \sum_{1\le i<j\le n} X_{ij}^t \log \pi_{z_i z_j} + (1 - X_{ij}^t) \log(1 - \pi_{z_i z_j}), \\ \tilde{Z}^t &= \operatorname*{arg\,max}_{z=(z_1,\dots,z_n)\in[\![1,Q]\!]^n} \sum_{1\le i<j\le n} \pi^*_{Z_i^t Z_j^t} \log \pi_{z_i z_j} + (1 - \pi^*_{Z_i^t Z_j^t}) \log(1 - \pi_{z_i z_j}). \end{aligned}$$

First term of the right-hand side of (2.5.4). We let

$$\begin{aligned} T_1 &:= \left| \frac{2}{n(n-1)T} \log \mathbb{P}_\theta(X^{1:T}) - \frac{2}{n(n-1)T} \log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T}) \right| \\ &\le \frac{2}{n(n-1)T} \sum_{t=1}^{T} \left| \log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) - \log \mathbb{P}_\theta(X^t \mid Z^t = \hat{z}^t) \right|. \end{aligned} \tag{2.5.7}$$

Lemma 2.5.1. *For every $t \in [\![1, T]\!]$, we have*

$$\left| \log \mathbb{P}_\theta(X^t | X^{1:t-1}) - \log \mathbb{P}_\theta(X^t | Z^t = \hat{z}^t) \right| \le \left| \log \mathbb{P}_\theta(Z^t = \hat{z}^t | X^{1:t-1}) \right|.$$

Going back to (2.5.7) and applying Lemma 2.5.1, we get

$$T_1 \le \frac{2}{n(n-1)T} \sum_{t=1}^{T} \left| \log \mathbb{P}_\theta(Z^t = \hat{z}^t \mid X^{1:t-1}) \right| = -\frac{2}{n(n-1)T} \sum_{t=1}^{T} \log \mathbb{P}_\theta(Z^t = \hat{z}^t \mid X^{1:t-1}).$$

Now, using classical dependency rules in directed acyclic graphs (see for e.g. Lauritzen, 1996) combined with Assumption 2, we get

$$
\begin{aligned}
T_1 &\leq -\frac{2}{n(n-1)T}\sum_{t=1}^{T}\log \sum_{z^{t-1}\in[\![1,Q]\!]^n} \mathbb{P}_\theta(Z^t=\hat{z}^t \mid Z^{t-1}=z^{t-1})\mathbb{P}_\theta(Z^{t-1}=z^{t-1}\mid X^{1:t-1}) \\
&\leq -\frac{2}{n(n-1)T}\sum_{t=1}^{T}\log \sum_{z^{t-1}\in[\![1,Q]\!]^n} \delta^n \mathbb{P}_\theta(Z^{t-1}=z^{t-1}\mid X^{1:t-1}) \\
&\leq -\frac{2}{n(n-1)T}\sum_{t=1}^{T} n\log\delta = \frac{2}{n-1}\log(1/\delta).
\end{aligned}
$$

This implies that $\mathbb{P}_{\theta^*}(\sup_{\theta\in\Theta} T_1 > \epsilon r_{n,T}/(3\sqrt{n})) = 0$ as soon as we have $\epsilon r_{n,T}/\sqrt{n} \geq 6\log(1/\delta)/(n-1)$. Then for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity, for any $\epsilon > 0$, we have that $\mathbb{P}_{\theta^*}(\sup_{\theta\in\Theta} T_1 > \epsilon r_{n,T}/(3\sqrt{n})) \to 0$ as n and T increase.

Second term of the right-hand side of (2.5.4). Let us denote

$$
\begin{aligned}
T_2(Z^{1:T}) := \Bigg| &\frac{2}{n(n-1)T}\log\mathbb{P}_\theta(X^{1:T}\mid Z^{1:T}=\hat{z}^{1:T}) \\
&- \frac{2}{n(n-1)T}\mathbb{E}_{\theta^*}\left[\log\mathbb{P}_\theta(X^{1:T}|Z^{1:T}=\tilde{Z}^{1:T}) \,\middle|\, Z^{1:T}\right]\Bigg|.
\end{aligned}
$$

For the sake of clarity, we study this term on the event $\{Z^{1:T}=z^{*1:T}\}$ where $z^{*1:T}\in[\![1,Q]\!]^{nT}$ is a fixed configuration. This event induces the definition of $\tilde{Z}^{1:T}$ following Equation (2.5.2) as

$$
\tilde{Z}^{1:T} = \underset{z^{1:T}\in[\![1,Q]\!]^{nT}}{\arg\max}\ \mathbb{E}_{\theta^*}\left[\log\mathbb{P}_\theta(X^{1:T}\mid Z^{1:T}=z^{1:T}) \,\middle|\, Z^{1:T}=z^{*1:T}\right],
$$

or equivalently for every $t\in[\![1,T]\!]$,

$$
\tilde{Z}^t = \underset{z=(z_1,\dots,z_n)\in[\![1,Q]\!]^n}{\arg\max} \sum_{1\leq i<j\leq n} \pi^*_{z_i^{*t}z_j^{*t}}\log\pi_{z_iz_j} + (1-\pi^*_{z_i^{*t}z_j^{*t}})\log(1-\pi_{z_iz_j}).
$$

By definition of $\hat{z}^{1:T}$ and $\tilde{Z}^{1:T}$ respectively, we have the two inequalities

$$
\log\mathbb{P}_\theta(X^{1:T}\mid Z^{1:T}=\hat{z}^{1:T}) \geq \log\mathbb{P}_\theta(X^{1:T}\mid Z^{1:T}=\tilde{Z}^{1:T})
$$

and

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) \,\middle|\, Z^{1:T} = z^{*1:T}\right]\\ \geq&\mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T}) \,\middle|\, Z^{1:T} = z^{*1:T}\right],\end{aligned}$$

implying the lower and upper bounds

$$\begin{aligned}&\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) - \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) \,\middle|\, Z^{1:T} = z^{*1:T}\right]\\ &\quad\leq \log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T}) - \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) \,\middle|\, Z^{1:T} = z^{*1:T}\right]\\ &\quad\leq \log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T}) - \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \hat{z}^{1:T}) \,\middle|\, Z^{1:T} = z^{*1:T}\right].\end{aligned}$$

Taking the absolute value gives us an upper bound for $T_2(z^{*1:T})$

$$\begin{aligned}T_2(z^{*1:T}) \leq \max_{z^{1:T}\in\{\hat{z}^{1:T},\tilde{Z}^{1:T}\}} \frac{2}{n(n-1)T}\Big|&\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T})\\ &- \mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T}) \,\middle|\, Z^{1:T} = z^{*1:T}\right]\Big|.\end{aligned}$$

Using Equations (2.5.5) and (2.5.6), we then obtain the following upper bound for $T_2(z^{*1:T})$

$$T_2(z^{*1:T}) \leq \max_{z^{1:T}\in\{\hat{z}^{1:T},\tilde{Z}^{1:T}\}} \left|\frac{2}{n(n-1)T}\sum_{t=1}^{T}\sum_{1\leq i<j\leq n}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^t z_j^t}}{1-\pi_{z_i^t z_j^t}}\right)\right|.$$

We use the following concentration result to conclude.

Lemma 2.5.2. *Let $\epsilon, \beta > 0$ and $\{x_{n,T}\}_{n,T\geq 1}$ a sequence of positive real numbers. We let $\mathbb{P}^*_{\theta^*}(\cdot)$ denote the probability conditional on $\{Z^{1:T} = z^{*1:T}\}$ under parameter θ^*, i.e. $\mathbb{P}^*_{\theta^*}(\cdot) = \mathbb{P}_{\theta^*}(\cdot \mid Z^{1:T} = z^{*1:T})$. Denoting $\Lambda = 2\log[(1-\zeta)/\zeta] > 0$ we have for any $\theta \in \Theta$*

$$\begin{aligned}&\mathbb{P}^*_{\theta^*}\left(\sup_{z^{1:T}\in[\![1,Q]\!]^{nT}}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{1\leq i<j\leq n}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^t z_j^t}}{1-\pi_{z_i^t z_j^t}}\right)\right| > \epsilon\right)\\ \leq&\mathbb{P}^*_{\theta^*}\left[\frac{(1+\beta)\Lambda}{\sqrt{n(n-1)T/2}} + \frac{\Lambda\sqrt{x_{n,T}/2}}{\sqrt{n(n-1)T/2}} + (1/\beta+1/3)\frac{(\Lambda/2)x_{n,T}}{n(n-1)T/2} > \epsilon\right] + 2e^{-x_{n,T}}\\ \leq&\mathbb{1}_{2\Omega/(n(n-1)T)>\epsilon} + 2e^{-x_{n,T}}\end{aligned} \tag{2.5.8}$$

with $\Omega = (1+\beta)\Lambda\sqrt{n(n-1)T/2} + \Lambda\sqrt{n(n-1)Tx_{n,T}/4} + (1/\beta + 1/3)(\Lambda/2)x_{n,T}$.

Let us choose $x_{n,T} = \log(n)$ in the above lemma. For any $\epsilon > 0$, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity, we have for n and T large enough

$$\frac{\epsilon r_{n,T}}{3\sqrt{n}} \geq \frac{2\Omega}{n(n-1)T}.$$

Then for n and T large enough, the first term in the right-hand side of inequality (2.5.8) is equal to 0 and we have

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\sup_{\theta\in\Theta} T_2(z^{*1:T}) > \frac{\epsilon r_{n,T}}{3\sqrt{n}}\right) \leq \frac{2}{n} \\
\text{and}\quad &\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta} T_2(Z^{1:T}) > \frac{\epsilon r_{n,T}}{3\sqrt{n}}\right) \leq \sum_{z^{*1:T}} \mathbb{P}^*_{\theta^*}\left(\sup_{\theta\in\Theta} T_2(z^{*1:T}) > \frac{\epsilon r_{n,T}}{3\sqrt{n}}\right)\mathbb{P}_{\theta^*}(Z^{1:T} = z^{*1:T}) \\
&\qquad\qquad \leq \frac{2}{n}.
\end{aligned}$$

Third term of the right-hand side of (2.5.4). Let us denote

$$\begin{aligned}
T_3(Z^{1:T}) &:= \left|\frac{2}{n(n-1)T}\mathbb{E}_{\theta^*}\left[\log\mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = \tilde{Z}^{1:T}) \,\middle|\, Z^{1:T}\right] - \mathbb{M}(\pi)\right| \\
&= \left|\frac{2}{n(n-1)T}\sum_{t=1}^{T}\mathbb{E}_{\theta^*}\left[\log\mathbb{P}_\theta(X^t \mid Z^t = \tilde{Z}^t) \,\middle|\, Z^t\right] - \mathbb{M}(\pi, \bar{A}_\pi)\right|.
\end{aligned}$$

For any fixed configuration $z^t \in [\![1, Q]\!]^n$, analogous to Equation (2.5.6), we write

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left[\log\mathbb{P}_\theta(X^t \mid Z^t = z^t) \,\middle|\, Z^t\right] \\
=&\sum_{1\leq i<j\leq n} \pi^*_{Z^t_i Z^t_j}\log\pi_{z^t_i z^t_j} + (1-\pi^*_{Z^t_i Z^t_j})\log(1-\pi_{z^t_i z^t_j}) \\
=&\frac{1}{2}\sum_{1\leq i\neq j\leq n} \pi^*_{Z^t_i Z^t_j}\log\pi_{z^t_i z^t_j} + (1-\pi^*_{Z^t_i Z^t_j})\log(1-\pi_{z^t_i z^t_j}) \\
=&\frac{1}{2}\sum_{1\leq q,l,q',l'\leq Q}\ \sum_{1\leq i\neq j\leq n}\left(\pi^*_{ql}\log\pi_{q'l'} + (1-\pi^*_{ql})\log(1-\pi_{q'l'})\right)\mathbb{1}_{\{Z^t_i=q, Z^t_j=l, z^t_i=q', z^t_j=l'\}} \\
=&\frac{1}{2}\sum_{1\leq q,l,q',l'\leq Q} C_{qq'}(Z^t, z^t)C_{ll'}(Z^t, z^t)\left(\pi^*_{ql}\log\pi_{q'l'} + (1-\pi^*_{ql})\log(1-\pi_{q'l'})\right),
\end{aligned}$$

where $C_{qq'}(Z^t, z^t) = |\{i \in [\![1, n]\!]; Z^t_i = q, z^t_i = q'\}|$ is the (random variable) number of nodes classified in group q in the current (random) configuration Z^t, while they belong to group q' in (deterministic) configuration z^t. Recall that $N_q(z^t)$ is the number of

nodes assigned to class q by the configuration z^t and let us denote $a^t_{qq'} = a_{qq'}(Z^t, z^t) = C_{qq'}(Z^t, z^t)/N_q(Z^t)$ the (random) proportion of vertices from class q in Z^t attributed to class q' by z^t. We write

$$\begin{aligned}
&\frac{2}{n(n-1)}\mathbb{E}_{\theta^*}\left[\log \mathbb{P}_\theta(X^t \mid Z^t = z^t) \,\middle|\, Z^t\right] \\
=&\sum_{1\le q,l,q',l'\le Q} \frac{N_q(Z^t)N_l(Z^t)}{n(n-1)} a^t_{qq'}a^t_{ll'}\left(\pi^*_{ql}\log\pi_{q'l'} + (1-\pi^*_{ql})\log(1-\pi_{q'l'})\right) \\
:=&\Phi^t(A^t,\pi),
\end{aligned}$$

with $A^t = (a^t_{qq'})_{1\le q,q'\le Q}$.

Now extending these notations to the case where $z^t = \tilde{Z}^t$, we let $\tilde{A}^t = (\tilde{a}^t_{qq'})_{1\le q,q'\le Q}$ where $\tilde{a}^t_{qq'} = a_{qq'}(Z^t, \tilde{Z}^t)$. We remark that the definition of $\tilde{Z}^t$ implies that $\tilde{A}^t = \arg\max_{A^t\in\mathcal{A}^t(Z^{1:T})} \Phi^t(A^t,\pi)$ with $\mathcal{A}^t(Z^{1:T})$ the (random) subset of stochastic matrices defined for every $t \in [\![1,T]\!]$ by

$$\mathcal{A}^t(Z^{1:T}) = \left\{A = (n_{ql}/N_q(Z^t))_{1\le q,l\le Q}; n_{ql} \in [\![0, N_q(Z^t)]\!], \sum_{l=1}^{Q} n_{ql} = N_q(Z^t)\right\}.$$

Let us also denote $\bar{A}^t_\pi = \arg\max_{A\in\mathcal{A}^t(Z^{1:T})} \mathbb{M}(\pi, A)$. Then

$$\begin{aligned}
\sup_{\theta\in\Theta} T_3(Z^{1:T}) &\le \sup_{\pi\in[\zeta,1-\zeta]^{Q^2}} \frac{1}{T}\sum_{t=1}^{T}\left|\Phi^t(\tilde{A}^t,\pi) - \mathbb{M}(\pi,\bar{A}_\pi)\right| \\
&\le \sup_{\pi\in[\zeta,1-\zeta]^{Q^2}} \frac{1}{T}\sum_{t=1}^{T}\left|\Phi^t(\tilde{A}^t,\pi) - \mathbb{M}(\pi,\bar{A}^t_\pi)\right| \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\left|\mathbb{M}(\pi,\bar{A}^t_\pi) - \mathbb{M}(\pi,\bar{A}_\pi)\right|. \qquad (2.5.9)
\end{aligned}$$

We start by stating a concentration lemma on the random variable $N_q(Z^t)$ for any $q \in [\![1,Q]\!]$ and any $t \in [\![1,T]\!]$.

Lemma 2.5.3. *For any $\theta \in \Theta$ and any $\eta \in (0,\delta)$, let*

$$\Omega_\eta(\theta) := \left\{z^{1:T} \in [\![1,Q]\!]^{nT}; \forall t \in [\![1,T]\!], \forall q \in [\![1,Q]\!], \frac{N_q(z^t)}{n} \ge \alpha_q - \eta\right\}.$$

Then $\mathbb{P}_\theta\left(Z^{1:T} \in \Omega_\eta(\theta)\right) \ge 1 - QT\exp(-2\eta^2 n)$.

Building on the previous concentration lemma, the following one gives the convergence in $\mathbb{P}_{\theta^*}$-probability of the second term in the right-hand side of (2.5.9).

Lemma 2.5.4. *For any $\epsilon > 0$, any $\eta \in (0, \delta)$ and $\{r_{n,T}\}_{n,T\geq 1}$ any positive sequence,*

$$\begin{aligned}&\mathbb{P}_{\theta^*}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|>\frac{\epsilon r_{n,T}}{6\sqrt{n}}\right)\\ \leq& QT\exp\left(-2\eta^2 n\right)+\mathbb{1}_{n\leq 6c\sqrt{n}/[\epsilon r_{n,T}(\delta-\eta)]}\end{aligned} \tag{2.5.10}$$

with $c = 6(1-\delta)^2(1-\zeta)\log(1/\zeta)Q^4$.

Then taking any $\eta \in (0, \delta)$, for any $\epsilon > 0$, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity, we have the following inequality for n and T large enough

$$r_{n,T} > \frac{6c\sqrt{n}}{\epsilon(\delta-\eta)n}, \tag{2.5.11}$$

implying that the probability in Lemma 2.5.4 converges to 0 as n and T increase for any $\epsilon > 0$, as long as $\log T = o(n)$. Now, for the first term in the right-hand side of (2.5.9), note that we have for every π and every t

$$\begin{cases}\Phi^t(\tilde{A}^t,\pi)\geq\Phi^t(\bar{A}_\pi^t,\pi) & \text{because } \tilde{A}^t=\arg\max_{A\in\mathcal{A}^t}\Phi^t(A,\pi)\\ \mathbb{M}(\pi,\bar{A}_\pi^t)\geq\mathbb{M}(\pi,\tilde{A}^t) & \text{because } \bar{A}_\pi^t=\arg\max_{A\in\mathcal{A}^t}\mathbb{M}(\pi,A).\end{cases}$$

Then, either $\mathbb{M}(\pi,\bar{A}_\pi^t)\leq\Phi^t(\tilde{A}^t,\pi)$ and

$$0\leq\Phi^t(\tilde{A}^t,\pi)-\mathbb{M}(\pi,\bar{A}_\pi^t)\leq\Phi^t(\tilde{A}^t,\pi)-\mathbb{M}(\pi,\tilde{A}^t)$$

or $\mathbb{M}(\pi,\bar{A}_\pi^t)\geq\Phi^t(\tilde{A}^t,\pi)$ and

$$0\leq\mathbb{M}(\pi,\bar{A}_\pi^t)-\Phi^t(\tilde{A}^t,\pi)\leq\mathbb{M}(\pi,\bar{A}_\pi^t)-\Phi^t(\bar{A}_\pi^t,\pi).$$

In both cases, we get that $\left|\Phi^t(\tilde{A}^t,\pi)-\mathbb{M}(\pi,\bar{A}_\pi^t)\right| \leq \sup_{A\in\mathcal{A}}|\Phi^t(A,\pi)-\mathbb{M}(\pi,A)|$ for every t and π, thus obtaining the upper bound

$$\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\frac{1}{T}\sum_{t=1}^{T}\left|\Phi^t(\tilde{A}^t,\pi)-\mathbb{M}(\pi,\bar{A}_\pi^t)\right|\leq\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\sup_{A^t\in\mathcal{A}}\left|\Phi^t(A^t,\pi)-\mathbb{M}(\pi,A^t)\right|.$$

Letting

$$\Delta(\zeta)=\sup_{\pi\in[\zeta,1-\zeta]}\sup_{\pi^*\in[\zeta,1-\zeta]}\left|\pi^*\log\pi+(1-\pi^*)\log(1-\pi)\right|\in(0,+\infty)$$

and recalling that $0 \leq a_{ql} \leq 1$ (for every $q, l \in [\![1, Q]\!]$) for every $A = (a_{ql})_{1 \leq q,l \leq Q} \in \mathcal{A}$, we have

$$\begin{aligned}
&\sup_{\pi \in [\zeta, 1-\zeta]^{Q^2}} \sup_{A^t \in \mathcal{A}} \left| \Phi^t(A^t, \pi) - \mathbb{M}(\pi, A^t) \right| \\
&\leq \sup_{\pi \in [\zeta, 1-\zeta]^{Q^2}} \sup_{A^t \in \mathcal{A}} \sum_{1 \leq q,l,q',l' \leq Q} \left| \left(\frac{N_q(Z^t) N_l(Z^t)}{n(n-1)} - \alpha_q^* \alpha_l^* \right) a^t_{qq'} a^t_{ll'} \right. \\
&\qquad \left. \times \left(\pi^*_{ql} \log \pi_{q'l'} + (1 - \pi^*_{ql}) \log(1 - \pi_{q'l'}) \right) \right| \\
&\leq \Delta(\zeta) Q^2 \sum_{1 \leq q,l \leq Q} \left| \frac{N_q(Z^t) N_l(Z^t)}{n(n-1)} - \alpha_q^* \alpha_l^* \right|.
\end{aligned}$$

Finally, we bound the first term of the right-hand-side of (2.5.9) as follows

$$\sup_{\pi \in [\zeta, 1-\zeta]^{Q^2}} \frac{1}{T} \sum_{t=1}^{T} \left| \Phi^t(\tilde{A}^t, \pi) - \mathbb{M}(\pi, \bar{A}^t_\pi) \right| \leq \Delta(\zeta) Q^2 \sum_{1 \leq q,l \leq Q} \frac{1}{T} \sum_{t=1}^{T} \left| \frac{N_q(Z^t) N_l(Z^t)}{n(n-1)} - \alpha_q^* \alpha_l^* \right|. \tag{2.5.12}$$

Applying Markov's Inequality, we obtain

$$\begin{aligned}
&\mathbb{P}_{\theta^*} \left(\sup_{\pi \in [\zeta, 1-\zeta]^{Q^2}} \frac{1}{T} \sum_{t=1}^{T} \left| \Phi^t(\tilde{A}^t, \pi) - \mathbb{M}(\pi, \bar{A}^t_\pi) \right| > \frac{\epsilon r_{n,T}}{6\sqrt{n}} \right) \\
&\leq \sum_{q,l} \mathbb{P}_{\theta^*} \left(\frac{1}{T} \sum_{t=1}^{T} \left| \frac{N_q(Z^t) N_l(Z^t)}{n(n-1)} - \alpha_q^* \alpha_l^* \right| > \frac{\epsilon r_{n,T}}{6\Delta(\zeta) Q^4 \sqrt{n}} \right) \\
&\leq \frac{6\Delta(\zeta) Q^4 \sqrt{n}}{\epsilon r_{n,T}} \sum_{q,l} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\theta^*} \left[\left| \frac{N_q(Z^t) N_l(Z^t)}{n(n-1)} - \alpha_q^* \alpha_l^* \right| \right] \\
&\leq \frac{6\Delta(\zeta) Q^4 \sqrt{n}}{\epsilon r_{n,T}} \sum_{q,l} \mathbb{E}_{\theta^*} \left[\left| \frac{N_q(Z^1) N_l(Z^1)}{n(n-1)} - \alpha_q^* \alpha_l^* \right| \right].
\end{aligned}$$

The following lemma gives an upper bound of the expectation appearing in the previous inequality, for any $q, l \in [\![1, Q]\!]$.

Lemma 2.5.5. *For any $q, l \in [\![1, Q]\!]$ and any $t \in [\![1, T]\!]$, we have the following inequality*

$$\mathbb{E}_{\theta^*} \left[\left| \frac{N_q(Z^t) N_l(Z^t)}{n(n-1)} - \alpha_q^* \alpha_l^* \right| \right] \leq \frac{2\sqrt{n}}{n-1}.$$

This leads to

$$\mathbb{P}_{\theta^*}\left(\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\frac{1}{T}\sum_{t=1}^{T}\left|\Phi^t(\tilde{A}^t,\pi)-\mathbb{M}(\pi,\bar{A}^t_\pi)\right|>\frac{\epsilon r_{n,T}}{6\sqrt{n}}\right)\leq\frac{12\Delta(\zeta)Q^6 n}{\epsilon r_{n,T}(n-1)}.$$

Then for any $\epsilon>0$, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity, we have the convergence

$$\mathbb{P}_{\theta^*}\left(\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\frac{1}{T}\sum_{t=1}^{T}\left|\Phi^t(\tilde{A}^t,\pi)-\mathbb{M}(\pi,\bar{A}^t_\pi)\right|>\epsilon r_{n,T}/(6\sqrt{n})\right)\xrightarrow[n,T\to\infty]{}0.$$

We proved the convergence to 0 of the three terms in the right-hand side of (2.5.4) for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity and as long as $\log T=o(n)$. This gives the expected result and concludes the proof. □

2.5.2 Proof of Corollary 2.3.1

To prove this corollary, we establish the following lemma that allows us to obtain a rate of convergence of $\hat{\pi}$ to π^* from a rate of convergence of $M_{n,T}$ to $\mathbb{M}$. Note that this lemma is a bit more general than what we need and gives an equivalent result when the number of time steps T is fixed, which will be useful for Corollary 2.3.2.

Lemma 2.5.6. *Let $\{F_{n,T}\}_{n,T\geq 1}$ be any random functions on the set Θ (resp. Θ^T) and $\mathbb{M}$ (resp. $\mathbb{M}^T$) defined as before. Assume that there exists a sequence $\{v_{n,T}\}_{n,T\geq 1}$ (resp. $\{v_n\}_{n\geq 1}$) a sequence decreasing to 0 such that for every $\epsilon>0$, we have the following convergence as $n,T\to\infty$ (resp. $n\to\infty$)*

$$\mathbb{P}_{\theta^*}\left(\sup_{(\Gamma,\pi)\in\Theta}|F_{n,T}(\Gamma,\pi)-\mathbb{M}(\pi)|>\epsilon v_{n,T}\right)\xrightarrow[n,T\to\infty]{}0$$

$$\left(\text{resp. }\mathbb{P}_{\theta^*}\left(\sup_{(\Gamma,\pi)\in\Theta^T}\left|F_{n,T}(\Gamma,\pi^{1:T})-\mathbb{M}^T(\pi^{1:T})\right|>\epsilon v_n\right)\xrightarrow[n\to\infty]{}0\right).$$

If for any n and T, $\hat{\theta}=(\hat{\Gamma},\hat{\pi})$ (resp. $\hat{\theta}=(\hat{\Gamma},\hat{\pi}^{1:T})$) is defined as the maximizer of $F_{n,T}$ on the set Θ (resp. Θ^T), we have the following convergence

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma-\pi^*\|_\infty>\epsilon\sqrt{v_{n,T}}\right)\xrightarrow[n,T\to\infty]{}0$$

$$\left(\text{resp. }\mathbb{P}_{\theta^*}\left(\min_{\sigma^1,\dots,\sigma^T\in\mathfrak{S}_Q}\|\hat{\pi}^{1:T}_{\sigma^{1:T}}-\pi^{*1:T}\|_\infty>\epsilon\sqrt{v_n}\right)\xrightarrow[n\to\infty]{}0\right)$$

with $\hat{\pi}^{1:T}_{\sigma^{1:T}} = (\hat{\pi}^t_{\sigma^t})_{t\in[\![1,T]\!]}$.

The result of Corollary 2.3.1 is then a direct consequence of Theorem 2.3.1 (choosing the sequence $\{r^2_{n,T}\}_{n,t\geq 1}$) and Lemma 2.5.6 applied with $F_{n,T} = M_{n,T}$. □

2.5.3 Proof of Theorem 2.3.2

The proof follows the lines of the proof of Theorem 3.8 in Celisse et al. (2012). Nonetheless, our result is sharper as we will establish an upper bound of the rate of convergence (in probability) of the quantity at stake. For any $\epsilon > 0$, any sequence $\{y_{n,T}\}_{n,T\geq 1}$ and $\eta \in (0, \delta)$, we write

$$\begin{aligned}
&\mathbb{P}_{\theta^*}(\mathcal{E}(Z^{1:T}, \breve{\theta}, \epsilon y_{n,T})) \\
&= \sum_{z^{*1:T}\in[\![1,Q]\!]^{nT}} \mathbb{P}_{\theta^*}(\mathcal{E}(z^{*1:T}, \breve{\theta}, \epsilon y_{n,T}); Z^{1:T} = z^{*1:T}) \leq \mathbb{P}_{\theta^*}(Z^{1:T} \in \Omega^c_\eta(\theta^*)) \\
&\quad + \sum_{z^{*1:T}\in\Omega_\eta(\theta^*)} \mathbb{P}_{\theta^*}\left(\frac{\mathbb{P}_{\breve{\theta}}\left(Z^{1:T} \neq z^{*1:T} \mid X^{1:T}\right)}{\mathbb{P}_{\breve{\theta}}\left(Z^{1:T} = z^{*1:T} \mid X^{1:T}\right)} > \epsilon y_{n,T} \,\middle|\, Z^{1:T} = z^{*1:T} \right) \mathbb{P}_{\theta^*}\left(Z^{1:T} = z^{*1:T}\right)
\end{aligned} \tag{2.5.13}$$

with $\Omega_\eta(\theta^*)$ as defined in Lemma 2.5.3. We will establish that there exist some positive constants C, C_1, C_2, C_3, C_4 such that for any fixed configuration $z^{*1:T} \in \Omega_\eta(\theta^*)$, any $\epsilon > 0$, any positive sequence $\{y_{n,T}\}_{n,T\geq 1}$ such that $\log(1/y_{n,T}) = o(n)$ and n and T large enough, we have

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left[\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T} \,\middle|\, Z^{1:T} = z^{*1:T} \right] \\
\leq &\mathbb{P}_{\theta^*}\left(\|\breve{\pi} - \pi^*\|_\infty > v_{n,T} \mid Z^{1:T} = z^{*1:T} \right) \\
&\quad + CnT\left\{ \exp\left[-(\delta-\eta)^2 C_1 n + C_2 \log(nT) + C_4 \log(1/(\epsilon y_{n,T})) \right] \right. \\
&\qquad \left. + \exp\left[-C_3 \frac{(\log(nT))^2}{n v^2_{n,T}} + 3n \log(nT) \right] \right\}.
\end{aligned} \tag{2.5.14}$$

Combined with (2.5.13) and applying Lemma 2.5.3, this gives the desired result. So now we focus on establishing (2.5.14).

In what follows, we consider a fixed configuration $z^{*1:T} \in \Omega_\eta(\theta^*)$ and introduce the Hamming distance between $z^{*1:T}$ and any other configuration $z^{1:T}$ defined as

$$\|z^{1:T} - z^{*1:T}\|_0 = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{1}_{z_i^t \neq z_i^{*t}}.$$

We let $\mathbb{P}^*_{\theta^*}(\cdot)$ denote the probability conditional on $\{Z^{1:T} = z^{*1:T}\}$ under parameter $\theta = \theta^*$, i.e. $\mathbb{P}^*_{\theta^*}(\cdot) = \mathbb{P}_{\theta^*}(\cdot \mid Z^{1:T} = z^{*1:T})$. In the following, we will often use the fact that the variables $\{X_{ij}^t\}$ are independent under $\mathbb{P}^*_{\theta^*}$ (with mean value $\pi^*_{z_i^{*t} z_j^{*t}}$) so that we can rely on Hoeffding's Inequality. We introduce a sequence $\{v_{n,T}\}_{n,T\geq 1}$ decreasing to 0 and $\Omega_{n,T}$ the event defined as

$$\Omega_{n,T} = \{\|\breve{\pi} - \pi^*\|_\infty \leq v_{n,T}\}.$$

We bound the probability of interest in (2.5.14) by splitting it on the two complementary events $\Omega_{n,T}$ and $\Omega_{n,T}^c$. For any $\epsilon > 0$ and any positive sequence $\{y_{n,T}\}_{n,T\geq 1}$

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left[\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}\right] \\
\leq &\mathbb{P}^*_{\theta^*}\left(\Omega_{n,T}^c\right) + \mathbb{P}^*_{\theta^*}\left[\left\{\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}\right\} \cap \Omega_{n,T}\right]. \qquad (2.5.15)
\end{aligned}$$

Thus, the proof of (2.5.14) boils down to establishing the desired upper bound on the second term appearing in the right-hand side of (2.5.15). We have

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left[\left\{\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}\right\} \cap \Omega_{n,T}\right] \\
\leq &\sum_{r=1}^{nT} \sum_{z^{1:T}; \|z^{1:T} - z^{*1:T}\|_0 = r} \mathbb{P}^*_{\theta^*}\left[\left\{\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}/(Q^r (nT)^{r+1})\right\} \cap \Omega_{n,T}\right],
\end{aligned}$$

by using the bound $(Q-1)^r\binom{nT}{r} \leq Q^r(nT)^r$ on the number of terms in the sum over $\{z^{1:T}; \|z^{1:T} - z^{*1:T}\|_0 = r\}$ (for each value of r). Then,

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left[\left\{\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}\right\} \cap \Omega_{n,T}\right] \\
&\leq \sum_{r=1}^{nT} \sum_{\substack{z^{1:T}; \\ \|z^{1:T} - z^{*1:T}\|_0 = r}} \mathbb{P}^*_{\theta^*}\left[\left\{\log \frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \log(\epsilon y_{n,T}) - r\log Q \right.\right. \\
&\qquad\qquad \left.\left. - (r+1)\log(nT)\right\} \cap \Omega_{n,T}\right] \\
&\leq \sum_{r=1}^{nT} \sum_{\substack{z^{1:T}; \\ \|z^{1:T} - z^{*1:T}\|_0 = r}} \mathbb{P}^*_{\theta^*}\left[\left\{\log \frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > -\log\left(\frac{1}{\epsilon y_{n,T}}\right) \right.\right. \\
&\qquad\qquad \left.\left. - 3r\log(nT)\right\} \cap \Omega_{n,T}\right], \quad (2.5.16)
\end{aligned}$$

as long as $nT \geq Q$. For any configuration $z^{1:T}$ such that $\|z^{1:T} - z^{*1:T}\|_0 = r$, we denote by $r(1), \dots, r(T)$ the number of differences between the two configurations at each time step $t \in [\![1, T]\!]$, i.e. $r(t) = \|z^t - z^{*t}\|_0$ such that $r = \sum_t r(t)$. Moreover, for any parameter π, we define $D_{n,T}(z^{1:T}, \pi)$ the subset of indexes $(i, j, t) \in [\![1, n]\!]^2 \times [\![1, T]\!]$ such that $i < j$ for which the parameter π differs between the configuration $z^{*1:T}$ and $z^{1:T}$, namely

$$D_{n,T}(z^{1:T}, \pi) := \left\{(i, j, t) \in I_{n,T}; \pi_{z_i^t z_j^t} \neq \pi_{z_i^{*t} z_j^{*t}}\right\},$$

with $I_{n,T} = \{(i, j, t) \in [\![1, n]\!]^2 \times [\![1, T]\!]; i < j\}$ the set of indexes over which we sum to compute the conditional log-likelihood. In what follows, we abbreviate to D^* (resp. $\breve{D}$), the set $D_{n,T}(z^{1:T}, \pi^*)$ (resp. $D_{n,T}(z^{1:T}, \breve{\pi})$). Next lemma gives a decomposition of the main term at stake in (2.5.16).

Lemma 2.5.7. *We have the decomposition*

$$\log \frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} = U_1 + U_2 - U_3,$$

where

$$U_1 := \sum_{(i,j,t)\in D^*} \left(X_{ij}^t \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} + (1 - X_{ij}^t) \log \frac{1 - \pi^*_{z_i^t z_j^t}}{1 - \pi^*_{z_i^{*t} z_j^{*t}}} \right) + \sum_{i=1}^{n} \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^{n} \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \tag{2.5.17}$$

$$U_2 := \sum_{(i,j,t)\in D^*\cup\breve{D}} \log \left[1 + \frac{(\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t})(X_{ij}^t - \pi^*_{z_i^t z_j^t})}{\pi^*_{z_i^t z_j^t}(1 - \pi^*_{z_i^t z_j^t})} \right] \tag{2.5.18}$$

$$U_3 := \sum_{(i,j,t)\in D^*\cup\breve{D}} \log \left[1 + \frac{(\breve{\pi}_{z_i^{*t} z_j^{*t}} - \pi^*_{z_i^{*t} z_j^{*t}})(X_{ij}^t - \pi^*_{z_i^{*t} z_j^{*t}})}{\pi^*_{z_i^{*t} z_j^{*t}}(1 - \pi^*_{z_i^{*t} z_j^{*t}})} \right]. \tag{2.5.19}$$

Combining (2.5.16) and Lemma 2.5.7, we obtain

$$\begin{aligned} &\mathbb{P}^*_{\theta^*} \left[\left\{ \sum_{z^{1:T} \neq z^{*1:T}} \frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T} \right\} \cap \Omega_{n,T} \right] \\ &\leq \sum_{r=1}^{nT} \sum_{z^{1:T}; \|z^{1:T} - z^{*1:T}\|_0 = r} \mathbb{P}^*_{\theta^*} \left[\{U_1 + U_2 - U_3 > -\log(1/(\epsilon y_{n,T})) - 3r \log(nT)\} \cap \Omega_{n,T} \right]. \end{aligned} \tag{2.5.20}$$

We then decompose

$$\begin{aligned} &\mathbb{P}^*_{\theta^*} \left[\{U_1 + U_2 - U_3 > -\log(1/(\epsilon y_{n,T})) - 3r \log(nT)\} \cap \Omega_{n,T} \right] \\ \leq &\mathbb{P}^*_{\theta^*} \left[\{U_1 + U_2 - U_3 > -\log(1/(\epsilon y_{n,T})) - 3r \log(nT)\} \cap \Omega_{n,T} \cap \{|U_3| \leq r \log(nT)\} \right] \\ &+ \mathbb{P}^*_{\theta^*} \left[\Omega_{n,T} \cap \{|U_3| > r \log(nT)\} \right] \\ \leq &\mathbb{P}^*_{\theta^*} \left[\{U_1 + U_2 > -\log(1/(\epsilon y_{n,T})) - 4r \log(nT)\} \cap \Omega_{n,T} \right] \\ &+ \mathbb{P}^*_{\theta^*} \left[\Omega_{n,T} \cap \{|U_3| > r \log(nT)\} \right] \\ \leq &\mathbb{P}^*_{\theta^*} \left[U_1 > -\log(1/(\epsilon y_{n,T})) - 5r \log(nT) \right] + \mathbb{P}^*_{\theta^*} \left[\Omega_{n,T} \cap \{|U_2| > r \log(nT)\} \right] \\ &+ \mathbb{P}^*_{\theta^*} \left[\Omega_{n,T} \cap \{|U_3| > r \log(nT)\} \right]. \end{aligned} \tag{2.5.21}$$

We handle these three terms separately in the following. From now on, we consider a configuration $z^{1:T}$ such that $\|z^{1:T} - z^{*1:T}\|_0 = r = \sum_t r(t)$.

First term in the right-hand side of (2.5.21). Recall that U_1 is given by (2.5.17). We can further decompose this term

$$\begin{aligned} U_1 = &\sum_{(i,j,t)\in D^*} \left((X_{ij}^t - \pi^*_{z_i^{*t} z_j^{*t}}) \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} \frac{1-\pi^*_{z_i^{*t} z_j^{*t}}}{1-\pi^*_{z_i^t z_j^t}} \right) \\ &+ \sum_{(i,j,t)\in D^*} \left(\pi^*_{z_i^{*t} z_j^{*t}} \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} + (1-\pi^*_{z_i^{*t} z_j^{*t}}) \log \frac{1-\pi^*_{z_i^t z_j^t}}{1-\pi^*_{z_i^{*t} z_j^{*t}}} \right) \\ &+ \sum_{i=1}^n \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}}. \end{aligned}$$

For n and T large enough such that $\breve{\Gamma} \in [\delta, 1-\delta]^{Q^2}$ (implying for the corresponding stationary distribution $\breve{\alpha} \in [\delta, 1-\delta]^Q$), we have

$$\begin{aligned} &\sum_{i=1}^n \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \\ =&\sum_{i=1}^n \mathbb{1}_{\{z_i^1 \neq z_i^{*1}\}} \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \mathbb{1}_{\{(z_i^t, z_i^{t+1}) \neq (z_i^{*t}, z_i^{*t+1})\}} \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \\ \leq& r(1) \log \frac{1-\delta}{\delta} + \sum_{t=1}^{T-1} [r(t) + r(t+1)] \log \frac{1-\delta}{\delta} \leq 2r \log \frac{1-\delta}{\delta}. \end{aligned}$$

To handle the term U_1, we need to lower bound the cardinality of the set D^*. This is the purpose of Lemma 2.5.8 which is a generalization of Proposition B.4 in Celisse et al. (2012). This can be done for all the configurations $z^{1:T}$ and all the configurations $z^{*1:T}$ that belong to some $\Omega_\eta(\theta)$.

Lemma 2.5.8. *For any $\eta \in (0, \delta)$, any parameter $\theta \in \Theta$, any configuration $z^{1:T}$ and any $z^{*1:T} \in \Omega_\eta(\theta)$ such that $\|z^{1:T} - z^{*1:T}\|_0 = r$, we have*

$$\left| D_{n,T}(z^{1:T}, \pi) \right| \geq \frac{(\delta - \eta)^2}{4} nr.$$

Combining Lemma 2.5.8 with the previous bound, we get that

$$\begin{aligned} (|D^*|)^{-1} \left(\sum_{i=1}^n \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \right) &\leq \frac{2r}{|D^*|} \log \frac{1-\delta}{\delta} \\ &\leq \frac{8}{n(\delta-\eta)^2} \log \frac{1-\delta}{\delta} \xrightarrow[n\to+\infty]{} 0. \quad (2.5.22) \end{aligned}$$

We also have

$$(|D^*|)^{-1} \sum_{(i,j,t)\in D^*} \left(\pi^*_{z_i^t z_j^t} \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} + (1-\pi^*_{z_i^{*t} z_j^{*t}}) \log \frac{1-\pi^*_{z_i^t z_j^t}}{1-\pi^*_{z_i^{*t} z_j^{*t}}} \right)$$
$$\leq \max_{q,l,q',l';\pi^*_{ql}\neq\pi^*_{q'l'}} -k(\pi^*_{ql}, \pi^*_{q'l'})$$

with $k(x,y) = x\log(x/y) + (1-x)\log[(1-x)/(1-y)]$ for $(x,y) \in (0,1)^2$. The function k is positive for every (x,y) such that $x \neq y$, hence, introducing the notation $K^* = \min_{q,l,q',l';\pi^*_{ql}\neq\pi^*_{q'l'}} k(\pi^*_{ql}, \pi^*_{q'l'})/2$,

$$\max_{q,l,q',l';\pi^*_{ql}\neq\pi^*_{q'l'}} -k(\pi^*_{ql}, \pi^*_{q'l'}) := -2K^* < 0.$$

So, by (2.5.22), we have for n large enough

$$(|D^*|)^{-1} \Bigg\{ \sum_{(i,j,t)\in D^*} \left(\pi^*_{z_i^t z_j^t} \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} + (1-\pi^*_{z_i^{*t} z_j^{*t}}) \log \frac{1-\pi^*_{z_i^t z_j^t}}{1-\pi^*_{z_i^{*t} z_j^{*t}}} \right)$$
$$+ \sum_{i=1}^n \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \Bigg\} \leq -K^*.$$

This leads to

$$\mathbb{P}^*_{\theta^*}(U_1 > u) \leq \mathbb{P}^*_{\theta^*} \left[\sum_{(i,j,t)\in D^*} \left((X^t_{ij} - \pi^*_{z_i^{*t} z_j^{*t}}) \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} \frac{1-\pi^*_{z_i^{*t} z_j^{*t}}}{1-\pi^*_{z_i^t z_j^t}} \right) - |D^*|K^* > u \right]$$

for any $u > 0$ and large enough n. Moreover, thanks to Hoeffding's Inequality and Assumption 3,

$$\mathbb{P}^*_{\theta^*}(U_1 > u) \leq \mathbb{P}^*_{\theta^*} \left(\sum_{(i,j,t)\in D^*} \left((X^t_{ij} - \pi^*_{z_i^{*t} z_j^{*t}}) \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} \frac{1-\pi^*_{z_i^{*t} z_j^{*t}}}{1-\pi^*_{z_i^t z_j^t}} \right) > u + |D^*|K^* \right)$$
$$\leq \exp\left[-\frac{u^2 + |D^*|^2 K^{*2} + 2u|D^*|K^*}{|D^*|C_\zeta} \right]$$
$$\leq \exp\left[-\frac{|D^*|^2 K^{*2} + 2u|D^*|K^*}{|D^*|C_\zeta} \right] = \exp\left[-\frac{2uK^*}{C_\zeta} \right] \exp\left[-\frac{|D^*|K^{*2}}{C_\zeta} \right],$$

where C_ζ is a constant depending on ζ. Finally using Lemma 2.5.8, we have

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(U_1 > -\log(1/(\epsilon y_{n,T})) - 5r\log(nT)\right) \\
\leq &\exp\left[\left[\log(1/(\epsilon y_{n,T})) + 5r\log(nT)\right]\frac{2K^*}{C_\zeta}\right]\exp\left[-\frac{|D^*|K^{*2}}{C_\zeta}\right] \\
\leq &\exp\left[\left[\log(1/(\epsilon y_{n,T})) + 5r\log(nT)\right]\frac{2K^*}{C_\zeta}\right]\exp\left[-nr\frac{(\delta-\eta)^2K^{*2}}{4C_\zeta}\right].
\end{aligned}$$

Second term in the right-hand side of (2.5.21). We have

$$\begin{aligned}
U_2 &:= \sum_{(i,j,t)\in D^*\cup\breve{D}} \log\left[1 + \frac{(\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t})(X_{ij}^t - \pi^*_{z_i^t z_j^t})}{\pi^*_{z_i^t z_j^t}(1-\pi^*_{z_i^t z_j^t})}\right] \\
&\leq \sum_{(i,j,t)\in D^*\cup\breve{D}} \frac{(\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t})(X_{ij}^t - \pi^*_{z_i^t z_j^t})}{\pi^*_{z_i^t z_j^t}(1-\pi^*_{z_i^t z_j^t})}.
\end{aligned}$$

For any $q, l, q', l' \in [\![1, Q]\!]$, we introduce the sets

$$\begin{aligned}
F_{qlq'l'} &= F_{qlq'l'}(z^{1:T}, z^{*1:T}) := \{(i,j,t) \in I_{n,T}; z_i^t = q, z_j^t = l, z_i^{*t} = q', z_j^{*t} = l'\} \\
F_{ql} &= F_{ql}(z^{1:T}) := \cup_{1\leq q', l'\leq Q} F_{qlq'l'} = \{(i,j,t) \in I_{n,T}; z_i^t = q, z_j^t = l\} \\
G_{qlq'l'} &= G_{qlq'l'}(z^{1:T}, z^{*1:T}, \pi^*, \breve{\pi}) := (D^* \cup \breve{D}) \cap F_{qlq'l'} \\
&= \{(i,j,t) \in I_{n,T}; z_i^t = q, z_j^t = l, z_i^{*t} = q', z_j^{*t} = l' \\
&\qquad\qquad \text{and } (\pi^*_{z_i^t z_j^t} \neq \pi^*_{z_i^{*t} z_j^{*t}} \text{ or } \breve{\pi}_{z_i^t z_j^t} \neq \breve{\pi}_{z_i^{*t} z_j^{*t}})\} \\
G_{ql} &= G_{ql}(z^{1:T}, z^{*1:T}, \pi^*, \breve{\pi}) := (D^* \cup \breve{D}) \cap F_{ql} \\
&= \{(i,j,t) \in I_{n,T}; z_i^t = q, z_j^t = l \text{ and } (\pi^*_{z_i^t z_j^t} \neq \pi^*_{z_i^{*t} z_j^{*t}} \text{ or } \breve{\pi}_{z_i^t z_j^t} \neq \breve{\pi}_{z_i^{*t} z_j^{*t}})\}.
\end{aligned}$$

Then we bound

$$
\begin{aligned}
|U_2| &\leq \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{(i,j,t)\in D^*\cup \breve{D}} (X^t_{ij}-\pi^*_{ql}) \mathbb{1}_{z^t_i=q, z^t_j=l} \right| \\
&\leq \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{(i,j,t)\in G_{ql}} (X^t_{ij}-\pi^*_{ql}) \right| \\
&\leq \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{(i,j,t)\in G_{ql}} (X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j}) \right| \\
&\quad + \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{(i,j,t)\in G_{ql}} (\pi^*_{z^{*t}_i z^{*t}_j}-\pi^*_{ql}) \right| \\
&\leq \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{(i,j,t)\in G_{ql}} (X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j}) \right| \\
&\quad + \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{q',l'} (\pi^*_{q'l'}-\pi^*_{ql}) |G_{qlq'l'}| \right| . \qquad (2.5.23)
\end{aligned}
$$

For every $u>0$, we thus have

$$
\begin{aligned}
&\mathbb{P}^*_{\theta^*}(\Omega_{n,T}\cap\{|U_2|>u\}) \\
\leq &\mathbb{P}^*_{\theta^*}\left(\left\{ \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{(i,j,t)\in G_{ql}} (X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j}) \right| > u/2 \right\} \cap \Omega_{n,T} \right) \\
&+ \mathbb{P}^*_{\theta^*}\left(\left\{ \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})} \left| \sum_{1\leq q',l'\leq Q} (\pi^*_{q'l'}-\pi^*_{ql}) |G_{qlq'l'}| \right| > u/2 \right\} \cap \Omega_{n,T} \right) . \qquad (2.5.24)
\end{aligned}
$$

We start by dealing with the first term of the right-hand side of (2.5.24). Notice that on the event $\Omega_{n,T}$, we have $\left|(\breve{\pi}_{ql}-\pi^*_{ql})/(\pi^*_{ql}(1-\pi^*_{ql}))\right| \leq v_{n,T}/\zeta^2$ for every $q,l\in[\![1,Q]\!]$. The next lemma establishes that any set $D_{n,T}(z^{1:T},\pi)$ is included in a larger set, whose cardinality is bounded. In particular, the random set $\breve{D}$ is included in a larger deterministic subset.

Lemma 2.5.9. *Let $z^{1:T}$ and $z^{*1:T}$ denote two configurations such that $\|z^{1:T}-z^{*1:T}\|_0=r$. Then for any parameter $\pi=(\pi_{ql})_{1\leq q,l\leq Q}$, we have*

$$
D_{n,T}(z^{1:T},\pi)\subset D_{n,T}(z^{1:T}) := \left\{(i,j,t)\in[\![1,n]\!]^2\times[\![1,T]\!]; (z^t_i,z^t_j)\neq(z^{*t}_i,z^{*t}_j)\right\}
$$

and $\left|D_{n,T}(z^{1:T})\right| \leq 2nr.$

As the set G_{ql} is random (because $\breve{D}$ is random), we write

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\left\{\sum_{1\le q,l\le Q}\frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})}\left|\sum_{(i,j,t)\in G_{ql}}(X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j})\right|>u/2\right\}\cap\Omega_{n,T}\right)\\
\le&\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{(i,j,t)\in G_{ql}}(X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j})\right|>\frac{u\zeta^2}{2v_{n,T}}\right)\\
\le&\sum_{D\subset D_{n,T}(z^{1:T})}\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{(i,j,t)\in F_{ql}\cap D}(X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j})\right|>\frac{u\zeta^2}{2v_{n,T}}\right),
\end{aligned}$$

where now D is a deterministic set. By a union bound and Hoeffding's inequality, we have for any $D\subset D_{n,T}(z^{1:T})$

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{(i,j,t)\in F_{ql}\cap D}(X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j})\right|>\frac{u\zeta^2}{2v_{n,T}}\right)\\
\le&Q^2\max_{1\le q,l\le Q}\mathbb{P}^*_{\theta^*}\left(\left|\sum_{(i,j,t)\in F_{ql}\cap D}(X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j})\right|>\frac{u\zeta^2}{2v_{n,T}}\right)\le 2Q^2\exp\left(-\frac{2u^2\zeta^4}{4v^2_{n,T}Q^4}\frac{1}{|D|}\right).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\left\{\sum_{1\le q,l\le Q}\frac{\left|\breve{\pi}_{ql}-\pi^*_{ql}\right|}{\pi^*_{ql}(1-\pi^*_{ql})}\left|\sum_{(i,j,t)\in G_{ql}}(X^t_{ij}-\pi^*_{z^{*t}_i z^{*t}_j})\right|>u/2\right\}\cap\Omega_{n,T}\right)\\
\le&\sum_{D\subset D_{n,T}(z^{1:T})}2Q^2\exp\left(-\frac{2u^2\zeta^4}{4v^2_{n,T}Q^4}\frac{1}{|D|}\right)\le\sum_{k=1}^{2nr}\sum_{D\subset D_{n,T}(z^{1:T});|D|=k}2Q^2\exp\left(-\frac{2u^2\zeta^4}{4v^2_{n,T}Q^4}\frac{1}{k}\right)\\
\le&2Q^2\sum_{k=1}^{2nr}(2nr)^k\exp\left(-\frac{2u^2\zeta^4}{4v^2_{n,T}Q^4}\frac{1}{2nr}\right)\le 2Q^2\exp\left(-\frac{u^2\zeta^4}{4v^2_{n,T}Q^4nr}\right)(2nr)^{2nr+1}.
\end{aligned}$$

For the second term of (2.5.24), we get from a union bound and from Lemma 2.5.9 (that gives an upper bound for $|D^*\cup\breve{D}|$) that

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\left\{\sum_{1\le q,l\le Q}\left|\frac{(\breve{\pi}_{ql}-\pi^*_{ql})}{\pi^*_{ql}(1-\pi^*_{ql})}\right|\left|\sum_{1\le q',l'\le Q}(\pi^*_{q'l'}-\pi^*_{ql})|G_{qlq'l'}|\right|>u/2\right\}\cap\Omega_{n,T}\right)\\
\le&\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{1\le q',l'\le Q}(\pi^*_{q'l'}-\pi^*_{ql})|G_{qlq'l'}|\right|>\frac{u\zeta^2}{2v_{n,T}}\right)\\
\le&Q^2\max_{1\le q,l\le Q}\mathbb{P}^*_{\theta^*}\left(\left|\sum_{1\le q',l'\le Q}(\pi^*_{q'l'}-\pi^*_{ql})|G_{qlq'l'}|\right|>\frac{u\zeta^2}{2v_{n,T}Q^2}\right)\le Q^2\mathbb{P}^*_{\theta^*}\left(2nr>\frac{u\zeta^2}{2v_{n,T}Q^2}\right),
\end{aligned}$$

because $|\pi^*_{q'l'} - \pi^*_{ql}| \leq 1$, implying that

$$\left|\sum_{q',l'}(\pi^*_{q'l'} - \pi^*_{ql})|G_{qlq'l'}|\right| \leq \sum_{q',l'}|G_{qlq'l'}| = |G_{ql}| = |F_{ql} \cap (D^* \cup \breve{D})| \leq |D_{n,T}(z^{1:T})| \leq 2nr.$$

Finally, we have the following upper bound for the second term of (2.5.21)

$$\begin{aligned}\mathbb{P}^*_{\theta^*}\left(\Omega_{n,T} \cap \{|U_2| > r\log(nT)\}\right) \leq & 2Q^2 \exp\left(-\frac{r\zeta^4(\log(nT))^2}{4Q^4 v^2_{n,T} n}\right)(2nr)^{2nr+1} \\ & + Q^2 \mathbb{P}^*_{\theta^*}\left(v_{n,T} > \frac{\zeta^2\log(nT)}{4Q^2 n}\right).\end{aligned}$$

Third term in the right-hand side of (2.5.21). We want to bound (in probability) the last term U_3. Distinguishing between the cases where $X^t_{ij} = 0$ and $X^t_{ij} = 1$, we have

$$\begin{aligned}U_3 &:= \sum_{(i,j,t)\in D^*\cup\breve{D}} \log\left[1 + \frac{(\breve{\pi}_{z_i^{*t}z_j^{*t}} - \pi^*_{z_i^{*t}z_j^{*t}})(X^t_{ij} - \pi^*_{z_i^{*t}z_j^{*t}})}{\pi^*_{z_i^{*t}z_j^{*t}}(1 - \pi^*_{z_i^{*t}z_j^{*t}})}\right] \\ &= \sum_{(i,j,t)\in D^*\cup\breve{D}} \left((1 - X^t_{ij})\log\left[1 - \frac{(\breve{\pi}_{z_i^{*t}z_j^{*t}} - \pi^*_{z_i^{*t}z_j^{*t}})}{(1 - \pi^*_{z_i^{*t}z_j^{*t}})}\right]\right. \\ &\qquad \left. + X^t_{ij}\log\left[1 + \frac{(\breve{\pi}_{z_i^{*t}z_j^{*t}} - \pi^*_{z_i^{*t}z_j^{*t}})}{\pi^*_{z_i^{*t}z_j^{*t}}}\right]\right) \\ &= \sum_{1\leq q,l\leq Q}\sum_{(i,j,t)\in D^*\cup\breve{D}} \left((1 - X^t_{ij})\log\left[1 - \frac{(\breve{\pi}_{ql} - \pi^*_{ql})}{(1 - \pi^*_{ql})}\right]\right. \\ &\qquad \left. + X^t_{ij}\log\left[1 + \frac{(\breve{\pi}_{ql} - \pi^*_{ql})}{\pi^*_{ql}}\right]\right)\mathbb{1}_{z_i^{*t}=q, z_j^{*t}=l}.\end{aligned}$$

For any $(q,l) \in [\![1,Q]\!]^2$, we further introduce the sets

$$\begin{aligned}F^*_{ql} &= \cup_{1\leq q',l'\leq Q} F_{q'l'ql} = \{(i,j,t) \in I_{n,T}; z_i^{*t} = q, z_j^{*t} = l\} \\ G^*_{ql} &= \cup_{1\leq q',l'\leq Q} G_{q'l'ql} = (D^* \cup \breve{D}) \cap F^*_{ql} = \{(i,j,t) \in D^* \cup \breve{D}; z_i^{*t} = q, z_j^{*t} = l\}.\end{aligned}$$

Centering the X_{ij}^t (under the distribution $\mathbb{P}_{\theta^*}^*$), we get

$$\begin{aligned}
U_3 = &\sum_{1\le q,l\le Q}\sum_{(i,j,t)\in D^*\cup\breve{D}}\Bigg((\pi_{ql}^*-X_{ij}^t)\log\left[1-\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{(1-\pi_{ql}^*)}\right]\\
&\qquad+(X_{ij}^t-\pi_{ql}^*)\log\left[1+\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{\pi_{ql}^*}\right]\Bigg)\mathbb{1}_{z_i^{*t}=q,z_j^{*t}=l}\\
&+\sum_{1\le q,l\le Q}\sum_{(i,j,t)\in D^*\cup\breve{D}}\Bigg((1-\pi_{ql}^*)\log\left[1-\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{(1-\pi_{ql}^*)}\right]\\
&\qquad+\pi_{ql}^*\log\left[1+\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{\pi_{ql}^*}\right]\Bigg)\mathbb{1}_{z_i^{*t}=q,z_j^{*t}=l}\\
=&\sum_{1\le q,l\le Q}\left(\log\left[1+\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{\pi_{ql}^*}\right]-\log\left[1-\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{(1-\pi_{ql}^*)}\right]\right)\sum_{(i,j,t)\in G_{ql}^*}(X_{ij}^t-\pi_{ql}^*)\\
&+\sum_{1\le q,l\le Q}|G_{ql}^*|\left((1-\pi_{ql}^*)\log\left[1-\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{(1-\pi_{ql}^*)}\right]+\pi_{ql}^*\log\left[1+\frac{(\breve{\pi}_{ql}-\pi_{ql}^*)}{\pi_{ql}^*}\right]\right).
\end{aligned}$$

Then, on the event $\Omega_{n,T}$ and for n and T large enough such that $|(\breve{\pi}_{ql}-\pi_{ql}^*)/(1-\pi_{ql}^*)|\le 1/2$ and $|(\breve{\pi}_{ql}-\pi_{ql}^*)/\pi_{ql}^*|\le 1/2$ for every q and l, using the fact that $|\log(1+x)|\le 2|x|$ for $x\in[-1/2,1/2]$, we have

$$|U_3|\le 4\frac{v_{n,T}}{\zeta}\sum_{1\le q,l\le Q}\left|\sum_{(i,j,t)\in G_{ql}^*}(X_{ij}^t-\pi_{ql}^*)\right|+4\frac{v_{n,T}}{\zeta}\sum_{1\le q,l\le Q}|G_{ql}^*|.$$

Then, for every $u>0$,

$$\begin{aligned}
\mathbb{P}_{\theta^*}^*\left(\Omega_{n,T}\cap\{|U_3|>u\}\right)\le&\mathbb{P}_{\theta^*}^*\left(\sum_{1\le q,l\le Q}\left|\sum_{(i,j,t)\in G_{ql}^*}(X_{ij}^t-\pi_{ql}^*)\right|>\frac{u\zeta}{8v_{n,T}}\right)\\
&+\mathbb{P}_{\theta^*}^*\left(v_{n,T}\sum_{1\le q,l\le Q}|G_{ql}^*|>\frac{u\zeta}{8}\right).
\end{aligned}\tag{2.5.25}$$

For the first term of (2.5.25), using Hoeffding's inequality as before,

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\sum_{1\leq q,l\leq Q}\left|\sum_{(i,j,t)\in G^*_{ql}}(X^t_{ij}-\pi^*_{ql})\right|>\frac{u\zeta}{8v_{n,T}}\right)\\
\leq&\sum_{k=1}^{2nr}\sum_{D\subset D_{n,T}(z^{1:T});|D|=k}\mathbb{P}^*_{\theta^*}\left(\sum_{1\leq q,l\leq Q}\left|\sum_{(i,j,t)\in D\cap F^*_{ql}}(X^t_{ij}-\pi^*_{ql})\right|>\frac{u\zeta}{8v_{n,T}}\right)\\
\leq&2Q^2(2nr)^{2nr+1}\exp\left(-\frac{u^2\zeta^2}{8^2Q^4v^2_{n,T}nr}\right).
\end{aligned}$$

For the second term of (2.5.25), we use

$$\mathbb{P}^*_{\theta^*}\left(v_{n,T}\sum_{1\leq q,l\leq Q}|G^*_{ql}|>\frac{u\zeta}{8}\right)\leq\mathbb{P}^*_{\theta^*}\left(v_{n,T}>\frac{u\zeta}{16nr}\right).$$

Finally, we have the following upper bound for the third term of (2.5.21)

$$\begin{aligned}
\mathbb{P}^*_{\theta^*}\left(\Omega_{n,T}\cap\{|U_3|>r\log(nT)\}\right)\leq&2Q^2(2nr)^{2nr+1}\exp\left(-\frac{r(\log(nT))^2\zeta^2}{8^2Q^4v^2_{n,T}n}\right)\\
&+\mathbb{P}^*_{\theta^*}\left(v_{n,T}>\frac{\log(nT)\zeta}{16n}\right).
\end{aligned}$$

Combining the 3 bounds on the right-hand-side of (2.5.21).

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\{U_1+U_2-U_3>-\log(1/(\epsilon y_{n,T}))-3r\log(nT)\}\cap\Omega_{n,T}\right)\\
\leq&\exp\left[\left[\log(1/(\epsilon y_{n,T}))+5r\log(nT)\right]\frac{2K^*}{C_\zeta}\right]\exp\left[-nr\frac{(\delta-\eta)^2K^{*2}}{4C_\zeta}\right]\\
&+2Q^2(2nr)^{2nr+1}\exp\left[-\frac{r\zeta^4(\log(nT))^2}{4Q^4v^2_{n,T}n}\right]+Q^2\mathbb{P}^*_{\theta^*}\left(v_{n,T}>\frac{\zeta^2\log(nT)}{4Q^2n}\right)\\
&+2Q^2(2nr)^{2nr+1}\exp\left[-\frac{r(\log(nT))^2\zeta^2}{8^2Q^4v^2_{n,T}n}\right]+\mathbb{P}^*_{\theta^*}\left(v_{n,T}>\frac{\log(nT)\zeta}{16n}\right).
\end{aligned}$$

Now we choose the sequence $v_{n,T}$ such that $v_{n,T}=o(\sqrt{\log(nT)}/n)$ which is sufficient to imply that the quantities $\mathbb{P}^*_{\theta^*}\left(v_{n,T}>\zeta^2\log(nT)/(4Q^2n)\right)$ and $\mathbb{P}^*_{\theta^*}\left(v_{n,T}>\log(nT)\zeta/(16n)\right)$ vanish as n and T increase. For large enough values of n and T and with C_1, C_2, C_3, C_4

and κ positive constants only depending on Q, ζ and K^*, we then have

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\{U_1 + U_2 - U_3 > -\log(1/(\epsilon y_{n,T})) - 3r\log(nT)\} \cap \Omega_{n,T}\right) \\
\leq &\exp\left[[\log(1/(\epsilon y_{n,T})) + 5r\log(nT)]\,\frac{2K^*}{C_\zeta}\right]\exp\left[-nr\frac{(\delta-\eta)^2K^{*2}}{4C_\zeta}\right] \\
&+ 2Q^2(2nr)^{2nr+1}\exp\left[-\frac{r\zeta^4(\log(nT))^2}{4Q^4v_{n,T}^2 n}\right] + 2Q^2(2nr)^{2nr+1}\exp\left[-\frac{r(\log(nT))^2\zeta^2}{8^2Q^4v_{n,T}^2 n}\right] \\
\leq &\exp\left[-(\delta-\eta)^2C_1 nr + C_2\log(nT)r + C_4\log(1/(\epsilon y_{n,T}))\right] \\
&+ \kappa\exp\left[3nr\log(nT) - C_3\frac{(\log(nT))^2 r}{nv_{n,T}^2}\right]. \qquad (2.5.26)
\end{aligned}$$

Let us introduce

$$\begin{aligned}
u_{nT} &= \exp\left[-(\delta-\eta)^2C_1 n + C_2\log(nT) + C_4\log(1/(\epsilon y_{n,T}))\right] \\
w_{nT} &= \exp\left[-C_3\frac{(\log(nT))^2}{nv_{n,T}^2} + 3n\log(nT)\right].
\end{aligned}$$

Now we go back to (2.5.20). Noticing that the number of configurations $z^{1:T}$ such that $\|z^{1:T} - z^{*1:T}\|_0 = r$ is equal to $\binom{nT}{r}(Q-1)^r$, we have

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\left\{\frac{\mathbb{P}_{\check{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\check{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}\right\} \cap \Omega_{n,T}\right) \\
\leq &\sum_{r=1}^{nT}\binom{nT}{r}(Q-1)^r u_{nT}^r + \sum_{r=1}^{nT}\binom{nT}{r}(Q-1)^r\kappa w_{nT}^r \\
\leq &[1+Qu_{nT}]^{nT} - 1 + \kappa\left([1+Qw_{nT}]^{nT} - 1\right).
\end{aligned}$$

Finally, notice that as long as $\log T = o(n)$ and $\log(1/y_{n,T}) = o(n)$ (resp. as long as $v_{n,T} = o(\sqrt{\log(nT)}/n)$), we have nTu_{nT} (resp. nTw_{nT}) converges to 0. Then we obtain for some universal positive constant C and large enough n and T

$$\mathbb{P}^*_{\theta^*}\left(\left\{\frac{\mathbb{P}_{\check{\theta}}(Z^{1:T} \neq z^{*1:T} \mid X^{1:T})}{\mathbb{P}_{\check{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} > \epsilon y_{n,T}\right\} \cap \Omega_{n,T}\right) \leq CnT(u_{nT} + w_{nT}).$$

This leads directly to inequality (2.5.14). □

2.5.4 Proof of Theorem 2.3.3

We fix some $\sigma \in \mathfrak{S}_Q$ and study the convergence in $\mathbb{P}_{\theta^*}-$probability of $\hat{\gamma}_{\sigma(q)\sigma(l)}$ to γ^*_{ql} with $\hat{\Gamma}$ as defined by the fixed point equation (2.3.2), i.e.

$$\hat{\gamma}_{\sigma(q)\sigma(l)} = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat{\theta}_\sigma}\left(Z_i^t = q, Z_i^{t+1} = l \mid X^{1:T}\right)}{\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat{\theta}_\sigma}\left(Z_i^t = q \mid X^{1:T}\right)}.$$

First, let us denote

$$A_{q,l} = \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat{\theta}_\sigma}\left(Z_i^t = q, Z_i^{t+1} = l \mid X^{1:T}\right),$$
$$B_q = \frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat{\theta}_\sigma}\left(Z_i^t = q \mid X^{1:T}\right).$$

Then we can write the quantity at stake as

$$\hat{\gamma}_{\sigma(q)\sigma(l)} - \gamma^*_{ql} = \frac{A_{q,l}}{B_q} - \gamma^*_{ql} = \frac{A_{q,l} - \alpha^*_q\gamma^*_{ql}}{B_q} + \alpha^*_q\gamma^*_{ql}\left(\frac{1}{B_q} - \frac{1}{\alpha^*_q}\right)$$

to obtain the following upper bound on the probability of interest

$$\begin{aligned}\mathbb{P}_{\theta^*}\left(\left|\hat{\gamma}_{\sigma(q)\sigma(l)} - \gamma^*_{ql}\right| > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq & \mathbb{P}_{\theta^*}\left(\left|\frac{A_{q,l} - \alpha^*_q\gamma^*_{ql}}{B_q}\right| > \frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \\ & + \mathbb{P}_{\theta^*}\left(\alpha^*_q\gamma^*_{ql}\left|\frac{1}{B_q} - \frac{1}{\alpha^*_q}\right| > \frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right).\end{aligned} \tag{2.5.27}$$

First term of the right-hand side of (2.5.27). For the first term in (2.5.27), for any $0 < \lambda < \delta$ (implying $\lambda < \alpha_q^*$ for any $q \in [\![1, Q]\!]$),

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{A_{q,l} - \alpha_q^*\gamma_{ql}^*}{B_q}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
=&\mathbb{P}_{\theta^*}\left(\left|\frac{A_{q,l} - \alpha_q^*\gamma_{ql}^*}{B_q}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \,\middle|\, B_q \geq \alpha_q^* - \lambda\right)\mathbb{P}_{\theta^*}\left(B_q \geq \alpha_q^* - \lambda\right)\\
&+ \mathbb{P}_{\theta^*}\left(\left|\frac{A_{q,l} - \alpha_q^*\gamma_{ql}^*}{B_q}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \,\middle|\, B_q < \alpha_q^* - \lambda\right)\mathbb{P}_{\theta^*}\left(B_q < \alpha_q^* - \lambda\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\left|A_{q,l} - \alpha_q^*\gamma_{ql}^*\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}(\alpha_q^* - \lambda)\right) + \mathbb{P}_{\theta^*}\left(B_q < \alpha_q^* - \lambda\right). \qquad (2.5.28)
\end{aligned}$$

First, we upper bound the probability $\mathbb{P}_{\theta^*}\left(\left|A_{q,l} - \alpha_q^*\gamma_{ql}^*\right| > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)$ for any $\epsilon > 0$, using the following lemma.

Lemma 2.5.10. *If $\log(T) = o(n)$, for any $\epsilon > 0$, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity such that $r_{n,T} = o\left(\sqrt{nT/\log n}\right)$ and any $\eta \in (0, \delta)$, we have for any $\sigma \in \mathfrak{S}_Q$*

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat{\theta}_\sigma}\left(Z_i^t = q, Z_i^{t+1} = l \mid X^{1:T}\right) - \alpha_q^*\gamma_{ql}^*\right| > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1)
\end{aligned}$$

with $v_{n,T}$ a sequence decreasing to 0 such that $v_{n,T} = o\left(\sqrt{\log(nT)}/n\right)$.

Then, for the second term of (2.5.28), notice that $B_q = \sum_{l=1}^{Q} A_{q,l}$ and $\sum_{l=1}^{Q}\gamma_{ql}^* = 1$. We then have, if $\log(T) = o(n)$ and $v_{n,T} = o\left(\sqrt{\log(nT)}/n\right)$, using Lemma 2.5.10 again,

$$\begin{aligned}
\mathbb{P}_{\theta^*}\left(B_q < \alpha_q^* - \lambda\right) =&\mathbb{P}_{\theta^*}\left(B_q - \alpha_q^* < -\lambda\right) = \mathbb{P}_{\theta^*}\left(\sum_{l=1}^{Q}(A_{q,l} - \alpha_q^*\gamma_{ql}^*) < -\lambda\right)\\
&\leq \sum_{l=1}^{Q}\mathbb{P}_{\theta^*}\left(A_{q,l} - \alpha_q^*\gamma_{ql}^* < -\lambda/Q\right) \leq \sum_{l=1}^{Q}\mathbb{P}_{\theta^*}\left(\left|A_{q,l} - \alpha_q^*\gamma_{ql}^*\right| > \lambda/Q\right)\\
&\leq Q\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1).
\end{aligned}$$

Finally, for the first term of (2.5.27), if $y_{n,T}$ is such that $1/y_{n,T} = o\left(\sqrt{nT/\log(n)}\right)$, if $v_{n,T} = o\left(\sqrt{\log(nT)/n}\right)$ and as long as $\log(T) = o(n)$, we obtain

$$\mathbb{P}_{\theta^*}\left(\left|\frac{A_{q,l} - \alpha_q^*\gamma_{ql}^*}{B_q}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq (Q+1)\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1). \tag{2.5.29}$$

Second term of the right-hand side of (2.5.27). For the second term of (2.5.27), we split it on two complementary events as before. For any $0 < \lambda < \delta$, we have

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\alpha_q^*\gamma_{ql}^*\left|\frac{1}{B_q} - \frac{1}{\alpha_q^*}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \\
=&\mathbb{P}_{\theta^*}\left(\alpha_q^*\gamma_{ql}^*\left|\frac{1}{B_q} - \frac{1}{\alpha_q^*}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \,\middle|\, B_q \geq \alpha_q^* - \lambda\right)\mathbb{P}_{\theta^*}\left(B_q \geq \alpha_q^* - \lambda\right) \\
&+ \mathbb{P}_{\theta^*}\left(\alpha_q^*\gamma_{ql}^*\left|\frac{1}{B_q} - \frac{1}{\alpha_q^*}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \,\middle|\, B_q < \alpha_q^* - \lambda\right)\mathbb{P}_{\theta^*}\left(B_q < \alpha_q^* - \lambda\right) \\
\leq&\mathbb{P}_{\theta^*}\left(\alpha_q^*\gamma_{ql}^*\left|\frac{1}{B_q} - \frac{1}{\alpha_q^*}\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \,\middle|\, B_q \geq \alpha_q^* - \lambda\right)\mathbb{P}_{\theta^*}\left(B_q \geq \alpha_q^* - \lambda\right) \\
&+ \mathbb{P}_{\theta^*}\left(B_q < \alpha_q^* - \lambda\right).
\end{aligned} \tag{2.5.30}$$

We already gave an upper bound on the second term in the right-hand side of (2.5.30). Let us give one for the first term. Notice that as $\alpha_q^* \geq \delta$ and if $B_q \geq \alpha_q^* - \lambda \geq \delta - \lambda > 0$, we have by the mean value theorem

$$\left|\frac{1}{B_q} - \frac{1}{\alpha_q^*}\right| \leq \frac{1}{(\delta-\lambda)^2}\left|B_q - \alpha_q^*\right|.$$

We can then write for the first term in the right-hand side of (2.5.30), as long as $\log(T) = o(n)$, for $\{y_{n,T}\}_{n,T\geq 1}$ such that $1/y_{n,T} = o\left(\sqrt{nT/\log n}\right)$ and with $v_{n,T}$ such

that $v_{n,T} = o\left(\sqrt{\log(nT)}/n\right)$, still using Lemma 2.5.10

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\alpha_q^*\gamma_{ql}^*\left|\frac{1}{B_q}-\frac{1}{\alpha_q^*}\right| > \frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\,\middle|\, B_q \geq \alpha_q^* - \lambda\right)\mathbb{P}_{\theta^*}\left(B_q \geq \alpha_q^* - \lambda\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\left|B_q - \alpha_q^*\right| > \frac{(\delta-\lambda)^2\epsilon}{2\alpha_q^*\gamma_{ql}^*}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\left|\sum_{l=1}^{Q}(A_{q,l} - \alpha_q^*\gamma_{ql}^*)\right| > \frac{(\delta-\lambda)^2\epsilon}{2\alpha_q^*\gamma_{ql}^*}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\sum_{l=1}^{Q}\mathbb{P}_{\theta^*}\left(\left|A_{q,l} - \alpha_q^*\gamma_{ql}^*\right| > \frac{(\delta-\lambda)^2\epsilon}{2\alpha_q^*\gamma_{ql}^*Q}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq Q\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1).
\end{aligned}$$

We finally obtain for the second term of the right-hand side of (2.5.27)

$$\mathbb{P}_{\theta^*}\left(\alpha_q^*\gamma_{ql}^*\left|\frac{1}{B_q}-\frac{1}{\alpha_q^*}\right| > \frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq 2Q\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1). \quad (2.5.31)$$

We conclude the proof by summing the upper bounds obtained in (2.5.29) and (2.5.31)

$$\mathbb{P}_{\theta^*}\left(\left|\hat{\gamma}_{\sigma(q)\sigma}(l) - \gamma_{ql}^*\right| > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq (3Q+1)\mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1)$$

and by noticing that $\mathbb{P}_{\theta^*}(\|\hat{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_{n,T}\sqrt{\log n}/\sqrt{nT}) \leq \sum_{1\leq q,l\leq Q}\mathbb{P}_{\theta^*}(|\hat{\gamma}_{\sigma(q)\sigma(l)} - \gamma_{ql}^*| > \epsilon r_{n,T}\sqrt{\log n}/\sqrt{nT})$. □

2.5.5 Proof of Corollary 2.3.3

Denoting by $\sigma_{n,T}$ the permutation minimizing the distance between $\hat{\pi}$ (permuted) and π^* for every $(n,T) \in [\![1,n]\!] \times [\![1,T]\!]$, i.e. $\sigma_{n,T} = \operatorname{argmin}_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma - \pi^*\|_\infty$, we apply Theorem 2.3.3 to $\hat{\theta}_{\sigma_{n,T}}$ in order to get

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\,\hat{\Gamma}_\sigma - \Gamma^*\|_\infty > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\|\,\hat{\Gamma}_{\sigma_{n,T}} - \Gamma^*\|_\infty > \epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&Q^2(3Q+1)\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1) \xrightarrow[n,T\to\infty]{} 0.
\end{aligned}$$

□

2.5.6 Proof of Theorem 2.4.1

We use the following lemma, that states that the quantity we optimize in the VEM algorithm and the log-likelihood are asymptotically equivalent.

Lemma 2.5.11. *We have the following inequality $\mathbb{P}_{\theta^*}$-a.s.*

$$\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\mathcal{J}(\hat{\chi}(\theta),\theta)-\frac{2}{n(n-1)T}\ell(\theta)\right|\leq\frac{2\log(1/\delta)}{n-1}.$$

We have that for any $\epsilon>0$, for n and T large enough,

$$\begin{aligned}&\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\mathcal{J}(\hat{\chi}(\theta),\theta)-\frac{2}{n(n-1)T}\ell(\theta)\right|>\frac{\epsilon r_{n,T}}{\sqrt{n}}\right)\\ \leq&\mathbb{P}_{\theta^*}\left(\frac{2\log(1/\delta)}{n-1}>\frac{\epsilon r_{n,T}}{\sqrt{n}}\right)=0\end{aligned}$$

We then conclude by combining this result with Theorem 2.3.1. □

2.5.7 Proof of Corollary 2.4.1

This is a direct consequence of Theorem 2.4.1 and Lemma 2.5.6 applied with the functions $F_{n,T}=\frac{2}{n(n-1)T}\mathcal{J}(\hat{\chi}(\cdot),\cdot)$. □

2.5.8 Proof of Theorem 2.4.2

This proof is quite similar to that of Theorem 2.3.3. We fix some $\sigma\in\mathfrak{S}_Q$ and study the convergence in $\mathbb{P}_{\theta^*}-$probability of $\tilde{\gamma}_{\sigma(q)\sigma(l)}$ to γ^*_{ql} with $\tilde{\Gamma}$ as defined by the fixed point equation (2.4.1), i.e.

$$\tilde{\gamma}_{\sigma(q)\sigma(l)}=\frac{\sum_{i=1}^n\sum_{t=1}^{T-1}\hat{\eta}^t_{iql}(\tilde{\theta}_\sigma)}{\sum_{i=1}^n\sum_{t=1}^{T-1}\hat{\tau}^t_{iq}(\tilde{\theta}_\sigma)}.$$

First, let us denote

$$\begin{aligned}A_{q,l}&=\frac{1}{n(T-1)}\sum_{i=1}^n\sum_{t=1}^{T-1}\hat{\eta}^t_{iql}(\tilde{\theta}_\sigma)=\frac{1}{n(T-1)}\sum_{i=1}^n\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z^t_i=q,Z^{t+1}_i=l),\\ B_q&=\frac{1}{n(T-1)}\sum_{i=1}^n\sum_{t=1}^{T-1}\hat{\tau}^t_{iq}(\tilde{\theta}_\sigma)=\frac{1}{n(T-1)}\sum_{i=1}^n\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z^t_i=q).\end{aligned}$$

Then we can write the quantity at stake as

$$\tilde{\gamma}_{\sigma(q)\sigma(l)} - \gamma^*_{ql} = \frac{A_{q,l}}{B_q} - \gamma^*_{ql} = \frac{A_{q,l} - \alpha^*_q \gamma^*_{ql}}{B_q} + \alpha^*_q \gamma^*_{ql} \left(\frac{1}{B_q} - \frac{1}{\alpha^*_q} \right).$$

We follow the line of the proof of Theorem 2.3.3, using Lemma 2.5.12 below instead of Lemma 2.5.10 in order to obtain the result.

Lemma 2.5.12. *For any $\epsilon > 0$, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$ increasing to infinity such that $r_{n,T} = o\left(\sqrt{nT/\log n}\right)$ and any $\eta \in (0, \delta)$, we have for any $\sigma \in \mathfrak{S}_Q$*

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left| \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T-1} \mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t = q, Z_i^{t+1} = l) - \alpha^*_q \gamma^*_{ql} \right| > \epsilon r_{n,T} \frac{\sqrt{\log n}}{\sqrt{nT}} \right) \\
\leq & 2\mathbb{P}_{\theta^*}\left(\|\tilde{\pi}_\sigma - \pi^*\|_\infty > v_{n,T} \right) + o(1)
\end{aligned}$$

with $v_{n,T}$ a sequence decreasing to 0 such that $v_{n,T} = o(\sqrt{\log(nT)}/n)$.

□

Chapter 3

Estimation of parameters in a space-evolving graph model based on Markov random fields

3.1 Introduction

Markov random fields (MRFs) are widely used models for the study of spatial data, for example in image processing, statistical physics or epidemiology, as they are convenient and flexible. A MRF is based on a known graph on the considered locations (generally a lattice), which we refer to as *location graph*, and the value of the field at one location depends only on the neighbour locations. The distribution of the Markov random field is a Gibbs distribution (by the Hammersley-Clifford Theorem, see for example Besag (1974)). Besides, random graphs are a suitable tool to model and describe interactions in many kinds of datasets such as biological, ecological, social or transport networks. We will focus here on the case where we observe networks in multiple locations linked through a MRF, i.e. space-evolving networks, and are interested in the classification structure of the nodes of these graphs. We will therefore combine unobserved Markov random fields and graph models.

This is motivated by an application in ecology, considering we observe graphs of interactions of species (which we refer to as *species interaction graphs*) in different locations and we want to study the connection behaviour of this species and cluster them into groups. Rather than considering each graph separately, we want to make use of geographical information about species at different locations. Edges in ecological networks

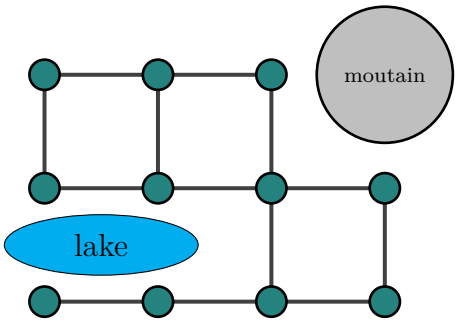


Fig. 3.1 Location graph for a species. This graph depends on the geography and environment. We assume here that the species is not present at high altitude (on the mountain) and cannot cross the lake.

can be of several types, such as predation, parasitism, mutualism[1] or commensalism[2]. See for example Delmas et al. (2019) for more information on the analysis of ecological networks. In the model we consider, we model the correlation over space using a known graph over the locations (that we will call *location graph*) that might be different for each species, and that is based on the environment. In such a graph, an edge exists between two locations if they are nearby and if there are no natural barrier between them that cannot be crossed by the species, such as mountains or rivers (see Figure 3.1). This is based on the idea that the species (which means a collection of individuals) moves freely between these two locations and thus tends to behave the same way (with respect to its interactions towards other species) at these two locations. The model is then flexible, allowing to represent the species heterogeneity regarding their movement behaviours between locations.

Random graphs have been intensively studied, and in particular several node clustering methods exist, allowing to describe the heterogeneous behaviour of nodes (and more precisely of groups of nodes) in a graph. We will focus in this work on the Stochastic Block Model (SBM, Holland et al., 1983) in which the nodes are partitioned into classes. In the SBM, class memberships of the nodes are represented by latent variables and the connection between two nodes is drawn from a distribution depending on the classes of these two nodes (a Bernoulli distribution in the case of binary graphs). Assuming that our observations are drawn from such a model, we can estimate the parameters and the groups of nodes that share the same connection behaviour. Another way to obtain a

[1] ecological interaction between two or more species where each species has a benefit, such as flowering plants being pollinated by animals

[2] interaction which is beneficial for a species and neutral for the other

clustering, and more specifically to recover communities[3], is to use spectral clustering methods, i.e. algorithms that cluster points using eigenvectors of matrices derived from the data (see for example Von Luxburg (2007) for more details on the spectral clustering, and Rohe et al. (2011) for a version of the spectral clustering for the recovery of a general clustering in the context of graphs and not only community detection). Methods based on the optimisation of a measure known as modularity have also been introduced (Newman and Girvan, 2004), this measure being proportional to the number of edges within groups minus the expected number in an equivalent network with random edges. Another way to identify communities is by running dynamical processes on the graph, such as random walks (Pons and Latapy, 2005), based on the idea that if the within-community edge density is high and the between-community one is low, a random walk will be trapped in each cluster for quite some time, before finding a way out of it.

When studying spatial data through MRF models, problems occur when performing maximum likelihood estimation. In particular, the complexity of the joint probability distribution leads to an intractable normalising constant (unless for very small graphs). Some methods have been introduced to circumvent the problems caused by the complexity and therefore intractability of the joint probability distribution in spatial data. Besag (1975) introduced the pseudolikelihood, approximating the joint probability distribution by the product of the conditional distributions at each location, given all the other locations (thus given the neighbours). An estimator of the distribution parameter can then be obtained by maximising the pseudolikelihood. The composite likelihood (Lindsay, 1988) extends the pseudolikelihood by approximating the joint distribution by the product of tractable joint distributions of variables of a small number of locations. See Varin et al. (2011) for a review of composite likelihood methods and Okabayashi et al. (2011) for the use of composite likelihood for the Potts model. Some Markov chain Monte Carlo methods have also been introduced in order to approximate the maximum likelihood estimator. For example, Younes (1988) proposes a stochastic gradient algorithm using a Gibbs sampler for the MCMC approximation of the gradient. Geyer and Thompson (1992) propose an algorithm to approximate the maximum likelihood based on a direct approximation of the likelihood from a MCMC sample and its maximisation. Some methods have also been introduced in order to calculate or approximate the normalising constant. Pettitt et al. (2003) and Friel and Pettitt (2004) compute the constant by wrapping the lattice on a cylinder. Reeves and Pettitt (2004) propose a recursive method based on an appropriate factorisation of the unnormalised joint probability (due to

[3]i.e. groups of vertices that have a high within-group connectivity and a low between-group connectivity

the Markov property) reducing the complexity of the summation and giving an exact computation of the normalising constant for lattices up to about 20 rows. Reduced dependence approximation (RDA) extends this forward recursion method in order to compute an approximation of the normalising constant on regular lattices of any size (Friel et al., 2009) or lattices with irregular boundaries (McGrory et al., 2012). In particular, such methods can be used in the task of inference. The normalising constant can also be approximated by using Markov Chain Monte Carlo (MCMC) samplers such as the Gibbs sampler, or using path sampling (Gelman and Meng, 1998).

Another type of methods, which will be our focus in this work, is the mean-field (Chandler, 1987) and mean-field like approximations, consisting of neglecting the fluctuations of the neighbours of the location we are considering by setting their values to their mean (for the mean-field approximation) or to another value, like the mode or a simulated value (for the mean-field like approximation). This results in a distribution that factorises over the locations, and solves the problem of the normalising constant. Note that in a Bayesian setting, the computation of the posterior distribution of the parameters of the MRF given the observations suffers from the same problems as for the maximum likelihood estimation. Some existing methods are for example auxiliary variable methods (Møller et al., 2006; Murray et al., 2006) or the use of composite likelihood (Friel, 2012; Stoehr and Friel, 2015). For a review on inference for discrete Markov random fields, see Stoehr (2017).

In the model we consider, the species are partitioned into a finite number of classes at each location, the classes following a Markov random field for each species, thus taking into account the spatial dependency of group memberships. But, as it is often the case, the groups are not observed, and we are then in the case of a hidden Markov random field (HMRF), where the observations (here the species interaction graphs at each location) depend on the latent Markov random field. Methods have been introduced to estimate the parameters and/or classification of a hidden Markov random field. In particular, this problem has been tackled in the field of image segmentation, i.e. when wanting to obtain a segmentation of a picture (Celeux et al., 2003; Forbes and Fort, 2007), of brain imaging (Zhang et al., 2001), and of gene clustering (Vignes and Forbes, 2009). Number of algorithms have been introduced for the study of HMRFs, namely methods based on Monte Carlo techniques, pseudolikelihood, the EM algorithm... Besag (1986) and then Lakshmanan and Derin (1989) (more generally, see Qian and Titterington (1991) describing Restoration-Maximization algorithms) propose iterative algorithms that consist of two steps. The first is a step of assignment in which we find the configuration maximising the probability given the observations and the current

parameter, using respectively an ICM (Iterated Conditional Modes) algorithm[4] and a simulated annealing. The second step consists in estimating the parameter maximising the complete likelihood, by using for example the pseudolikelihood as in Besag (1986). Chalmond (1989) introduced the Gibbsian EM, a variation of the EM algorithm based on the pseudolikelihood and where the conditional expectations are approximated by using a Monte Carlo method based on the Gibbs sampler. Pieczynski (1992) (see also Pieczynski (1994)) proposed a method for parameter estimation with latent variables called ICE (Iterative Conditional Expectation) which, assuming that we can estimate the parameters given the complete variable (observed and latent), iteratively approximates the expectation of this estimator given the observations (based on simulations from the conditional distribution of the latent variables given the observations).

In this work, we are interested in the estimation of parameters, and we follow Celeux et al. (2003) who propose EM-like algorithms for image segmentation (that have been illustrated in Celeux et al. (2002)). They use mean-field like approximations, generalising the mean field EM by Zhang (1992). They introduce in particular the simulated field EM algorithm (which we will also refer to as simulated EM), in which at every iteration of the algorithm, a configuration of the hidden Markov field given the observations is simulated with the current parameter, this configuration being used for the mean-field like approximation. Other possibilities are the mean field algorithm and mode field algorithm, using respectively the mean field estimate and the mode field estimate of the conditional distribution. In this work, we propose an algorithm based on the simulated EM, adapted to the observation of species interaction graphs generated under a SBM at each location. The simulated EM has been used for example in Vignes and Forbes (2009) to cluster genes, with Gaussian observations. In Vignes and Forbes (2009), the parameters are estimated using a simulated EM and the class memberships are then estimated using a Maximum Posterior Marginal principle[5]. More recently, Forbes and Fort (2007) introduced the Monte Carlo VEM algorithm, that combines the mean-field like approach and Monte Carlo techniques, and proved its convergence. Ranalli et al. (2018) and Lai and Lim (2015) use methods based on the composite likelihood. In particular, Ranalli et al. (2018) propose an EM algorithm based on the complete composite likelihood. Note that as before, in a Bayesian setting for an HMRF, the computation of the posterior distribution of the parameters given the observations suffers from the same problems as for the maximum likelihood estimation. Some examples of Bayesian estimation in the context of HMRF are Friel et al. (2009); McGrory et al. (2009) or Everitt (2012).

[4] corresponding to a simulated annealing with a temperature equel to 0

[5] consisting in assigning a gene to the class that maximises the probability of the membership to this class given the observations, under the estimated parameter

This article is organised as follows. Section 3.2 introduces our model and notation. More precisely, Section 3.2.1 describes the spatial graph model and notations, Section 3.2.2 gives the assumptions we make on the model parameters and Section 3.3 establishes the generic identifiability of the model under certain conditions. Section 3.4 describes the parameter estimation method, that is based on the simulated EM. Finally, Section 3.5 presents some results on simulations.

3.2 Model and notation

3.2.1 Definition of the model

We consider n species observed at L locations. At each location l, we observe interactions between the species, represented by an undirected binary graph with no self-loops, with adjacency matrix $X^l = \{X^l_{ij}\}_{1\leq i,j\leq n}$ such that for every nodes $1 \leq i, j \leq n$, we have $X^l_{ii} = 0$ and $X^l_{ij} = X^l_{ji}$. The case of a set of directed graphs, with or without self-loops, may be handled similarly. The species are divided into $Q \geq 2$ groups at each location, represented by the latent variables $(Z_1, \ldots, Z_n)$ with $Z_i = \{Z^l_i\}_{1\leq l\leq L} \in \{1, \ldots, Q\}^L$. We will also denote $Z^l = (Z^l_1, \ldots, Z^l_n)$ and $Z^{1:L} = (Z^1, \ldots, Z^L) = (Z^l_i)_{1\leq l\leq L, 1\leq i\leq n}$. All the Z_is are independent and each one follows a Gibbs distribution (more precisely we consider a Potts model), given by the known location graph $\mathcal{G}_i = (\{1, \ldots, L\}; E_i)$ with E_i the set of edges. We allow the species to have different location graphs in order to model species with different movement behaviours between the observed locations. Figure 3.2 gives a graphical representation of the model.

We have for every species i according to the Potts model, denoting $\psi_i = (\alpha_i, \beta_i)$,

$$\begin{aligned}\mathbb{P}_{\psi_i}(Z_i) =& \mathbb{P}_{\psi_i}(Z^1_i, \ldots, Z^L_i) = \frac{1}{S_i(\alpha_i, \beta_i)} \exp\left[\sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{Z^l_i=q} + \beta_i \sum_{(l,l')\in E_i} \mathbb{1}_{Z^l_i=Z^{l'}_i}\right] \\ :=& \frac{1}{S_i(\alpha_i, \beta_i)} \exp\left[-H_i(Z_i)\right] \end{aligned} \tag{3.2.1}$$

with $S_i(\alpha_i, \beta_i)$ the normalising constant, $\alpha_i = (\alpha_{iq})_{1\leq q\leq Q}$, where $\alpha_{iq}, \beta_i \in \mathbb{R}$, and defining H_i the energy function

$$H_i(z_i) = H_i(z^1_i, \ldots, z^L_i) = \sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{z^l_i=q} + \beta_i \sum_{(l,l')\in E_i} \mathbb{1}_{z^l_i=z^{l'}_i}.$$

The parameter $\alpha = (\alpha_{iq})_{1\leq q\leq Q}$ is the parameter of the external field, i.e. species i is more likely to be in groups associated with large values of $\alpha_i = (\alpha_{iq})_{1\leq q\leq Q}$. Note

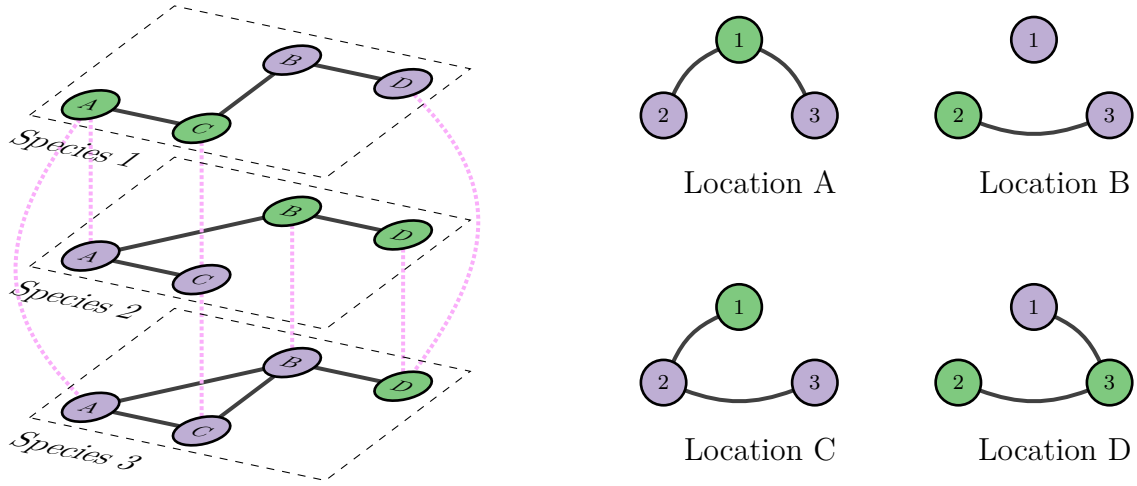


Fig. 3.2 Example of representation of the model for $Q = 2$, $L = 4$ and $n = 3$. On the left, the three layers represent the graphs on locations of the three species (i.e. the three location graphs), and for each location (A, B, C and D), the dotted edges between layers represent the connection between species at that location. On the right are the observed interaction graphs between the 3 species at each of the four locations. Note that the representation on the right only contains the interactions graphs at each location, and not the space dependency between these locations. The Q classes are represented by colors on the nodes (green and purple).

that if $\alpha_{iq} = 0$ for all $q \in [\![1, Q]\!]$, all the groups are equally probable (a priori). The parameter β_i determines the influence on the group of species i at a given location of the groups of the same species at neighbour locations (or the strength of interaction between neighbour locations). This means that the group membership of a species at a location and therefore its behaviour towards the other ones is influenced by the group membership of this species at the neighbour locations. More precisely, the membership to a group for a species i at location l is dependent on whether or not the species is in that same group at a neighbour location in the location graph $\mathcal{G}_i$. If β_i is positive, the model encourages neighbours to be in the same group, and on the contrary, a negative β_i encourages neighbours to be in different groups.

We write $S_i(\alpha_i, \beta_i)$ as follows

$$S_i(\alpha_i, \beta_i) = \sum_{z_i^{1:L}} \exp\left[\sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{z_i^l = q} + \beta_i \sum_{(l,l') \in E_i} \mathbb{1}_{z_i^l = z_i^{l'}}\right].$$

Note that the normalising constant is intractable in general because it requires a summation over all the Q^{nL} configurations. We can write the distribution of $Z^{1:L}$ as follows,

the Z^l being independent,

$$\mathbb{P}_\psi(Z^{1:L}) = \prod_{i=1}^{n} \frac{1}{S_i(\alpha_i, \beta_i)} \exp\left[-H_i(Z_i)\right] = \frac{1}{\prod_{i=1}^{n} S_i(\alpha_i, \beta_i)} \exp\left[-\sum_{i=1}^{n} H_i(Z_i)\right]$$

where $\psi = (\psi_i)_{1 \leq i \leq n}$.

Each adjacency matrix X^l then follows a stochastic block model so that, conditional on the latent groups $\{Z_i^l\}_{1 \leq i \ \leq n}$, the $\{X_{ij}^l\}_{1 \leq i < j \leq n}$ are independent Bernoulli random variables

$$X_{ij}^l \mid Z_i^l = q, Z_j^l = q' \sim \mathcal{B}(\pi_{qq'}^l), \quad \text{i.e.} \quad \mathbb{P}_\pi(X^l \mid Z^l) = \prod_{1 \leq i < j \leq n} (\pi_{Z_i^l Z_j^l}^l)^{X_{ij}^l} (1 - \pi_{Z_i^l Z_j^l}^l)^{1 - X_{ij}^l}, \tag{3.2.2}$$

and $\pi_{qq'}^l \in [0,1]$ are the connectivity parameters, satisfying $\pi_{qq'}^l = \pi_{q'q}^l$ for every $1 \leq q, q' \leq Q$. The parameter of the model is $\theta = (\psi, \pi) = (\alpha, \beta, \pi)$ with

- $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{nQ}$ with $\alpha_i = (\alpha_{iq})_{1 \leq q \leq Q}$ the parameters of the external fields,
- $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$ the interaction parameters,
- $\pi = (\pi_{qq'}^l)_{1 \leq q, q' \leq Q, 1 \leq l \leq L} \in [0,1]^{LQ(Q+1)/2}$ the symmetric matrices of connection probabilities.

We will denote in the following $z_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ for any $i \in [\![1, n]\!]$, i.e. the latent variables for every node except i, and identically we will denote $z^{-l} = (z^1, \ldots, z^{l-1}, z^{l+1}, \ldots, z^L)$ for any $l \in [\![1, L]\!]$, i.e. the latent variables at every location except l. Let us also denote $Z_{iq}^l = \mathbb{1}_{Z_i^l = q}$ for every l, i and q, and $\mathbf{Z}_i^l = (Z_{iq}^l)_{1 \leq q \leq Q} \in \{0,1\}^Q$.

3.2.2 Assumptions

We impose some constraints on these model parameters for identifiability purposes (see Section 3.3). The assumptions we make are the following.

1. For every $i \in [\![1, n]\!]$, the parameter α_i satisfies $\sum_{q=1}^{Q} \alpha_{iq} = 0$.
2. For every $q \in [\![1, Q]\!]$, for any $l, l' \in [\![1, L]\!]$, we have $\pi_{qq}^l = \pi_{qq}^{l'} \coloneqq \pi_{qq}$.
3. For every $l \in [\![1, L]\!]$, the values $\{\pi_{qq'}^l\}_{1 \leq q \leq q' \leq Q}$ are $Q(Q+1)/2$ distinct values.

We make Assumption 1 in order for the Potts model to be identifiable. Indeed, adding the same constant to all the α_{iq} for any $i \in [\![1, n]\!]$ leads to the same distribution. Note that when considering the estimation of the parameters (in Section 3.4), we will equivalently impose that $\alpha_{i1} = 0$ for every $i \in [\![1, n]\!]$. Assumptions 2 and 3 are sufficient for the identifiability of the model, to avoid label switching issues between the different locations. We denote by Θ the set of parameters satisfying these constraints.

3.3 Identifiability

In this section, we establish the generic identifiability of our model under the assumptions of Section 3.2.2. We first state the identifiability of the Potts model, for any $i \in [\![1, n]\!]$.

Lemma 3.3.1. *For each $i \in [\![1, n]\!]$, the parameters $\psi_i = (\alpha_i, \beta_i)$ of the Potts model as defined in (3.2.1) satisfying Assumption 1 are identifiable from the distribution of $Z_i^{1:L}$, as long as there is at least one edge in the location graph $\mathcal{G}_i$ (implying that $L \geq 2$).*

We believe that this result is known, but did not find a written proof of it. We then give the proof of Lemma 3.3.1 in Appendix B.1 for completeness. Then, for the identifiability of our model, we follow the proof of Theorem 2 in Allman et al. (2011) (see also Theorem 2 in Becker and Holzmann (2018)) based on Kruskal's theorem (see for example Theorem 16 in Allman et al. (2011)) to prove the generic identifiability of our model, meaning that the nonidentifiable parameters form a set of Lebesgue measure zero. Note that we obtain this identifiability up to label permutation of the groups (as in any latent groups model), meaning that we can only recover θ_σ with σ in $\mathfrak{S}_Q$ the set of permutations on $[\![1, Q]\!]$, where

$$\theta_\sigma = (\alpha_\sigma, \beta, \pi_\sigma) = \left((\alpha_{i\sigma(q)})_{1\leq i\leq n, 1\leq q\leq Q}, (\beta_i)_{1\leq i\leq n}, (\pi_{\sigma(q)\sigma(q')})_{1\leq q,q'\leq Q}\right).$$

Note that the permutation does not affect β, at it applies only on the latent groups.

Proposition 3.3.1. *Under Assumptions 1 to 3, the parameter $\theta = (\alpha, \beta, \pi)$ is generically identifiable up to label permutation from the distribution of $X^{1:L}$, if there is at least one edge in each location graph and $n \geq m^2$, where*

$$\begin{cases} m \geq Q - 1 + \left(\frac{Q+2}{2}\right)^2 & \text{if } Q \text{ is even,} \\ m \geq Q - 1 + \frac{(Q+1)(Q+3)}{4} & \text{if } Q \text{ is odd.} \end{cases}$$

Remark 3.3.1. Following the proof of Theorem 5 in Allman et al. (2009) would allow to get a better result for the case $Q = 2$, obtaining the identifiability for $n \geq 16$ instead of $n \geq 25$. We will not give more details about that.

Note that as in Allman et al. (2011) the generic aspect concerns only the part of the parameter space with the connectivity parameter, i.e. the $\pi_{qq'}$s. This means that α and β are identifiable (not only generically) from the distribution of $X^{1:L}$. In particular, a Potts model without external field (i.e. with $\alpha_{iq} = 0$ for every i and q) is still generically identifiable (even though this restriction on the parameter reduces the parameter space to a subspace of smaller dimension).

Proof of Proposition 3.3.1. Following the lines of the proof of Theorem 2 in Allman et al. (2011), for any number m of species, we denote by A^l the $Q^m \times 2^{\binom{m}{2}}$ matrix of conditional probabilities of observing all possible species interaction graph configurations $X^l \in \{0,1\}^{\binom{m}{2}}$, conditioned on node states $Z^l = (Z_1^l, \ldots, Z_m^l) \in \{1, \ldots, Q\}^m$ at location l for each $l \in [\![1, L]\!]$. We then write, denoting by $\otimes$ the Kronecker product,

$$A = A^1 \otimes A^2 \otimes \ldots \otimes A^L, \tag{3.3.1}$$

this $Q^{mL} \times 2^{L\binom{m}{2}}$ matrix being the matrix of conditional probabilities of observing all possible sets of species interaction graphs $X^{1:L}$, conditioned on node states at all locations $Z^{1:L}$. It is established in the proof of Theorem 2 in Allman et al. (2011) that each matrix A^l has generically full row rank as soon as

$$\begin{cases} m \geq Q - 1 + \left(\frac{Q+2}{2}\right)^2 & \text{if } Q \text{ is even,} \\ m \geq Q - 1 + \frac{(Q+1)(Q+3)}{4} & \text{if } Q \text{ is odd.} \end{cases}$$

Then from its definition 3.3.1, A has also generically full row rank. We will then conclude the proof thanks to a lemma similar to Lemma 16 in Allman et al. (2009) that we write below. Let us define before K_n the graph on the nL nodes $\{(l,i)\}_{1 \leq i \leq n, 1 \leq l \leq L}$ that is the union of the L complete graphs at each location, i.e. such that there is an edge between (l,i) and (l',i') if and only if $l = l'$ (and $i \neq i'$ so that there is no self loop).

Lemma 3.3.2. *If the $Q^{mL} \times 2^{L\binom{m}{2}}$ matrix A defined in (3.3.1) has rank Q^{mL}, then with $n = m^2$ there exist pairwise edge-disjoint subgraphs G_1, G_2, G_3 of K_n such that for each G_k ($1 \leq k \leq 3$), the matrix of probabilities B_k of observing subgraphs of G_k conditioned on node state assignments has rank Q^{nL}.*

Proof of Lemma 3.3.2. The proof of that lemma is similar to that of Lemma 16 from Allman et al. (2009), except for the construction of the G_ks. We do the same partition

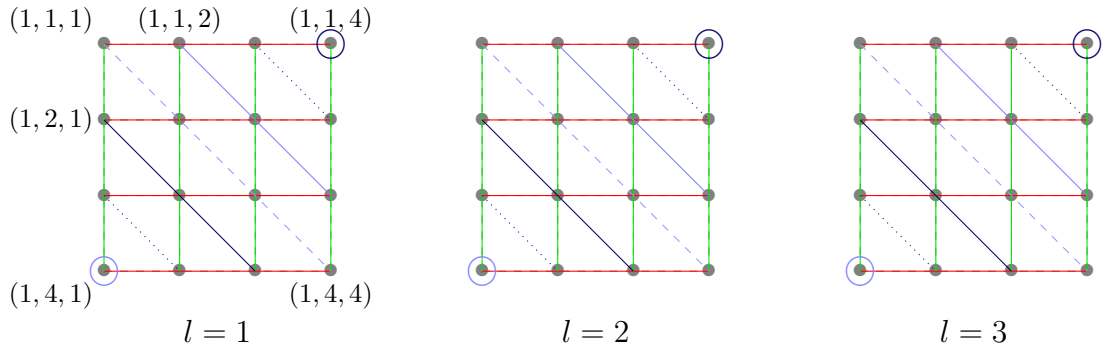


Fig. 3.3 Partitions for $m = 4$ and $L = 3$. In red, the $mL = 12$ sets of the partition leading to G_1, in green the mL sets of the partition leading to G_2, and in blue, the mL sets of the partition leading to G_3. More precisely, the sets of the partition leading to G_3 are for each l $\{(l,1,1),(l,2,2),(l,3,3),(l,4,4)\}$, $\{(l,1,2),(l,2,3),(l,3,4),(l,4,1)\}$, $\{(l,1,3),(l,2,4),(l,3,1),(l,4,2)\}$, $\{(l,1,4),(l,2,1),(l,3,2),(l,4,3)\}$.

as in Allman et al. (2009) separately for each l, i.e. we picture the nodes as lattice points in L square grids and take as the partition leading to G_1 the rows of the grids, as the partition leading to G_2 the columns of the grids and as the partition leading to G_3 the diagonals of the grids, each G_k then being the union of mL complete subgraphs. Explicitly, we label the nodes by $(l,i,j) \in \{1,\dots,L\} \times \{1,\dots,m\} \times \{1,\dots,m\}$ such that $\{(l,i,j)\}_{1\leq i,j\leq m}$ is the set of $n = m^2$ nodes at location l for any $l \in L$ (and two nodes (l,i,j) and (l',i,j) corresponding to the same species at locations l and l'). We can then write $P_k = \{V^k_{l,j}; l \in \{1,\dots,L\}, j \in \{1,\dots,m\}\}$ denoting the partition of the node set leading to G_k with

$$\begin{aligned} V^1_{l,j} &= \{(l,j,i); i \in \{1,\dots,m\}\}, \\ V^2_{l,j} &= \{(l,i,j); i \in \{1,\dots,m\}\}, \\ V^3_{l,j} &= \{(l,i,i+j \mod m); i \in \{1,\dots,m\}\} \end{aligned}$$

and each G_k is the union over $(l,j) \in \{1,\dots,L\} \times \{1,\dots,m\}$ of the complete graphs on node set $V^k_{l,j}$. By construction, the G_ks have no edge in common, and are thus independent given $Z^{1:L}$. See example in Figure 3.3.

Then, up to a reordering of the rows and columns, the matrix B_k of conditional probabilities of observing all possible subgraphs of G_k conditioned on node states can be written as the m-th Kronecker power of A

$$B_k = A^{\otimes m} = (A^1 \otimes A^2 \otimes \dots \otimes A^L)^{\otimes m}. \tag{3.3.2}$$

We already stated that A has generically full row rank under our assumptions, and then so has B_k. This concludes the proof of Lemma 3.3.2. □

Going back to the proof of Proposition 3.3.1 and as in Allman et al. (2011), we can apply Kruskal's theorem that gives that the set $(v := \mathbb{P}_\theta(Z^{1:L} = \cdot), \{B_k = \mathbb{P}_\theta(G_k = \cdot \mid Z^{1:L} = \cdot)\}_{1\leq k\leq 3})$ is generically identifiable up to label swapping, from the distribution of the random variables G_1, G_2, G_3 and thus from that of $X^{1:L}$ (for n large enough with respect to Q). To conclude, we need to identify the parameters (those of the Potts models, i.e. $(\alpha_{iq})_{1\leq i\leq n, 1\leq q\leq Q}$ and $(\beta_i)_{1\leq i\leq n}$, and the connection probability matrices $(\pi^l_{qq'})_{1\leq l\leq L, 1\leq q,q'\leq Q}$). Recall that we assume that $\forall q \in [\![1, Q]\!]$, the within-group connection probability is the same for every location l, i.e. $\pi^l_{qq} = \pi^{l'}_{qq} := \pi_{qq}$ for any $l, l' \in [\![1, L]\!]$, and also that the values $\{\pi_{qq}\}_{1\leq q\leq Q}$ are all distinct. We also assume that for any $l \in [\![1, L]\!]$ the $(\pi^l_{qq'})_{1\leq q\leq q'\leq Q}$ are all distinct. To identify the parameters, we rely on the same technique as in the proof of Theorem 2 in Becker and Holzmann (2018) and on the conclusion of the proof of Theorem 2 in Allman et al. (2011). We focus on the matrix B_1, and doing marginalisations as in those proofs for every edge in G_1, we can identify the values $(\pi_{qq})_{1\leq q\leq Q}$ (choosing an arbitrary labeling) and then $(\pi^l_{qq'})_{1\leq q\neq q'\leq Q}$.

Having identified all the connection probabilities, we can isolate the entries of v corresponding to the configurations such that $Z_i^{1:L}$ is fixed (equal to some $c \in [\![1, Q]\!]^L$). We describe how we identify those entries. To find the rows corresponding to configurations in which $Z^l_i = q$ for some $q \in [\![1, Q]\!]$, $i \in [\![1, n]\!]$ and $l \in [\![1, L]\!]$, we choose some $j, k \in [\![1, n]\!]$ such that i, j, and k are in the same set in the partition used to construct G_1 and we sum the columns for which $X^l_{ij} = 1$, those for which $X^l_{ik} = 1$, and those for which $X^l_{jk} = 1$. Doing so, we obtain the values of $\pi^l_{Z^l_i Z^l_j}$, $\pi^l_{Z^l_i Z^l_k}$ and $\pi^l_{Z^l_j Z^l_k}$ and can therefore identify for every row of B_1 the values of Z^l_i, Z^l_j and Z^l_k thanks to the assumption that $(\pi^l_{qq'})_{1\leq q,q'\leq Q}$ are distinct values and to the fact that we identified these quantities earlier. Doing this for all $i \in [\![1, n]\!]$ and $l \in [\![1, L]\!]$, we can obtain the entry in v corresponding to any configuration of the nL nodes. Summing the entries corresponding to the configurations such that $Z_i^{1:L} = c$, we obtain the probability $\mathbb{P}_\theta(Z_i^{1:L} = c)$. Doing this for every possible c and for every i, we have the probability distribution of each Markov field. Now using Lemma 3.3.1, stating that we can identify the parameters of the Potts model for a species i from the distribution of $Z_i^{1:L}$, allows us to recover the α_{iq}s and β_is and to conclude the proof. □

Particular case of the affiliation model: a counter-example The constraint imposed on each π^l in Assumption 3 (that its values $(\pi^l_{qq'})_{1\leq q\leq q'\leq Q}$ are distinct) implies that the result of Proposition 3.3.1 does not apply to the particular case of the affiliation

model for the conditional distribution of the graphs X^l. Indeed, in the affiliation model, the parameter π is determined by the values $\pi^l_{qq} = \pi^l_{\text{in}}$ for every $q \in [\![1, Q]\!]$ (within-group connection probability) and $\pi^l_{qq'} = \pi^l_{\text{out}}$ for every $q \neq q' \in [\![1, Q]\!]$ (between-groups connection probability) at each location l.

The identifiability result of Proposition 3.3.1 actually does not hold in general for the affiliation model. Indeed we exhibit in the proposition below a particular example of an affiliation model (satisfying Assumptions 1 and 2 but not Assumption 3) without external field[6] whose parameters are not identifiable. Note that this is a partial result, as the model might be identifiable in the affiliation case at the cost of some additional assumptions. The exhibition of such counter-example allows us to have some insights about the constraints we could impose in order to obtain identifiability.

Proposition 3.3.2. *Assume that all the species have the same location graphs $\mathcal{G}_i = \mathcal{G} = (V, E)$ with no cycle of odd length and the same parameter $\beta_i = \beta$, and that there is no external field, i.e. $\alpha_{iq} = 0$ for every $i \in [\![1, n]\!]$ and $q \in [\![1, Q]\!]$. Assume also that $Q = 2$ and that $\pi^l_{qq} = \pi_{\text{in}}$ (so that Assumption 2 is satisfied) and that $\pi^l_{qq'} = \pi^l_{\text{out}}$ for any $q \neq q' \in [\![1, Q]\!]$ and $l \in [\![1, L]\!]$. Then, the parameter $\theta = (\beta, \pi)$ of the model cannot be identified.*

This result holds in particular for first order lattices as location graphs, such lattices containing only cycles of even length.

Proof of Proposition 3.3.2. We are going to show that for any parameter $\theta = (\beta, \pi)$ satisfying the assumptions in the proposition, the parameter $\tilde{\theta} = (\tilde{\beta}, \tilde{\pi})$, defined such that $\tilde{\beta} = -\beta$ and $\tilde{\pi} = \pi$, leads to the same distribution as θ for $X^{1:L}$. For any possible configuration $z^{1:L}$, we construct a transformed version $\tilde{z}^{1:L}$ of $z^{1:L}$ as follows. Let us choose a starting location $l \in [\![1, L]\!]$ (for example location 1 in Figure 3.4) and set $\tilde{z}^l = z^l$ (i.e. $\tilde{z}^l_i = z^l_i$ for every $i \in [\![1, n]\!]$). Then for every neighbour location l' of l (in our case, locations 2 and 3), let us set $\tilde{z}^{l'}_i = 1$ if $z^{l'}_i = 2$ and $\tilde{z}^{l'}_i = 2$ if $z^{l'}_i = 1$ for every $i \in [\![1, n]\!]$[7]. Then for any neighbour location l'' of each l' that has not been visited yet (in our case, locations 4,5 and 6), let us set $\tilde{z}^{l''} = z^{l''}$. Continue until the whole $\tilde{z}^{1:L}$ has been defined. In short, this means that we must define $\tilde{z}^{1:L}$ such that for any two neighbour locations, $\tilde{z}$ must necessarily be permuted at exactly one of the two locations. This is possible thanks to the absence of cycles of odd length. See figure 3.4 for a graphical representation.

[6] Recall that the generic aspect of the identifiability result we obtained in Proposition 3.3.1 concerns only the part of the parameter space with the connectivity parameter, so that the result holds without external field.

[7] i.e. $\tilde{z}^{l'}$ is a permutation of $z^{l'}$, switching groups 1 and 2

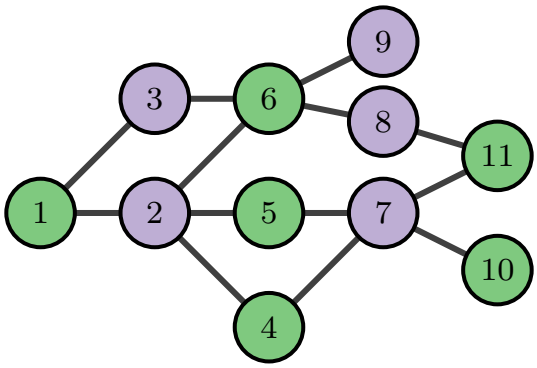


Fig. 3.4 Example of a location graph $\mathcal{G}$ with 10 locations and no cycle of odd length. For any configuration $z^{1:L}$, for the locations in green, $\tilde{z}^l$ is the same than z^l. For the locations in purple, $\tilde{z}^l$ is equal to a permutation (switching the groups 1 and 2) of z^l. The absence of cycles of odd length implies that two nodes of the same colour cannot be neighbours.

We can write the probability distribution of $X^{1:L}$ as follows

$$\begin{aligned}
\mathbb{P}_\theta(X^{1:L}) &= \sum_{z^{1:L}} \mathbb{P}_\pi(X^{1:L} \mid Z^{1:L} = z^{1:L}) \mathbb{P}_\psi(Z^{1:L} = z^{1:L}) \\
&= \sum_{z^{1:L}} \left\{ \left[\prod_{l=1}^{L} \prod_{i<j} (\pi^l_{z^l_i z^l_j})^{X^l_{ij}} (1 - \pi^l_{z^l_i z^l_j})^{1-X^l_{ij}} \right] \prod_{i=1}^{n} \frac{1}{S(\beta)} \exp\left(\sum_{(l,l')\in E} \beta \mathbb{1}_{z^l_i = z^{l'}_i} \right) \right\} \\
&= \sum_{z^{1:L}} \left\{ \left[\prod_{l=1}^{L} \prod_{i<j} \left((\pi_{\text{in}})^{\mathbb{1}_{z^l_i = z^l_j}} (\pi^l_{\text{out}})^{\mathbb{1}_{z^l_i \neq z^l_j}} \right)^{X^l_{ij}} \left(1 - (\pi_{\text{in}})^{\mathbb{1}_{z^l_i = z^l_j}} (\pi^l_{\text{out}})^{\mathbb{1}_{z^l_i \neq z^l_j}} \right)^{1-X^l_{ij}} \right] \right. \\
&\qquad \left. \times \prod_{i=1}^{n} \frac{1}{S(\beta)} \exp\left(\sum_{(l,l')\in E} \beta \mathbb{1}_{z^l_i = z^{l'}_i} \right) \right\}.
\end{aligned}$$

Then, using the definitions of $\tilde{\beta}$, $\tilde{\pi}$ and $\tilde{z}$, and in particular noticing that

- if locations l and l' are neighbours, $z^l_i = z^{l'}_i \iff \tilde{z}^l_i \neq \tilde{z}^{l'}_i$ for any $i \in [\![1, n]\!]$,
- $z^l_i = z^l_j \iff \tilde{z}^l_i = \tilde{z}^l_j$ for any $i, j \in [\![1, n]\!]$ and any $l \in [\![1, L]\!]$,

we have

$$\begin{aligned}
\mathbb{P}_\theta(X^{1:L}) = \sum_{z^{1:L}} \left\{ \left[\prod_{l=1}^{L} \prod_{i<j} \left((\tilde{\pi}_{\text{in}})^{\mathbb{1}_{\tilde{z}^l_i = \tilde{z}^l_j}} (\tilde{\pi}^l_{\text{out}})^{\mathbb{1}_{\tilde{z}^l_i \neq \tilde{z}^l_j}} \right)^{X^l_{ij}} \left(1 - (\tilde{\pi}_{\text{in}})^{\mathbb{1}_{\tilde{z}^l_i = \tilde{z}^l_j}} (\tilde{\pi}^l_{\text{out}})^{\mathbb{1}_{\tilde{z}^l_i \neq \tilde{z}^l_j}} \right)^{1-X^l_{ij}} \right] \right. \\
\left. \times \prod_{i=1}^{n} \frac{1}{S(\beta)} \exp\left(- \sum_{(l,l')\in E} \tilde{\beta} \mathbb{1}_{\tilde{z}^l_i \neq \tilde{z}^{l'}_i} \right) \right\}.
\end{aligned}$$

Finally, using the fact that summing over the $\tilde{z}^{1:L}$ is equivalent to summing over the $z^{1:L}$ and that $\mathbb{1}_{\tilde{z}_i^l \neq \tilde{z}_i^{l'}} = 1 - \mathbb{1}_{\tilde{z}_i^l = \tilde{z}_i^{l'}}$

$$\begin{aligned}
\mathbb{P}_\theta(X^{1:L}) &= \sum_{\tilde{z}^{1:L}} \left\{ \left[\prod_{l=1}^{L} \prod_{i<j} \left((\tilde{\pi}_{\text{in}})^{\mathbb{1}_{\tilde{z}_i^l = \tilde{z}_j^l}} (\tilde{\pi}_{\text{out}}^l)^{\mathbb{1}_{\tilde{z}_i^l \neq \tilde{z}_j^l}} \right)^{X_{ij}^l} \left(1 - (\tilde{\pi}_{\text{in}})^{\mathbb{1}_{\tilde{z}_i^l = \tilde{z}_j^l}} (\tilde{\pi}_{\text{out}}^l)^{\mathbb{1}_{\tilde{z}_i^l \neq \tilde{z}_j^l}} \right)^{1 - X_{ij}^l} \right] \right. \\
&\qquad \left. \times \prod_{i=1}^{n} \frac{1}{S(\beta) \exp(|E|\tilde{\beta})} \exp\left(\sum_{(l,l') \in E} \tilde{\beta} \mathbb{1}_{\tilde{z}_i^l = \tilde{z}_i^{l'}} \right) \right\} \\
&= \sum_{z^{1:L}} \left\{ \left[\prod_{l=1}^{L} \prod_{i<j} \left((\tilde{\pi}_{\text{in}})^{\mathbb{1}_{z_i^l = z_j^l}} (\tilde{\pi}_{\text{out}}^l)^{\mathbb{1}_{z_i^l \neq z_j^l}} \right)^{X_{ij}^l} \left(1 - (\tilde{\pi}_{\text{in}})^{\mathbb{1}_{z_i^l = z_j^l}} (\tilde{\pi}_{\text{out}}^l)^{\mathbb{1}_{z_i^l \neq z_j^l}} \right)^{1 - X_{ij}^l} \right] \right. \\
&\qquad \left. \times \prod_{i=1}^{n} \frac{1}{S(\beta) \exp(|E|\tilde{\beta})} \exp\left(\sum_{(l,l') \in E} \tilde{\beta} \mathbb{1}_{z_i^l = z_i^{l'}} \right) \right\}.
\end{aligned}$$

We can then express this probability distribution in terms of the parameter $\tilde{\theta}$, the two parameters then leading to the same distribution. This relies on the fact that the probability of observing any $z^{1:L}$ under the parameter β is the same as that of observing $\tilde{z}^{1:L}$ under the parameter $\tilde{\beta}$, and that the distribution of $X^{1:L}$ given $Z^{1:L} = z^{1:L}$ under parameter π is the same as that of $X^{1:L}$ given $Z^{1:L} = \tilde{z}^{1:L}$ under the parameter $\tilde{\pi} = \pi$ (due to the fact that we are in an affiliation model). We remark that the normalising constant of the Potts model with $Q = 2$, with parameter $\tilde{\beta} = -\beta$ and location graph $\mathcal{G} = (V, E)$ (with no cycles of odd length) is equal to $S(\beta) \exp(|E|\tilde{\beta})$ where $S(\beta)$ is the normalising constant of the Potts model with $Q = 2$, with parameter β and location graph $\mathcal{G}$.

□

3.4 Estimation

3.4.1 Likelihood

We can write the conditional log-likelihood and the likelihood, using the conditional independence of the X_{ij}^ls (given the Z_i^ls), as

$$\begin{aligned}
\log \mathbb{P}_\theta(X^{1:L} | Z^{1:L} = z^{1:L}) &= \sum_{l=1}^{L} \log \mathbb{P}_\pi(X^l | Z^l = z^l) \\
&= \sum_{l=1}^{L} \sum_{i<j} X_{ij}^l \log \pi_{z_i^l z_j^l}^l + (1 - X_{ij}^l) \log(1 - \pi_{z_i^l z_j^l}^l) \qquad (3.4.1)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_\theta(X^{1:L}) \qquad (3.4.2)\\
&= \sum_{z^{1:L}} \mathbb{P}_\pi(X^{1:L}|Z^{1:L}=z^{1:L})\mathbb{P}_\psi(Z^{1:L}=z^{1:L})\\
&= \sum_{z^{1:L}} \left[\left\{ \prod_{l=1}^{L} \mathbb{P}_\pi(X^l|Z^l=z^l) \right\} \prod_{i=1}^{n} \frac{1}{S_i(\alpha_i,\beta_i)} \exp\left(\sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{z_i^l=q} + \beta_i \sum_{(l,l')\in E_i} \mathbb{1}_{z_i^l=z_i^{l'}} \right) \right]\\
&= \sum_{z^{1:L}} \left[\left\{ \prod_{l=1}^{L} \prod_{i<j} (\pi^l_{z_i^l z_j^l})^{X_{ij}^l} (1-\pi^l_{z_i^l z_j^l})^{1-X_{ij}^l} \right\} \right.\\
&\qquad \left. \times \prod_{i=1}^{n} \frac{1}{S_i(\alpha_i,\beta_i)} \exp\left(\sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{z_i^l=q} + \beta_i \sum_{(l,l')\in E_i} \mathbb{1}_{z_i^l=z_i^{l'}} \right) \right]. \qquad (3.4.3)
\end{aligned}$$

The log-likelihood is then written

$$\log \mathbb{P}_\theta(X^{1:L}) = \log \left(\sum_{z^{1:L}} \mathbb{P}_\pi(X^{1:L}|Z^{1:L}=z^{1:L})\mathbb{P}_\psi(Z^{1:L}=z^{1:L}) \right). \qquad (3.4.4)$$

This is the quantity we would like to maximise to estimate the model parameters, but as we will see in the following, we will not be able to maximise it directly.

3.4.2 Maximum likelihood approach

We want to estimate the model parameters relying on a maximum likelihood approach. We then want to maximise the following quantity

$$\begin{aligned}
\mathbb{P}_\theta(X^{1:L}) = \sum_{z^{1:L}} \left[\left\{ \prod_{l=1}^{L} \prod_{i<j} (\pi^l_{z_i^l z_j^l})^{X_{ij}^l} (1-\pi^l_{z_i^l z_j^l})^{1-X_{ij}^l} \right\} \right.\\
\left. \times \prod_{i=1}^{n} \frac{1}{S_i(\alpha_i,\beta_i)} \exp\left(\sum_{q=1}^{Q} \alpha_{iq} \sum_{l=1}^{L} \mathbb{1}_{z_i^l=q} + \beta_i \sum_{(l,l')\in E_i} \mathbb{1}_{z_i^l=z_i^{l'}} \right) \right].
\end{aligned}$$

We cannot maximise it directly because the computation of this quantity involves a summation over all the Q^{nL} possible latent configurations, in addition to the normalising constants $S_i(\alpha_i,\beta_i)$ being intractable. We cannot either use the Expectation-Maximization (EM) algorithm to approximate it because it involves the computation of the conditional distribution of the latent variables given the observations which is not tractable, and it still involves the intractable normalising constants. We will then rely on an alternative of the EM algorithm, the simulated field algorithm as defined in Celeux et al. (2003). It

consists of simulating a latent configuration and using a mean-field like approximation (for the conditional distribution of the latent variables given the observations, and the distribution of the latent variables) in order to make the considered conditional distribution and the normalising constant tractable. The approximation consists in replacing the intractable distributions by distributions that factorise over the locations, and such that for each location, the neighbours of this location are fixed at the values of the simulated configuration.

3.4.3 EM algorithm

We first describe briefly the EM algorithm (Dempster et al., 1977) in this case and see why we need to use an alternative. EM is an iterative algorithm used to approximate maximum likelihood estimates of parameters in statistical models, in the presence of latent variables. We start with some initial value of the parameter $\theta^{(0)} = (\psi^{(0)}, \pi^{(0)}) = (\alpha^{(0)}, \beta^{(0)}, \pi^{(0)})$. At iteration t of the EM algorithm, we want to maximise with respect to θ the quantity

$$\begin{aligned} Q(\theta|\theta^{(t-1)}) =& \mathbb{E}_{\theta^{(t-1)}}\left[\log \mathbb{P}_\theta\left(Z^{1:L}, X^{1:L}\right)\middle| X^{1:L}\right] \\ =& \mathbb{E}_{\theta^{(t-1)}}\left[\log \mathbb{P}_\pi\left(X^{1:L} \mid Z^{1:L}\right)\middle| X^{1:L}\right] + \mathbb{E}_{\theta^{(t-1)}}\left[\log \mathbb{P}_\psi\left(Z^{1:L}\right)\middle| X^{1:L}\right] \\ :=& Q_1(\pi|\theta^{(t-1)}) + Q_2(\alpha, \beta|\theta^{(t-1)}). \end{aligned} \tag{3.4.5}$$

Notice that the first term Q_1 only depends on the connection probabilities (or emission parameter) π whereas the second term Q_2 only depends on the parameters α and β of the Gibbs distribution. We can write for the first term

$$\begin{aligned} Q_1(\pi|\theta^{(t-1)}) &= \mathbb{E}_{\theta^{(t-1)}}\left[\sum_{l=1}^{L}\sum_{i<j} X_{ij}^l \log \pi^l_{Z_i^l Z_j^l} + (1 - X_{ij}^l)\log(1 - \pi^l_{Z_i^l Z_j^l})\middle| X^{1:L}\right] \\ &= \sum_{l=1}^{L}\sum_{z^l}\sum_{i<j}\left(X_{ij}^l \log \pi^l_{z_i^l z_j^l} + (1 - X_{ij}^l)\log(1 - \pi^l_{z_i^l z_j^l})\right)\mathbb{P}_{\theta^{(t-1)}}\left(z^l \,\middle|\, X^{1:L}\right). \end{aligned} \tag{3.4.6}$$

The second term can be written as follows

$$Q_2(\alpha,\beta|\theta^{(t-1)}) = \mathbb{E}_{\theta^{(t-1)}}\left[\sum_{i=1}^{n}\left(-\log S_i(\alpha_i,\beta_i) + \sum_{q=1}^{Q}\alpha_{iq}\sum_{l=1}^{L}\mathbb{1}_{Z_i^l=q} + \sum_{(l,l')\in E_i}\beta_i\mathbb{1}_{Z_i^l=Z_i^{l'}}\right)\middle| X^{1:L}\right]$$

$$= \sum_{i=1}^{n}\left(-\log S_i(\alpha_i,\beta_i) + \sum_{q=1}^{Q}\alpha_{iq}\sum_{l=1}^{L}\mathbb{P}_{\theta^{(t-1)}}\left(Z_i^l = q \,\middle|\, X^{1:L}\right)\right.$$

$$\left. + \beta_i \sum_{(l,l')\in E_i}\mathbb{P}_{\theta^{(t-1)}}\left(Z_i^l = Z_i^{l'} \,\middle|\, X^{1:L}\right)\right). \tag{3.4.7}$$

Each iteration is composed of two consecutive steps (E-step and M-step). The E step consists in computing the quantity $Q(\theta|\theta^{(t-1)})$, and the M (Maximization) step consists in maximising this quantity with respect to θ to obtain $\theta^{(t)}$, the estimate at step t. It is proven that the log-likelihood increases at each iteration. Here, the E step is intractable because neither the normalising constants $S_i(\alpha_i,\beta_i)$ appearing in Q_2 nor the distribution of the latent variables given the observations (needed in both Q_1 and Q_2) can be computed. Using a mean-field like approximation for both the distribution of the latent variables and of the latent variables given the observations will solve these problems. We will describe this method in the following section.

3.4.4 Mean-field like approximation

The idea of this approach is to replace the intractable distributions (of the latent variables, and of the latent variables given the observations) with simpler distributions.

We use a mean-field like approximation in the EM algorithm, both for the distribution of the latent variables and that of the latent variables given the observations. The mean field like approximation, in a classical setup (when considering a single HMRF), consists in considering approximations of the distributions that factorise over the locations, where for each location l, the values of the latent variable at neighbour locations of l are set to deterministic values. This approach is a generalisation of the mean-field approximation (see Chandler (1987); Celeux et al. (2003) and Appendix B.2), in which the fixed values of the neighbours are their means. In our setup, we will consider approximate distributions factorising over both the locations and species. More precisely, given a fixed configuration

$\tilde{z}^{1:L}$, the distribution of $Z^{1:L}$ appearing in Q_2 is approximated by the following distribution

$$\mathbb{P}_{\psi}^{\tilde{z}}(Z^{1:L} = z^{1:L}) \coloneqq \prod_{l=1}^{L}\prod_{i=1}^{n} \mathbb{P}_{\psi}(Z_i^l = z_i^l \mid Z_i^{-l} = \tilde{z}_i^{-l}) = \prod_{l=1}^{L}\prod_{i=1}^{n} \mathbb{P}_{\psi}(Z_i^l = z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)}) \tag{3.4.8}$$

where $\mathcal{N}_i(l)$ is the set of neighbours of l in the location graph $\mathcal{G}_i$ of species i, i.e.

$$\mathcal{N}_i(l) = \{l'; (l, l') \in E_i\}.$$

Moreover note that

$$\mathbb{P}_{\psi}(Z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)}) = \frac{\exp\left(\alpha_{iZ_i^l} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{Z_i^l = \tilde{z}_i^{l'}}\right)}{\sum_{q=1}^{Q} \exp(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{q = \tilde{z}_i^{l'}})}. \tag{3.4.9}$$

Now, for the distribution of the latent variable $Z^{1:L}$ given the observations $X^{1:L}$ (that appears in the expectation in both Q_1 and Q_2), we consider instead given a fixed configuration $\tilde{z}^{1:L}$ the following approximation

$$\begin{aligned}
&\mathbb{P}_{\theta^{(t-1)}}^{\tilde{z}}(Z^{1:L} = z^{1:L} \mid X^{1:L}) \\
\propto &\mathbb{P}_{\pi^{(t-1)}}^{\tilde{z}}(X^{1:L} \mid Z^{1:L} = z^{1:L}) \mathbb{P}_{\psi^{(t-1)}}^{\tilde{z}}(Z^{1:L} = z^{1:L}) \\
\propto &\prod_{l=1}^{L}\prod_{i=1}^{n} \mathbb{P}_{\pi^{(t-1)}}(X_{i\bullet}^l \mid Z_i^l = z_i^l, Z_{-i}^l = \tilde{z}_{-i}^l) \mathbb{P}_{\psi^{(t-1)}}(Z_i^l = z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)}) \\
\propto &\prod_{l=1}^{L}\prod_{i=1}^{n} \mathbb{P}_{\theta^{(t-1)}}(z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)}, Z_{-i}^l = \tilde{z}_{-i}^l, X_{i\bullet}^l),
\end{aligned} \tag{3.4.10}$$

where $\propto$ means "proportional to", and $X_{i\bullet}^l = (X_{i1}^l, \ldots, X_{in}^l)$. Note that we make two approximations here, for the probability distribution of $Z^{1:L}$ in (3.4.8), as in the classical mean field like approximation, and an additional approximation for the conditional probability distribution of $X^{1:L}$ given $Z^{1:L}$, assuming that the neighbours of a location are fixed (set to the values $\tilde{z}_i^{\mathcal{N}_i(l)}$), but also the values of the other species at the same location (set to $\tilde{z}_{-i}^l$).

3.4.5 Simulated EM Algorithm

We consider an EM algorithm with a mean-field like approximation, and in particular the simulated EM (Celeux et al., 2003), in which the mean field like approximation is

based on a realisation of the conditional distribution of the latent variables given the observations.

At each step t of the algorithm, we approximate the distribution of the latent variables given the observations and the distribution of the latent variables by their mean-field like approximations as defined in (3.4.10) and (3.4.8) respectively, where $\tilde{z}^{1:L}$ is drawn from the distribution $\mathbb{P}_{\theta^{(t-1)}}(Z^{1:L} \mid X^{1:L})$ thanks to a Gibbs sampler. This will allow us to compute the expectations with respect to that distribution of $Z^{1:L}$ given $X^{1:L}$ (see Equation (3.4.5)) and to get rid of the intractable normalising constant in the term $\mathbb{P}_{\psi}(Z^{1:L})$ appearing in Q_2.

Each iteration t of the algorithm is then divided into two steps. One consists in simulating a configuration $\tilde{z}^{1:L}$ from a current parameter value $\theta^{(t-1)}$ and from the observations $X^{1:L}$ using a Gibbs sampler, in order to replace the distributions (of $Z^{1:L}$ and of $Z^{1:L}$ given the observations $X^{1:L}$) by their mean-field like approximations. The second one is an EM algorithm iteration (using the mean-field like approximations). Our algorithm is summarised in Algorithm 6.

3.4.6 Step 1: Simulation of a configuration for the mean-field like approximation

The first step is the simulation of a configuration $\tilde{z}^{1:L}$ from the conditional distribution $\mathbb{P}_{\theta^{(t-1)}}(Z^{1:L} \mid X^{1:L})$ relying on a Gibbs sampler.

Gibbs sampling The Gibbs sampling algorithm is such that at each iteration $m \in [\![1, M]\!]$, for each $l \in [\![1, L]\!]$ and $i \in [\![1, n]\!]$, $z_i^{(m)l}$ is simulated from the distribution

$$\begin{aligned}\mathbb{P}_{\theta^{(t-1)}}(Z_i^l \mid X^l, z^{-l}, z_{-i}^l) &= \mathbb{P}_{\theta^{(t-1)}}(Z_i^l \mid X^l, z_i^{\mathcal{N}_i(l)}, z_{-i}^l) \\ &\propto \mathbb{P}_{\pi^{(t-1)}}(X^l \mid Z_i^l, z_{-i}^l)\mathbb{P}_{\psi^{(t-1)}}(Z_i^l \mid z_i^{\mathcal{N}_i(l)}) \end{aligned} \tag{3.4.11}$$

where $z^{1:L}$ is the current configuration in the algorithm, and recalling that the probability of the observations at location l given the latent classes of the species at this location is given in (3.2.2) and that the probability $\mathbb{P}_{\psi^{(t-1)}}(Z_i^l \mid z_i^{-l})$ of Z_i^l given the values of z_i^{-l} at the other locations depends only on the neighbour locations and is given in (3.4.9). More precisely, we use Algorithm 5. Note that in order for the Gibbs sampler to be valid, we must update at each step m the nL components of the configuration one at a time and not simultaneously.

Algorithm 5: Conditional Gibbs sampler

input : A number of iterations M, a collection of location graphs $\mathcal{G}_1, \dots \mathcal{G}_n$, a parameter $\theta^{(t-1)} = (\alpha^{(t-1)}, \beta^{(t-1)}, \pi^{(t-1)})$

output : A realisation of the random variable $Z^{1:L} \sim \mathbb{P}_{\theta^{(t-1)}}(\cdot \mid X^{1:L})$

Initialise an arbitrary configuration $z^{(0)1:L}$;
for $m = 1$ **to** M **do**
  for $l = 1$ **to** L **do**
    for $i = 1$ **to** n **do**
      $z^{1:L} \leftarrow (z^{(m)1:l-1}, z^{(m)l}_{1:i-1}, z^{(m-1)l}_{i:n}, z^{(m-1)l+1:L})$;
      Draw $z_i^{(m)l}$ from the conditional distribution $\mathbb{P}_{\theta^{(t-1)}}(Z_i^{(m)l} \mid X^{1:L}, z^l_{-i}, z^{-l})$ in (3.4.11) ;
    end
  end
end
return $z^{(M)1:L}$

In practice, at each iteration of the simulated EM algorithm, we only carry out a single (M=1) or just a few iterations of the Gibbs sampler, starting from the previous configuration.

3.4.7 Step 2: EM iteration

As mentioned before, the mean-field like approximation based on the configuration drawn in the first step of the algorithm is what makes this second step tractable. Indeed, replacing the conditional distribution $\mathbb{P}_{\theta^{(t-1)}}(Z^{1:L} \mid X^{1:L})$ by its approximation in (3.4.10) solves the problem of the computation of the conditional expectations appearing in Q_1 and Q_2, and replacing the marginal distribution $\mathbb{P}_\psi(Z_i^{1:L})$ by $\prod_{l=1}^L \mathbb{P}_\psi(Z_i^l \mid Z^{\mathcal{N}_i(l)} = \tilde{z}^{\mathcal{N}_i(l)})$ for every $i \in [\![1, n]\!]$ solves the problem of the computation of the normalising constants appearing in Q_2. Recalling that we denote by $\theta^{(t-1)} = (\alpha^{(t-1)}, \beta^{(t-1)}, \pi^{(t-1)})$ the previous parameter (at iteration $t-1$), we have at iteration t the following two steps of the algorithm.

3.4.7.1 E step

We want to compute the approximations of the quantities $Q_1(\pi|\theta^{(t-1)})$ and $Q_2(\alpha, \beta|\theta^{(t-1)})$ (see (3.4.5)). Recall that Q_1 and Q_2 are given by (3.4.6) and (3.4.7) respectively.

Lemma 3.4.1. *The approximations $\tilde{Q}_1$ and $\tilde{Q}_2$ of Q_1 and Q_2 respectively under the mean-field like approximation based on $\tilde{z}^{1:L}$ are given by*

$$\tilde{Q}_1(\pi|\theta^{(t-1)}) = \sum_{1\le q<q'\le Q}\sum_{l=1}^{L}\sum_{i\neq j}\left\{\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_i^l = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_j^l = q' \mid X^l) f(X_{ij}^l, \pi_{qq'}^l)\right\} + \sum_{q=1}^{Q}\sum_{l=1}^{L}\sum_{i<j}\left\{\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_i^l = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_j^l = q \mid X^l) f(X_{ij}^l, \pi_{qq})\right\} \quad (3.4.12)$$

and

$$\tilde{Q}_2(\alpha,\beta|\theta^{(t-1)}) = \sum_{i=1}^{n}\sum_{l=1}^{L}\left\{\left(\sum_{q=1}^{Q}\alpha_{iq}\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}\left(Z_i^l = q \,\middle|\, X^l\right)\right) + \beta_i \sum_{l'\in\mathcal{N}_i(l)}\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}\left(Z_i^l = \tilde{z}_i^{l'} \,\middle|\, X^l\right) - \log\left[\sum_{q=1}^{Q}\exp\left(\alpha_{iq} + \beta_i \sum_{l'\in\mathcal{N}_i(l)}\mathbb{1}_{q=\tilde{z}_i^{l'}}\right)\right]\right\}. \quad (3.4.13)$$

where $f(X_{ij}^l, \pi_{qq'}^l)$ is the probability of X_{ij}^l conditional to $\pi^l_{Z_i^l Z_j^l} = \pi_{qq'}^l$, i.e.

$$f(X_{ij}^l, \pi_{qq'}^l) = \mathbb{P}_\pi(X_{ij}^l \mid Z_i^l = q, Z_j^l = q') = X_{ij}^l \log \pi_{qq'}^l + (1 - X_{ij}^l)\log(1 - \pi_{qq'}^l)$$

and where $\mathbb{P}^{\tilde{z}}$ stands for a probability distribution under the mean field like approximation based on $\tilde{z}^{1:L}$ and is given in this case by

$$\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_i^l = q \mid X^l) = \frac{\mathbb{P}_{\pi^{(t-1)}}(X_{i\cdot}^l \mid Z_{-i}^l = \tilde{z}_{-i}^l, Z_i^l = q)\mathbb{P}_{\psi^{(t-1)}}(Z_i^l = q \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)})}{\sum_{z_i^l\in[\![1,Q]\!]}\mathbb{P}_{\pi^{(t-1)}}(X_{i\cdot}^l \mid Z_i^l = z_i^l, Z_{-i}^l = \tilde{z}_{-i}^l)\mathbb{P}_{\psi^{(t-1)}}(Z_i^l = z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)})}.$$

A proof of this lemma can be found in Section 3.6. Note that all the quantities appearing in $\tilde{Q}_1$ and $\tilde{Q}_2$ can now be computed easily, and we will then be able to maximise them in the following M-step.

3.4.7.2 M step

The aim of the M-step is then to maximise respectively with respect to π and (α, β) the approximations of the quantities $Q_1(\pi|\theta^{(t-1)})$ and $Q_2(\alpha,\beta|\theta^{(t-1)})$ obtained above. This is the role of the two following lemmas.

Lemma 3.4.2. *The parameter* $\tilde{\pi}$ *maximising the approximation* $\tilde{Q}_1(\pi|\theta^{(t-1)})$ *(see Equation (3.4.12)) of* $Q_1(\pi|\theta^{(t-1)})$ *satisfies for every* $l \in [\![1, L]\!]$ *and any* $1 \leq q \neq q' \leq Q$

$$\tilde{\pi}^l_{qq'} = \frac{\sum_{i \neq j} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_j = q' \mid X^l)X^l_{ij}}{\sum_{i \neq j} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_j = q' \mid X^l)} \tag{3.4.14}$$

and for any $q \in [\![1, Q]\!]$

$$\tilde{\pi}_{qq} = \frac{\sum_{l=1}^{L} \sum_{i<j} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_j = q \mid X^l)X^l_{ij}}{\sum_{l=1}^{L} \sum_{i<j} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_j = q \mid X^l)}. \tag{3.4.15}$$

The proof of this lemma is immediate, consisting in differentiating $\tilde{Q}_1(\pi|\theta^{(t-1)})$ (in Equation (3.4.12)) with respect to $\pi^l_{qq'}$ for any $q, q' \in [\![1, Q]\!]$ and $l \in [\![1, L]\!]$ and setting these derivatives equal to zero, and is therefore omitted. We then state a result for the maximisation of $\tilde{Q}_2(\alpha, \beta|\theta^{(t-1)})$ with respect to α, β. We recall that we impose the constraint $\alpha_{i1} = 0$ on the parameter.

Lemma 3.4.3. *The parameter* $(\tilde{\alpha}, \tilde{\beta})$ *maximising the approximation* $\tilde{Q}_2(\alpha, \beta|\theta^{(t-1)})$ *(see Equation (3.4.13)) of* $Q_2(\alpha, \beta|\theta^{(t-1)})$ *satisfies for any* $i \in [\![1, n]\!]$

$$\tilde{\alpha}_{i1} = 0, \qquad \frac{\partial \tilde{Q}_2(\tilde{\alpha}, \tilde{\beta}|\theta^{(t-1)})}{\partial \alpha_{iq}} = 0 \ \ \forall q \in [\![2, Q]\!] \qquad \textit{and} \qquad \frac{\partial \tilde{Q}_2(\tilde{\alpha}, \tilde{\beta}|\theta^{(t-1)})}{\partial \beta_i} = 0$$

where

$$\frac{\partial \tilde{Q}_2(\alpha, \beta|\theta^{(t-1)})}{\partial \alpha_{iq}} = \sum_{l=1}^{L} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l) - \sum_{l=1}^{L} \frac{\exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}^{l'}_i = q}\right)}{\sum_{q'=1}^{Q} \exp\left(\alpha_{iq'} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}^{l'}_i = q'}\right)} \tag{3.4.16}$$

and

$$\begin{aligned}\frac{\partial \tilde{Q}_2(\alpha, \beta|\theta^{(t-1)})}{\partial \beta_i} = &\sum_{l=1}^{L} \sum_{l' \in \mathcal{N}_i(l)} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = \tilde{z}^{l'}_i \mid X^l) \\ &- \sum_{l=1}^{L} \frac{\sum_{q=1}^{Q} \left(\sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}^{l'}_i = q}\right) \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}^{l'}_i = q}\right)}{\sum_{q=1}^{Q} \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}^{l'}_i = q}\right)}.\end{aligned} \tag{3.4.17}$$

The proof of this lemma is also immediate and is omitted. Note that we do not obtain a closed-form expression for the parameters α and β for which the derivatives in (3.4.16) and (3.4.17) are equal to zero, and need to rely on numerical methods. We choose to use a Newton-Raphson algorithm to find the zeros of these derivatives.

Remark 3.4.1. Note that we could also consider the particular case where the parameters of the MRF (α and β) are the same for every species (and then are denoted by $\alpha = (\alpha_q)_{1\leq q\leq Q} \in \mathbb{R}^Q$ and $\beta \in \mathbb{R}$). In that case, the derivatives with respect to α_q and β are respectively

$$\frac{\partial \tilde{Q}_2(\alpha,\beta|\theta^{(t-1)})}{\partial \alpha_q} = \sum_{i=1}^{n}\sum_{l=1}^{L} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_i^l = q \mid X^l) - \sum_{i=1}^{n}\sum_{l=1}^{L} \frac{\exp\left(\alpha_q + \beta \sum_{l'\in\mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}_i^{l'}=q}\right)}{\sum_{q'=1}^{Q} \exp\left(\alpha_{q'} + \beta \sum_{l'\in\mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}_i^{l'}=q'}\right)},$$

and

$$\frac{\partial \tilde{Q}_2(\alpha,\beta|\theta^{(t-1)})}{\partial \beta} = \sum_{i=1}^{n}\sum_{l=1}^{L}\sum_{l'\in\mathcal{N}_i(l)} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_i^l = \tilde{z}_i^{l'} \mid X^l) - \sum_{i=1}^{n}\sum_{l=1}^{L} \frac{\sum_{q=1}^{Q}\left(\sum_{l'\in\mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}_i^{l'}=q}\right)\exp\left(\alpha_{iq} + \beta_i \sum_{l'\in\mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}_i^{l'}=q}\right)}{\sum_{q=1}^{Q}\exp\left(\alpha_{iq} + \beta_i \sum_{l'\in\mathcal{N}_i(l)} \mathbb{1}_{\tilde{z}_i^{l'}=q}\right)}.$$

Algorithm 6: Simulated EM for the space-evolving SBM

input : A collection of observed graphs $X^{1:L}$, a collection of location graphs $\mathcal{G}_1,\dots,\mathcal{G}_n$, a number of groups Q, an initial parameter $\theta^{(0)} = (\alpha^{(0)},\beta^{(0)},\pi^{(0)})$

output : A parameter estimate $\tilde{\theta}$

Initialise an arbitrary configuration $\tilde{z}^{(0)1:L}$ and $t = 0$;
while *the algorithm has not converged* **do**
  Set $t \leftarrow t+1$;
  Simulation step: ;
  Draw a configuration $\tilde{z}^{(t)1:L}$ with the Gibbs sampler algorithm (Algorithm 5) with parameter $\theta^{(t-1)}$;
  EM step: ;
  Compute $\pi^{(t)}$ the estimator of π from the formulas (3.4.14) and (3.4.15) ;
  Compute $\alpha^{(t)}$ and $\beta^{(t)}$ by setting the quantities in (3.4.16) and (3.4.17) equal to zero with a Newton-Raphson algorithm ;
end
return $\tilde{\theta} := \theta^{(t)}$

3.4.8 Initialisation and stopping criterion of the algorithm

It is well known that the EM algorithm may only find a local maximum of the likelihood, depending on the initial parameter. The usual strategy is then to run the algorithm multiple times with different initialisations and compare the obtained estimators in terms of the likelihood. Here we cannot compute the likelihood, but we can rely on the approximation $\tilde{Q}$ of Q obtained in the final iteration of the algorithm in order to choose the best estimator. We can also use this approximation for stopping the algorithm.

We can start from some random initial parameters, or from a "reasonable" value of the parameter, as we now describe. We start by performing an absolute spectral clustering of the nodes at each location and thus obtain an estimated clustering $\hat{z}^l$ at each location l. At each location l, the initial connection probabilities $\pi^{(0)l}$ are computed from $\hat{z}^l$[8], and the rows and columns of $\pi^{(0)l}$ are permuted such that the diagonal is in ascending order (in order to take into account the potential label switching between the different locations). The initial parameter $\alpha^{(0)}$ is defined as follows for every $i \in [\![1, n]\!]$

$$\alpha_{i1}^{(0)} = 0 \quad \text{and} \quad \alpha_{iq}^{(0)} = \frac{1}{L}\left(\sum_{l=1}^{L} \mathbb{1}_{\hat{z}_i^l = q} - \sum_{l=1}^{L} \mathbb{1}_{\hat{z}_i^l = 1}\right) \ \forall q \in [\![2, Q]\!].$$

The parameter β is initialised at $(0, \ldots, 0)$ (no interaction). In Section 3.5, we will start from four different initialisations, three random ones and one "reasonable" initialisation that we just described.

We observe that the estimators obtained at the different steps of our algorithm are not stable, i.e. do not seem to converge to a fixed value, but their trajectories seem however to exhibit a similar limiting behaviour around a fixed value. As pointed out in Forbes and Fort (2007), the convergence of this algorithm might need to be understood in terms of the ergodic behaviour of the process of the estimators at each step, as it is the case for the stochastic EM (see Diebolt and Celeux (1993); Feodor Nielsen (2000)) in which this process is an homogeneous Markov chain that is ergodic under mild conditions and converges to its stationary distribution. We thus choose to compute a mean of the parameter estimators of the last iterations of the algorithm. We use a stopping condition that is $|\tilde{Q}^{(t)} - \tilde{Q}^{(t-1)}|/|\tilde{Q}^{(t-1)}| < 10^{-5}$, then let the algorithm perform 30 more iterations and take the mean of the parameters and of $\tilde{Q}$ over these 30 iterations. We finally keep the estimator corresponding to the initialisation leading to the highest mean of $\tilde{Q}$ over the last 30 iterations.

[8]For $q_1, q_2 \in [\![1, Q]\!]$, $\pi_{q_1 q_2}^{(0)l}$ is the proportion of present edges between species that are in groups q_1 and q_2 respectively in the configuration $\hat{z}^l$.

3.5 Illustration of the method on synthetic datasets

We illustrate the performance of our algorithm on simulated datasets. For that, we fix $n = 18$ species (leading to $n^2 = 324$ observations of potential edges at each location) and $L = 218$ locations. Moreover, we use location graphs that are grids of size 12×18. We also fix $Q = 2$. For these $n = 18$ species, we consider 3 different values for α_{i2} that are $-0.8, 0, 0.8$ (recall that $\alpha_{i1} = 0$) and three different values for β_i that are $-0.5, 0, 0.5$. We thus have 9 different scenarios for the parameters governing the MRF. We distribute these values such that for each of the 9 pairs of parameters (α_{i2}, β_i), 2 species are governed by a MRF with (α_{i2}, β_i). We fix the diagonal of π at $(0.3, 0.7)$[9], and we also choose 3 different values for π^l_{12} that we fix at $0.2, 0.5, 0.8$, each of these three values being assigned to 72 locations.

With this setup, we run 60 simulations. For each simulated dataset, we start from four different initialisations. Three out of the four iterations are random. The other is defined as in Section 3.4.8. We use a stopping condition that is $|\tilde{Q}^{(t)} - \tilde{Q}^{(t-1)}|/|\tilde{Q}^{(t-1)}| < 10^{-5}$, then let the algorithm perform 30 more iterations and take the mean of the parameters and of $\tilde{Q}$ over these 30 iterations.

We finally keep the estimator corresponding to the initialisation leading to the highest mean of $\tilde{Q}$ over the last 30 iterations. We plotted the boxplots of the estimators of α_{i2} for the three values of the true α_{i2} in Figure 3.5, those of the esimators of β_i for the three true values of β_i in Figure 3.6, those of the estimators of π_{11} and π_{22} in Figure 3.7, and those of the estimators of π^l_{12} for the three true values of π^l_{12} in Figure 3.8. The x-axis indicates the true value of the parameter.

These experiments give good results, in the sense that we can recover the different behaviours from the estimated parameters. We remark that the estimation seems more precise for the connection parameters π than for the parameters of the Potts model.

[9] that is constant over the locations

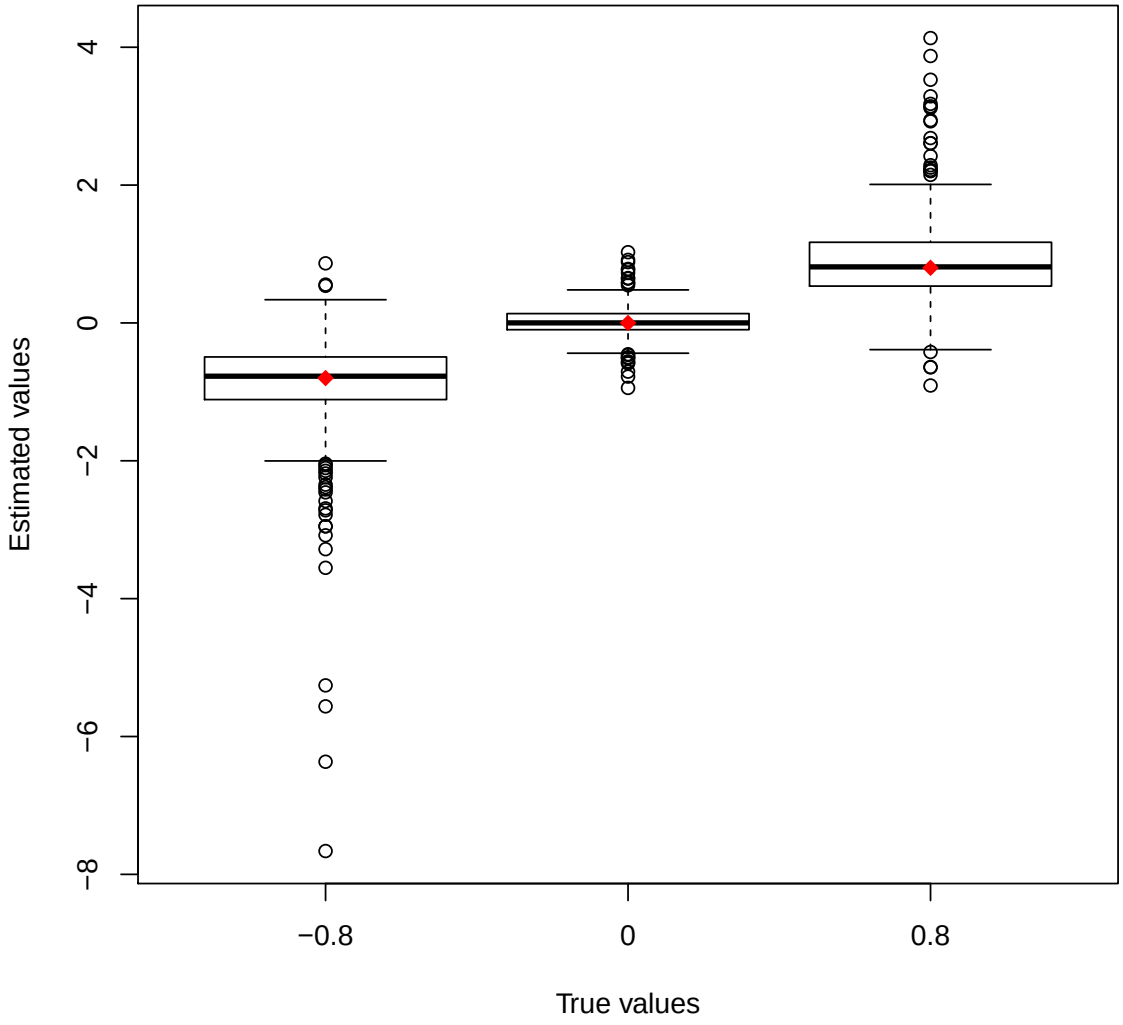


Fig. 3.5 Results of the simulated EM for α_{i2}. The 3 boxplots correspond to the 3 different values $\alpha_{i2} = \{-0.8, 0, 0.8\}$. The true values are marked in red. Note that each true value of the parameter α_{i2} corresponds to 6 species.

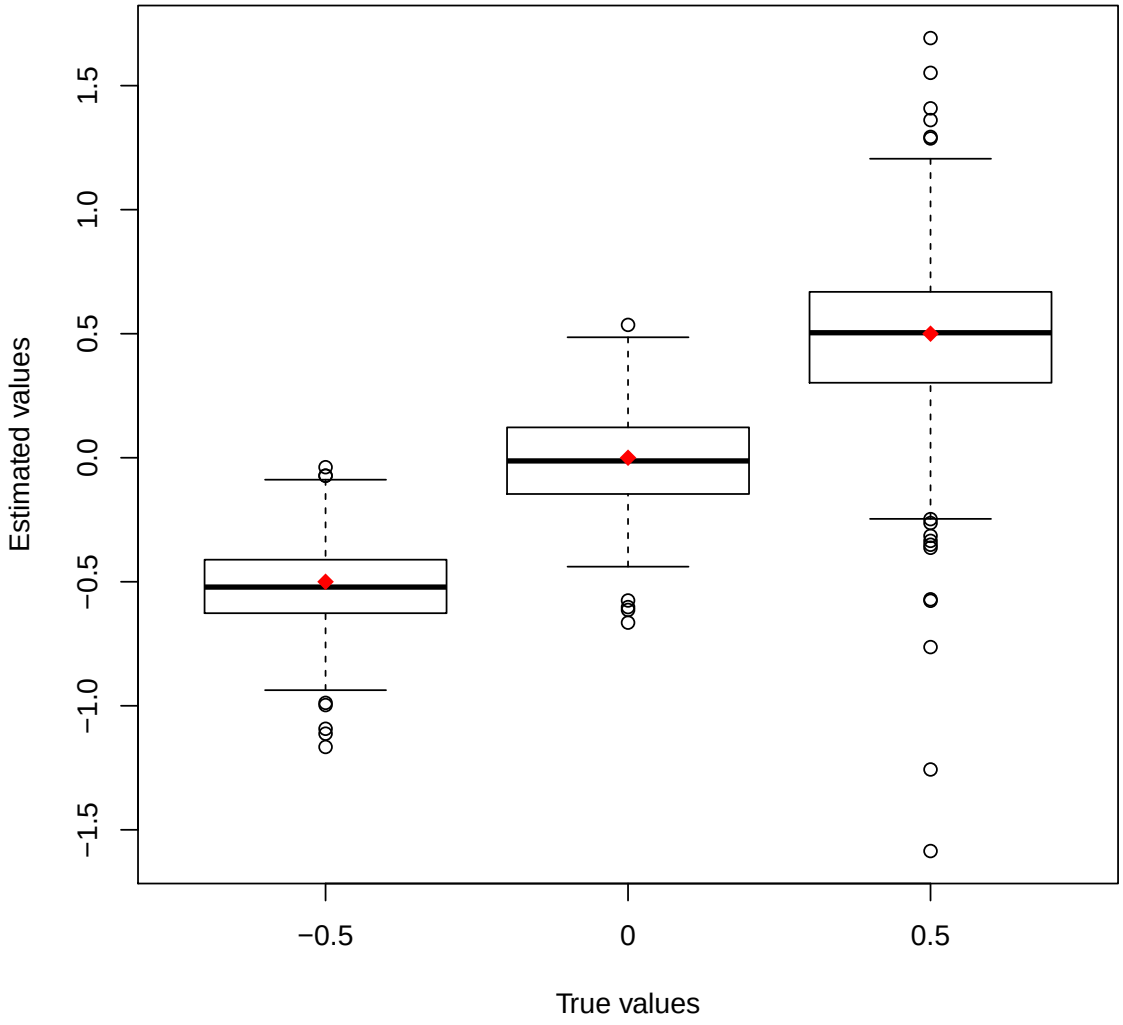


Fig. 3.6 Results of the simulated EM for β_i. The 3 boxplots correspond to the 3 different values $\beta_i = \{-0.5, 0, 0.5\}$. The true values are marked in red. Note that each true value of the parameter β_i corresponds to 6 species.

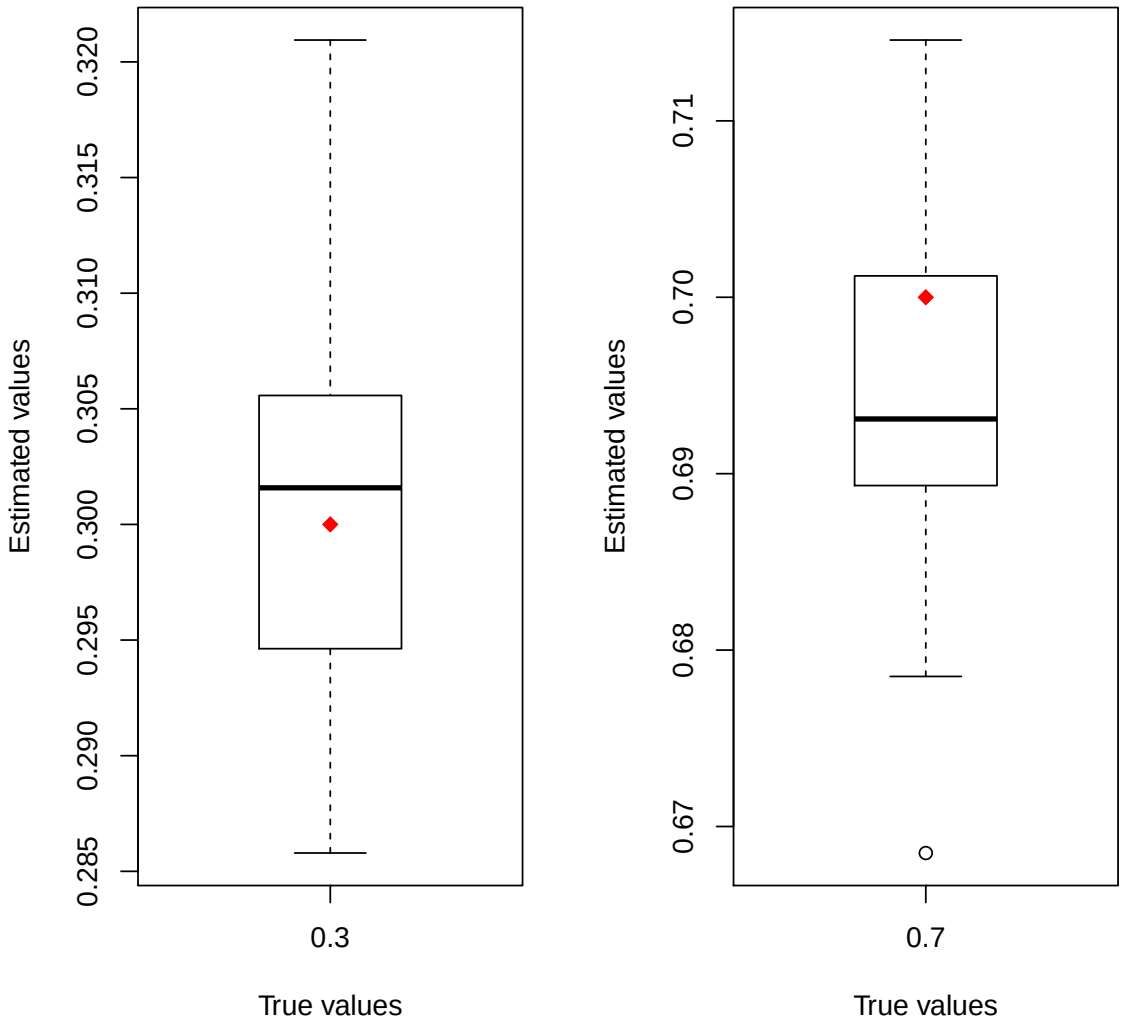


Fig. 3.7 Results of the simulated EM for π_{11} and π_{22}. The 2 boxplots correspond to the 2 different values $\pi_{11} = 0.3$ and $\pi_{22} = 0.7$. The true values are marked in red. Note that the scale is different on these 2 boxplots.

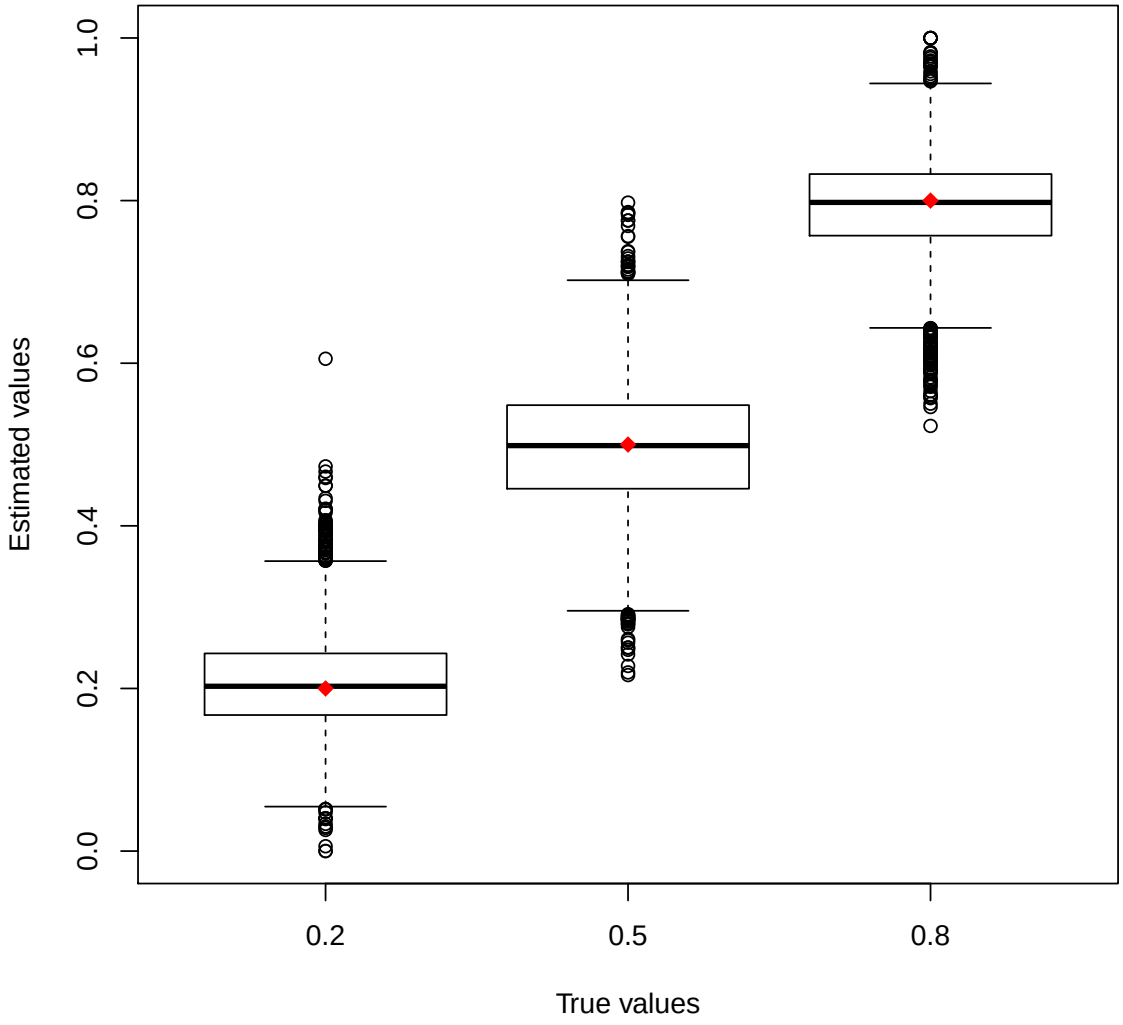


Fig. 3.8 Results of the simulated EM for π_{12}^l. The 3 boxplots correspond to the 3 different values $\pi_{12}^l = \{0.2, 0.5, 0.8\}$. The true values are marked in red. Note that each true value of the parameter π_{12}^l corresponds to 72 locations.

3.6 Proofs

Proof of Lemma 3.4.1. The mean-field like approximations of Q_1 consists in replacing the expectation with respect to the conditional distribution of the latent variables $Z^{1:L}$ given the observations $X^{1:L}$ by the expectation under the mean field like approximation (based on $\tilde{z}^{1:L}$) of this conditional distribution (see Equation (3.4.10)). It can then be written as follows

$$\begin{aligned}
&\tilde{Q}_1(\pi|\theta^{(t-1)}) \\
=&\mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\log \mathbb{P}_\pi\left(X^{1:L}|Z^{1:L}\right) \mid X^{1:L}\right] \\
=&\mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\sum_{l=1}^{L}\sum_{i<j} X^l_{ij}\log \pi^l_{Z^l_i Z^l_j} + (1-X^l_{ij})\log(1-\pi^l_{Z^l_i Z^l_j})\middle| X^{1:L}\right] \\
=&\mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\sum_{1\le q,q'\le Q}\sum_{l=1}^{L}\sum_{i<j} Z^l_{iq}Z^l_{jq'}\left(X^l_{ij}\log \pi^l_{qq'} + (1-X^l_{ij})\log(1-\pi^l_{qq'})\right)\middle| X^{1:L}\right],
\end{aligned}$$

where $\mathbb{E}^{\tilde{z}}$ stands for an expectation under the mean field approximation (with neighbours set to the values in $\tilde{z}^{1:L}$). Using the fact that the matrices π^l are symmetric ($\pi^l_{qq'} = \pi^l_{q'q}$), and so are the matrices X^l, we can rewrite the summation differently and obtain

$$\begin{aligned}
\tilde{Q}_1(\pi|\theta^{(t-1)}) = &\sum_{1\le q<q'\le Q}\sum_{l=1}^{L}\sum_{i\neq j}\Bigg\{\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_j = q' \mid X^l) \\
&\qquad\times\left(X^l_{ij}\log \pi^l_{qq'} + (1-X^l_{ij})\log(1-\pi^l_{qq'})\right)\Bigg\} \\
&+\sum_{q=1}^{Q}\sum_{l=1}^{L}\sum_{i<j}\Bigg\{\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_i = q \mid X^l)\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z^l_j = q \mid X^l) \\
&\qquad\times\left(X^l_{ij}\log \pi^l_{qq} + (1-X^l_{ij})\log(1-\pi^l_{qq})\right)\Bigg\},
\end{aligned}$$

where $\mathbb{P}^{\tilde{z}}$ stands for a probability distribution under the mean field like approximation. This quantity is written as (see Equation (3.4.10))

$$\begin{aligned}
&\mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}(Z_i^l = q \mid X^l) \\
=&\frac{\mathbb{P}_{\pi^{(t-1)}}(X^l_{i\bullet} \mid Z^l_{-i} = \tilde{z}^l_{-i}, Z^l_i = q)\mathbb{P}_{\psi^{(t-1)}}(Z_i^l = q \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)})}{\mathbb{P}_{\theta^{(t-1)}}(X^l_{i\bullet} \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)}, Z^l_{-i} = \tilde{z}^l_{-i})} \\
=&\frac{\mathbb{P}_{\pi^{(t-1)}}(X^l_{i\bullet} \mid Z^l_{-i} = \tilde{z}^l_{-i}, Z^l_i = q)\mathbb{P}_{\psi^{(t-1)}}(Z_i^l = q \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)})}{\sum_{z_i^l \in [\![1,Q]\!]} \mathbb{P}_{\pi^{(t-1)}}(X^l_{i\bullet} \mid Z^l_i = z^l_i, Z^l_{-i} = \tilde{z}^l_{-i})\mathbb{P}_{\psi^{(t-1)}}(Z_i^l = z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)})}.
\end{aligned}$$

We also write the approximation of Q_2 as follows, by replacing both the expectation with respect to the conditional distribution of the latent variables $Z^{1:L}$ given the observations $X^{1:L}$ by the expectation under the mean field like approximation (based on $\tilde{z}^{1:L}$) of this conditional distribution (see Equation (3.4.10)) and the distribution of the latent variables $Z^{1:L}$ by its approximation under the mean field like approximation (also based on $\tilde{z}^{1:L}$) (see Equation (3.4.9)),

$$\begin{aligned}
\tilde{Q}_2(\alpha, \beta | \theta^{(t-1)}) =& \mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\log \mathbb{P}^{\tilde{z}}_{\psi}\left(Z^{1:L}\right)\middle| X^{1:L}\right] \\
=& \sum_{i=1}^{n}\sum_{l=1}^{L} \mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\log \mathbb{P}_{\psi}\left(Z_i^l \mid Z_i^{\mathcal{N}_i(l)} = \tilde{z}_i^{\mathcal{N}_i(l)}\right)\middle| X^{1:L}\right] \\
=& \sum_{i=1}^{n}\sum_{l=1}^{L} \mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\log \frac{\exp\left(\alpha_{iZ_i^l} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{Z_i^l = \tilde{z}_i^{l'}}\right)}{\sum_{q=1}^{Q} \exp(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{q = \tilde{z}_i^{l'}})}\middle| X^{1:L}\right] \\
=& \sum_{i=1}^{n}\sum_{l=1}^{L} \left\{ \mathbb{E}^{\tilde{z}}_{\theta^{(t-1)}}\left[\alpha_{iZ_i^l} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{Z_i^l = \tilde{z}_i^{l'}}\middle| X^{1:L}\right] \right. \\
& \qquad \left. - \log\left[\sum_{q=1}^{Q} \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{q = \tilde{z}_i^{l'}}\right)\right]\right\} \\
=& \sum_{i=1}^{n}\sum_{l=1}^{L} \left\{ \left(\sum_{q=1}^{Q} \alpha_{iq} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}\left(Z_i^l = q\middle| X^l\right)\right) + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{P}^{\tilde{z}}_{\theta^{(t-1)}}\left(Z_i^l = \tilde{z}_i^{l'}\middle| X^l\right) \right. \\
& \qquad \left. - \log\left[\sum_{q=1}^{Q} \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \mathbb{1}_{q = \tilde{z}_i^{l'}}\right)\right]\right\}.
\end{aligned}$$

□

Conclusions and perspectives

The aim of this thesis was to study and propose methods for the analysis of graphs and more precisely for node clustering, in the context of multiple networks with a space or time dependency. In particular, we studied the consistency of parameter estimators in a dynamic SBM and proposed a spatial SBM together with a method to estimate its parameter.

In Chapter 2, we obtained consistency results for the maximum likelihood and variational estimators in a dynamic version of the SBM based on hidden Markov chains under certain conditions. This follows the work of Celisse et al. (2012) who obtained consistency results in the static SBM. However, these conditions exclude the sparse case (as in Celisse et al. (2012)), and are therefore quite restrictive since many large real-world network exhibit sparsity, and the study of this case would be of great interest. In Bickel et al. (2013), in the static case, they introduce a density parameter defined as $\rho = \mathbb{P}(X_{ij} = 1)$ and analyse the asymptotic behaviour when $\rho \equiv \rho_n \to 0$ when n increases. Note that we tried generalising the work of Bickel et al. (2013) to the dynamic SBM, but have not managed to obtain results yet.

Moreover, we obtained bounds for the rates of convergence. For the case where the connectivity parameter is fixed over time, we proved that the estimators of the connectivity parameter converges faster than $r_{n,T}/n^{1/4}$ with $\{r_{n,T}\}_{n,T\geq 1}$ any sequence increasing to infinity, and that the estimators of the transition matrix converges faster than $r_{n,T}\sqrt{\log n}/\sqrt{n}$ with $\{r_{n,T}\}_{n,T\geq 1}$ any sequence increasing to infinity. We believe that these rates are not optimal, and that they should be at least as good as those obtained in Bickel et al. (2013) in the static case, when their density parameter is constant (dense setup), which is n^{-1} for the connectivity parameter, and $n^{-1/2}$ for the parameter of the distribution of the latent variables. Furthermore, our consistency result for the transition matrix estimators requires an additional assumption that the connectivity parameter estimators converges at a rate that is $o(\sqrt{\log(nT)}/n)$. Obtaining a sharper bound as mentioned before for the connectivity parameter estimators would then solve the problem

of this additional assumption.

In Chapter 3, we introduced a spatial SBM based on hidden Markov random fields and proposed an algorithm based on the simulated EM of Celeux et al. (2003) for the estimation of parameters in this model. While it gave promising results on synthetic data in the case where $Q = 2$, more experiments have to be conducted, namely with a larger number of classes, or with more general location graphs than lattices. Moreover, we have been interested in parameter estimation, but we would also like to know if we can recover a satisfying clustering of the nodes at each location. Such estimated clustering could be obtained using the fact that we can simulate according to the conditional distribution $\mathbb{P}_{\tilde{\theta}}(Z^{1:L} \mid X^{1:L})$ of the latent variables given the observations, under the estimated parameter $\tilde{\theta}$ (outputted by the algorithm). Experiments should then be conducted, comparing an estimated clustering with the true one, using for example the Adjusted Rand Index (Hubert and Arabie, 1985).

Moreover, we would like to study the effect of the strength of interaction on the quality of parameter estimation. Indeed, for large absolute values of the interaction strength parameter β, we tend to observe large parts or the whole Markov random field equal to the same value (for positive values) or for example a check pattern on a first order lattice (for negative values) (see Figure 1.13). We can then reasonably think that large absolute values of β make the estimation task harder.

We also obtained the generic identifiability of the model parameters under certain conditions. These conditions exclude the affiliation case, and we gave a counter-example (in Section 3.3) to show that without further assumptions, the parameters are in fact not identifiable in this case. However, from the proof of non-identifiability of our particular counter-example, we may wonder if we could obtain the identifiability by imposing for example a non-negativity condition on β, different values for the parameter α in all the groups, and/or to have at least a triangle (or cycle of odd length) in the location graph. Further investigation has to be done for the identifiability in the particular affiliation case.

In addition, we would like to propose a model selection criterion for the choice of the number of classes. Some simple possibility would be to approximate the BIC based on the mean field like approximation as in Forbes and Peyrard (2003) (see Section 1.7.9). However the results obtained with this criterion have been found to be unstable. Another option would be to adapt the criterion proposed by Forbes and Peyrard (2003) based on an approximation of the partition function (see also Section 1.7.9). We plan to investigate

a combination of this approach with criteria used for the SBM (see Section 1.4.5.2), such as the ICL of Daudin et al. (2008).

We would also like to extend the model to make it more flexible. A quite straightforward extension would be to allow the set of nodes to be (a bit) different at the different locations. A less straightforward one would be to allow the model to have different numbers of classes at different locations. This would however be problematic for obtaining a satisfying criterion for the choice of the number of classes.

It would also be interesting to take into account the asymptotic behaviour of the obtained algorithm (as it does not seem to converge to a fixed value) in order to obtain a smarter stopping rule. Indeed, the obtained sequence of estimators (and hence of our stopping criterion) is not "stable" and looking at the difference of the criterion between two time steps is probably not optimal. Moreover, the convergence of this algorithm could be studied.

We could compare the performance of our method with that of separate SBM for each location (i.e. without taking into account the interaction between the locations), as no other methods for such spatial network data has been introduced. In particular, we could compare the performance for different values of the strength of interaction, as we expect our method to outperform a method considering the graphs at each location separately when there is a significant interaction between the locations.

Finally, we would like to apply the method to a real dataset, with "expert" information to determine the location graphs for each species.

References

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86.

Abbe, E., Bandeira, A. S., and Hall, G. (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487.

Agarwal, A. and Xue, L. (2019). Model-based clustering of nonparametric weighted networks with application to water pollution analysis. *Technometrics*, 0(0):1–21.

Aicher, C., Jacobs, A. Z., and Clauset, A. (2013). Adapting the stochastic block model to edge-weighted networks. *ICML Workshop on Structured Learning*.

Aiello, W., Chung, F., and Lu, L. (2001). Random evolution in massive graphs. In *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*, pages 510–519. IEEE Computer Soc., Los Alamitos, CA.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014.

Airoldi, E. M., Costa, T. B., and Chan, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory (Petrov, B. N. and Csaki, F., eds.)*, pages 267–281. Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581 – 598.

Allman, E. S., Matias, C., and Rhodes, J. A. (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, 141(5):1719–1736.

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Asuncion, A., Liu, Q., Ihler, A., and Smyth, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 33–40, Chia Laguna Resort, Sardinia, Italy. PMLR.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 21–30, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bajardi, P., Poletto, C., Balcan, D., Hu, H., Goncalves, B., Ramasco, J. J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W., et al. (2009). Modeling vaccination campaigns and the fall/winter 2009 activity of the new A (H1N1) influenza in the northern hemisphere. *Emerging Health Threats Journal*, 2(1):7093.

Balcan, D., Hu, H., Goncalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., Paolotti, D., Perra, N., Tizzoni, M., Van den Broeck, W., et al. (2009). Seasonal transmission potential and activity peaks of the new influenza A (H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC medicine*, 7(1):45.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barrat, A. and Cattuto, C. (2013). Temporal networks of face-to-face human interactions. In *Temporal Networks*, pages 191–216. Springer.

Bartolucci, F., Marino, M. F., and Pandolfi, S. (2018). Dealing with reciprocity in dynamic stochastic block models. *Comput. Stat. Data Anal.*, 123(C):86–100.

Bauke, H. (2007). Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B*, 58(2):167–173.

Beal, M. J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 453–463. Oxford Univ. Press, New York.

Becker, A. and Holzmann, H. (2018). Nonparametric identification in the dynamic stochastic block model. *IEEE Transactions on Information Theory*, 65:4335–4344.

Berge, C. (1976). *Graphs and hypergraphs, revised ed.* North-Holland Publishing Co., Amsterdam.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(3):179–195.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):75–83.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.*, 41(4):1922–1943.

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.

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.

Biernacki, C., Celeux, G., and Govaert, G. (1998). Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. Technical Report RR-3521, INRIA.

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.

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.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Blundell, C., Beck, J., and Heller, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2600–2608. Curran Associates, Inc.

Borgs, C., Chayes, J., Cohn, H., and Zhao, Y. (2019). An l^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062.

Borgs, C., Chayes, J. T., Lovász, L., Sós, V. T., and Vesztergombi, K. (2008). Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. *Advances in Mathematics*, 219(6):1801–1851.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Brault, V. and Channarond, A. (2016). Fast and consistent algorithm for the latent block model. *arXiv preprint arXiv:1610.09005*.

Brault, V., Keribin, C., and Mariadassou, M. (2020). Consistency and Asymptotic Normality of Latent Block Model Estimators. *Electron. J. Statist.* To appear.

Brault, V. and Mariadassou, M. (2015). Co-clustering through latent block model: A review. *Journal de la Société Française de Statistique*, 156(3):120–139.

Brémaud, P. (2013). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media.

Britton, T., Deijfen, M., and Martin-Löf, A. (2006). Generating simple random graphs with prescribed degree distribution. *J. Stat. Phys.*, 124(6):1377–1397.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, 10(1):1–10.

Bubeck, S. (2010). *Jeux de bandits et fondations du clustering*. PhD thesis, Université Lille 1.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1):155–200.

Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41 – 55.

Celeux, G., Forbes, F., and Peyrard, N. (2001). EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation. Research Report RR-4105, INRIA.

Celeux, G., Forbes, F., and Peyrard, N. (2002). EM-based image segmentation using Potts models with external field. Research Report RR-4456, INRIA.

Celeux, G., Forbes, F., and Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern recognition*, 36(1):131–144.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.*, 6:1847–1899.

Chakraborty, T., Dalmia, A., Mukherjee, A., and Ganguly, N. (2017). Metrics for community analysis: A survey. *ACM Computing Surveys (CSUR)*, 50(4):54.

Chalmond, B. (1989). An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6):747 – 761.

Chan, S. and Airoldi, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216.

Chandler, D. (1987). *Introduction to modern statistical mechanics*. Oxford University Press.

Channarond, A., Daudin, J.-J., and Robin, S. (2012). Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electron. J. Statist.*, 6:2574–2601.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214.

Chatterjee, S., Diaconis, P., et al. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

Chen, T., Singh, P., and Bassler, K. E. (2018). Network community detection using modularity density measures. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(5):053406.

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284.

Chung, F. R. and Graham, F. C. (1997). *Spectral graph theory*. Number 92. American Mathematical Soc.

Chung, K.-M., Lam, H., Liu, Z., and Mitzenmacher, M. (2012). Chernoff-Hoeffding bounds for Markov chains: generalized and simplified. In Dürr, C. and Wilke, T., editors, *29th International Symposium on Theoretical Aspects of Computer Science (STACS 2012)*, volume 14 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 124–135, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703.

Clifford, P. (1990). Markov random fields in statistics. *Disorder in physical systems: A volume in honour of John M. Hammersley*, 19.

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.

Comets, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.*, 20(1):455–468.

Corneli, M., Latouche, P., and Rossi, F. (2016). Block modelling in dynamic networks with non-homogeneous Poisson processes and exact ICL. *Social Network Analysis and Mining*, 6(1):55.

Cucala, L. and Marin, J.-M. (2013). Bayesian inference on a mixture model with spatial dependence. *Journal of Computational and Graphical Statistics*, 22(3):584–597.

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183.

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.

Delmas, E., Besson, M., Brice, M.-H., Burkle, L. A., Dalla Riva, G. V., Fortin, M.-J., Gravel, D., Guimarães Jr., P. R., Hembry, D. H., Newman, E. A., Olesen, J. M., Pires, M. M., Yeakel, J. D., and Poisot, T. (2019). Analysing ecological networks of species interactions. *Biological Reviews*, 94(1):16–36.

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):1–38.

Descombes, X., Morris, R. D., Zerubia, J., and Berthod, M. (1999). Estimation of Markov random field prior parameters using Markov chain Monte Carlo maximum likelihood. *IEEE Transactions on Image Processing*, 8(7):954–963.

Desmarais, B. A. and Cranmer, S. J. (2012). Statistical inference for valued-edge networks: The generalized exponential random graph model. *PloS one*, 7(1).

Devi, J. C. and Poovammal, E. (2016). An analysis of overlapping community detection algorithms in social networks. *Procedia Computer Science*, 89:349 – 358.

Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl., VII. Ser.*, 28(1):33–61.

Diebolt, J. and Celeux, G. (1993). Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Communications in Statistics. Stochastic Models*, 9(4):599–613.

DuBois, C., Butts, C., and Smyth, P. (2013). Stochastic blockmodeling of relational event dynamics. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31, pages 238–246. PMLR.

Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Physical review E*, 72(2):027104.

Duminil-Copin, H. (2015). Order/disorder phase transitions: the example of the Potts model. *Current developments in mathematics*, 2015(1):27–71.

Durante, D., Dunson, D. B., et al. (2016). Locally adaptive dynamic networks. *The Annals of Applied Statistics*, 10(4):2203–2232.

Ebel, H., Mielsch, L.-I., and Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical review E*, 66(3):035103.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.*, 3(6):1189–1242.

Efron, B. (1978). The geometry of exponential families. *Ann. Statist.*, 6(2):362–376.

Erdős, P. and Gallai, T. (1961). Graphs with points of prescribed degree. (Graphen mit Punkten vorgeschriebenen Grades.). *Mat. Lapok*, 11:264–274.

Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, 6:290–297.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Everitt, R. G. (2012). Bayesian parameter estimation for latent Markov random fields and social networks. *Journal of Computational and graphical Statistics*, 21(4):940–960.

Feodor Nielsen, S. (2000). The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489.

Forbes, F. and Fort, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Transactions on Image Processing*, 16(3):824–837.

Forbes, F. and Peyrard, N. (2003). Hidden Markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101.

Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1 – 44. Community detection in networks: A user guide.

Frank, O. and Harary, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77(380):835–840.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Friel, N. (2012). Bayesian inference for Gibbs random fields using composite likelihoods. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–8. IEEE.

Friel, N. and Pettitt, A. N. (2004). Likelihood estimation and inference for the autologistic model. *Journal of Computational and Graphical Statistics*, 13(1):232–246.

Friel, N., Pettitt, A. N., Reeves, R., and Wit, E. (2009). Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics*, 18(2):243–261.

Friel, N., Rastelli, R., Wyse, J., and Raftery, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences*, 113(24):6629–6634.

Friel, N. and Rue, H. (2007). Recursive computing and simulation-free inference for general factorizable models. *Biometrika*, 94(3):661–672.

Frieze, A. and Karoński, M. (2016). *Introduction to random graphs*. Cambridge University Press.

Gao, C., Lu, Y., and Zhou, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.*, 43(6):2624–2652.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017a). Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024.

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.*, 46(5):2153–2185.

Gao, F., van der Vaart, A., Castro, R., and van der Hofstad, R. (2017b). Consistent estimation in general sublinear preferential attachment trees. *Electron. J. Statist.*, 11(2):3979–3999.

Gao, F. and van der Vaart, A. (2017). On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees. *Stochastic Processes and their Applications*, 127(11):3754 – 3775.

Gaucher, S. and Klopp, O. (2019). Maximum likelihood estimation of sparse networks with missing observations. Technical report, manuscript.

Gazal, S., Daudin, J.-J., and Robin, S. (2012). Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation*, 82(6):849–862.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Georgii, H.-O. (2011). *Gibbs measures and phase transitions*, volume 9. Walter de Gruyter.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156 – 163. Interface Foundation of North America.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3):657–699.

Gidas, B. (1988). Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. In Fleming, W. and Lions, P.-L., editors, *Stochastic Differential Systems, Stochastic Control Theory and Applications*, pages 129–145, New York, NY. Springer New York.

Gilbert, E. N. (1959). Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233.

Goldstein, M. L., Morris, S. A., and Yen, G. G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258.

Good, B. H., De Montjoye, Y.-A., and Clauset, A. (2010). Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106.

Goodman, L. A. (1961). Snowball sampling. *Ann. Math. Statist.*, 32(1):148–170.

Govaert, G. (2003). *Analyse des données*, volume 12. Lavoisier.

Govaert, G. and Nadif, M. (2008). Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233 – 3245.

Govaert, G. and Nadif, M. (2013). *Co-clustering: models, algorithms and applications*. John Wiley & Sons.

Guimera, R. and Amaral, L. A. N. (2005a). Cartography of complex networks: modules and universal roles. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(02):P02001.

Guimera, R. and Amaral, L. A. N. (2005b). Functional cartography of complex metabolic networks. *nature*, 433(7028):895.

Gunawardana, A. and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *Journal of machine learning research*, 6:2049–2073.

Guyon, X. and Künsch, H. R. (1992). Asymptotic comparison of estimators in the Ising model. In *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, pages 177–198. Springer.

Hajek, B., Wu, Y., and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937.

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.

Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In Breiger, R., Carley, K., and Pattison, P., editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 229–240, Washington, DC. The National Academies Press.

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.

Handcock, M. S., Robins, G., Snijders, T., Moody, J., and Besag, J. (2003). Assessing degeneracy in statistical models of social networks. *Journal of the American Statistical Association*, 76:33–50.

Hanneke, S., Fu, W., and Xing, E. P. (2010). Discrete temporal models of social networks. *Electron. J. Statist.*, 4:585–605.

Hanneke, S. and Xing, E. P. (2007). Discrete temporal models of social networks. In Airoldi, E., Blei, D. M., Fienberg, S. E., Goldenberg, A., Xing, E. P., and Zheng, A. X., editors, *Statistical Network Analysis: Models, Issues, and New Directions*, pages 115–125. Springer Berlin Heidelberg.

Haq, N. F., Moradi, M., and Wang, Z. J. (2019). Community structure detection from networks with weighted modularity. *Pattern Recognition Letters*, 122:14 – 22.

Haraldsdottir, S., Gupta, S., and Anderson, R. M. (1992). Preliminary studies of sexual networks in a male homosexual community in Iceland. *Journal of acquired immune deficiency syndromes*, 5(4):374–381.

He, J., Chen, D., and Chongjing Sun (2016). A fast simulated annealing strategy for community detection in complex networks. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 2380–2384.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3(5):1163–1174.

Ho, Q., Song, L., and Xing, E. P. (2011). Evolving cluster mixed-membership blockmodel for time-varying networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *JMLR: W&CP*, San Diego, CA, USA.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pages 657–664.

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.

Hofstad, R. v. d. (2016). *Random Graphs and Complex Networks*, volume 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137.

Holland, P. W. and Leinhardt, S. (1977). A dynamic model for social networks. *The Journal of Mathematical Sociology*, 5(1):5–20.

Holme, P. (2015). Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):234.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*, 2.

Huang, F. and Ogata, Y. (1999). Improvements of the maximum pseudo-likelihood estimators in various spatial statistical models. *Journal of Computational and Graphical Statistics*, 8(3):510–530.

Huang, F. and Ogata, Y. (2002). Generalized pseudo-likelihood estimates for Markov random fields on lattice. *Annals of the Institute of Statistical Mathematics*, 54(1):1–18.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hunter, D. R. (2007). Curved exponential family models for social networks. *Social networks*, 29(2):216–230.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.

Hunter, D. R. and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.

Ikeda, N. (2009). Estimation of power-law exponent of degree distribution using mean vertex degree. *Modern Physics Letters B*, 23(17):2073–2088.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.

Jiang, Q., Zhang, Y., and Sun, M. (2009). Community detection on weighted networks: A variational Bayesian method. In Zhou, Z.-H. and Washio, T., editors, *Advances in Machine Learning*, pages 176–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Junuthula, R., Haghdan, M., Xu, K. S., and Devabhaktuni, V. (2019). The block point process model for continuous-time event-based dynamic networks. In *The World Wide Web Conference*, pages 829–839.

Kallenberg, O. (2006). *Probabilistic symmetries and invariance principles.* Springer Science & Business Media.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107.

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.

Keribin, C., Brault, V., Celeux, G., Govaert, G., et al. (2012). Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, volume 2012.

Keribin, C., Govaert, G., and Celeux, G. (2010). Estimation d'un modèle à blocs latents par l'algorithme SEM. In *42èmes Journées de Statistique*, Marseille, France, France.

Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018). A review of dynamic network models with latent variables. *Statist. Surv.*, 12:105–135.

Klaus, A., Yu, S., and Plenz, D. (2011). Statistical analyses support power law distributions found in neuronal avalanches. *PLOS ONE*, 6(5):1–12.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: measurements, models, and methods. In *International Computing and Combinatorics Conference*, pages 1–17. Springer.

Klopp, O., Tsybakov, A. B., and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.*, 45(1):316–354.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and Models*. Springer.

Kolaczyk, E. D. (2017). *Topics at the Frontier of Statistics and Network Analysis:(re) visiting the Foundations*. Cambridge University Press.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Krivitsky, P. N. (2012). Exponential-family random graph models for valued networks. *Electron. J. Statist.*, 6:1100–1128.

Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281 – 293.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18(2):95 – 138.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. (2000). Stochastic models for the web graph. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 57–65. IEEE.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Extracting large-scale knowledge bases from the web. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, page 639–650, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lai, T. L. and Lim, J. (2015). Asymptotically efficient parameter estimation in hidden markov spatio-temporal random fields. *Statistica Sinica*, pages 403–421.

Lakshmanan, S. and Derin, H. (1989). Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):799–813.

Lancichinetti, A. and Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical review E*, 84(6):066122.

Latapy, M., Rotenberg, E., Crespelle, C., and Tarissan, F. (2017). Rigorous measurement of the internet degree distribution. *Complex Systems*, 26(1):1–29.

Latapy, M., Viard, T., and Magnien, C. (2018). Stream Graphs and Link Streams for the Modeling of Interactions over Time. *Social Networks Analysis and Mining*, 8(1):61:1–61:29.

Latouche, P., Birmelé, E., and Ambroise, C. (2010). Bayesian methods for graph clustering. In Fink, A., Lausen, B., Seidel, W., and Ultsch, A., editors, *Advances in Data Analysis, Data Handling and Business Intelligence*, pages 229–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Latouche, P., Birmele, E., and Ambroise, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Ann. Appl. Stat.*, 5(1):309–336.

Latouche, P. and Robin, S. (2016). Variational Bayes model averaging for graphon functions and motif frequencies inference in w-graph models. *Statistics and Computing*, 26(6):1173–1185.

Lauritzen, S. L. (1996). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York.

Lee, J., Heaukulani, C., Ghahramani, Z., James, L. F., and Choi, S. (2017). Bayesian inference on random simple graphs with power law degree distributions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2004–2013. JMLR.org.

Lee, K. H., Xue, L., and Hunter, D. R. (2020). Model-based clustering of time-evolving networks through temporal exponential-family random graph models. *Journal of Multivariate Analysis*, 175:104540.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237.

Li, S., Chen, Y., Du, H., and Feldman, M. W. (2010). A genetic algorithm with local search strategy for improved detection of community structure. *Complexity*, 15(4):53–60.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80:221–239.

Liu, F., Choi, D., Xie, L., and Roeder, K. (2018). Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 115(5):927–932.

Liu, J. S. (1996). Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83(3):681–682.

Lloyd, J., Orbanz, P., Ghahramani, Z., and Roy, D. M. (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 998–1006. Curran Associates, Inc.

Long, H. (2019). Edge intensity-based community measurement in complex networks. *Physics Letters A*, 383(11):1167 – 1173.

Longepierre, L. and Matias, C. (2019). Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model. *Electron. J. Statist.*, 13(2):4157–4223.

Lovász, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Soc.

Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957.

Luczak, T. (1990). On the equivalence of two basic models of random graph. In *Proceedings of Random graphs'87, (Michal Karonski, Jerzy Jaworski, Andrzej Rucinski, eds.)*, volume 87, pages 151–159. Wiley, Chichester.

Lusher, D., Koskinen, J., and Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press.

Mariadassou, M. and Matias, C. (2015). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: A variational approach. *Ann. Appl. Stat.*, 4(2):715–742.

Martínez, V., Berzal, F., and Cubero, J.-C. (2016). A survey of link prediction in complex networks. *ACM Computing Surveys (CSUR)*, 49(4):1–33.

Mase, S. (1995). Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes. *Ann. Appl. Probab.*, 5(3):603–612.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin.

Massoulié, L. (2014). Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703.

Masuda, N., Takaguchi, T., Sato, N., and Yano, K. (2013). Self-exciting point process modeling of conversation event sequences. In *Temporal Networks*, pages 245–264. Springer.

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(4):1119–1141.

Matias, C., Rebafka, T., and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680.

McGrory, C. A., Pettitt, A. N., Reeves, R., Griffin, M., and Dwyer, M. (2012). Variational Bayes and the reduced dependence approximation for the autologistic model on an irregular grid with applications. *Journal of Computational and Graphical Statistics*, 21(3):781–796.

McGrory, C. A., Titterington, D. M., Reeves, R., and Pettitt, A. N. (2009). Variational Bayes for estimating the parameters of a hidden Potts model. *Statistics and Computing*, 19(3):329.

Medus, A., Acuña, G., and Dorso, C. O. (2005). Detection of community structures in networks via global optimization. *Physica A: Statistical Mechanics and its Applications*, 358(2):593 – 604.

Milgram, S. (1967). The small world problem. *Psychology today*, 2(1):60–67.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

Moreno, J. L. and Jennings, H. H. (1938). Statistics of social configurations. *Sociometry*, 1(3/4):342–374.

Mossel, E., Neeman, J., and Sly, A. (2012). Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*.

Mossel, E., Neeman, J., and Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461.

Mossel, E., Neeman, J., and Sly, A. (2018). A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708.

Mukherjee, S. (2020). Degeneracy in sparse ERGMs with functions of degrees as sufficient statistics. *Bernoulli*, 26(2):1016–1043.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 359–366, Arlington, Virginia, USA. AUAI Press.

Nascimento, M. C. and De Carvalho, A. C. (2011). Spectral methods for graph clustering–a survey. *European Journal of Operational Research*, 211(2):221–231.

Nettasinghe, B. and Krishnamurthy, V. (2019). Maximum likelihood estimation of power-law degree distributions using friendship paradox based sampling. *arXiv preprint arXiv:1908.00310*.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087.

Okabayashi, S., Johnson, L., and Geyer, C. J. (2011). Extending pseudo-likelihood for Potts models. *Statistica Sinica*, pages 331–347.

Orbanz, P. and Roy, D. M. (2014). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):437–461.

Pattison, P. and Wasserman, S. (1999). Logit models and logistic regressions for social networks: II. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193.

Pattison, P. E., Robins, G. L., Snijders, T. A., and Wang, P. (2013). Conditional estimation of exponential random graph models from snowball sampling designs. *Journal of Mathematical Psychology*, 57(6):284 – 296. Social Networks.

Paul, S. and Chen, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Statist.*, 10(2):3807–3870.

Pensky, M. (2019). Dynamic network models and graphon estimation. *Ann. Statist.*, 47(4):2378–2403.

Pensky, M., Zhang, T., et al. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709.

Pettitt, A. N., Friel, N., and Reeves, R. (2003). Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):235–246.

Peyrard, N. (2001). *Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales*. PhD thesis, Université Joseph Fourier Grenoble 1.

Pieczynski, W. (1992). Statistical image segmentation. *Machine graphics and vision*, 1(1/2):261–268.

Pieczynski, W. (1994). Champs de Markov cachés et estimation conditionnelle itérative. *Traitement du Signal*, 11(2):141–153.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In Yolum, p., Güngör, T., Gürgen, F., and Özturan, C., editors, *Computer and Information Sciences - ISCIS 2005*, pages 284–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Potamianos, G. and Goutsias, J. (1997). Stochastic approximation algorithms for partition function estimation of Gibbs random fields. *IEEE Trans. Inform. Theory*, 43(6):1948–1965.

Potts, R. B. (1952). Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pages 106–109. Cambridge University Press.

Prasse, B., Achterberg, M. A., Ma, L., and Van Mieghem, P. (2020). Network-Based Prediction of the 2019-nCoV Epidemic Outbreak in the Chinese Province Hubei. *arXiv e-prints*, page arXiv:2002.04482.

Qian, W. and Titterington, D. (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):407–428.

Ranalli, M., Lagona, F., Picone, M., and Zambianchi, E. (2018). Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):575–598.

Reeves, R. and Pettitt, A. N. (2004). Efficient recursions for general factorisable models. *Biometrika*, 91(3):751–757.

Rhodes, J. A. (2010). A concise proof of Kruskal's theorem on tensor decomposition. *Linear Algebra and its Applications*, 432(7):1818 – 1824.

Rinaldo, A., Fienberg, S. E., and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Statist.*, 3:446–484.

Robins, G., Elliott, P., and Pattison, P. (2001). Network models for social selection processes. *Social Networks*, 23(1):1 – 30.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a). An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191.

Robins, G., Pattison, P., and Wasserman, S. (1999). Logit models and logistic regressions for social networks: III. valued relations. *Psychometrika*, 64(3):371–394.

Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b). Recent developments in exponential random graph (p*) models for social networks. *Social networks*, 29(2):192–215.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.*, 39(4):1878–1915.

Salter-Townshend, M. and Murphy, T. B. (2015). Role analysis in networks using mixtures of exponential random graph models. *Journal of Computational and Graphical Statistics*, 24(2):520–538.

Saramäki, J. and Moro, E. (2015). From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *The European Physical Journal B*, 88(6):164.

Sarkar, P. and Moore, A. W. (2006). Dynamic social network analysis using latent space models. In *Advances in Neural Information Processing Systems*, pages 1145–1152.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370.

Schweinberger, M. (2020). Consistent structure estimation of exponential-family random graph models with block structure. *Bernoulli*, 26(2):1205–1233.

Schweinberger, M. and Handcock, M. S. (2015). Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):647–676.

Schweinberger, M. and Stewart, J. (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *Ann. Statist.*, 48(1):374–396.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

Sewell, D. K. and Chen, Y. (2016). Latent space models for dynamic networks with weighted edges. *Social Networks*, 44:105 – 116.

Sizemore, A. E. and Bassett, D. S. (2018). Dynamic graph metrics: Tutorial, toolbox, and tale. *NeuroImage*, 180:417 – 427. Brain Connectivity Dynamics.

Snijders, T., Pattison, P., Robins, G., and Handcock, M. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99 – 153.

Snijders, T. A. (1996). Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172.

Snijders, T. A. (2001). The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395.

Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Stanford, D. C. (1999). *Fast automatic unsupervised image segmentation and curve detection in spatial point patterns*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of Washington.

Stanford, D. C. and Raftery, A. E. (2002). Approximate Bayes factors for image segmentation: the pseudolikelihood information criterion (PLIC). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(11):1517–1520.

Stivala, A., Robins, G., and Lomi, A. (2020). Exponential random graph model parameter estimation for very large directed networks. *PloS one*, 15(1):e0227804.

Stivala, A. D., Koskinen, J. H., Rolls, D. A., Wang, P., and Robins, G. L. (2016). Snowball sampling for estimating exponential random graph models for large networks. *Social Networks*, 47:167 – 188.

Stoehr, J. (2017). A review on statistical inference methods for discrete Markov random fields. *arXiv e-prints*, page arXiv:1704.03331.

Stoehr, J. and Friel, N. (2015). Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 921–929, San Diego, California, USA. PMLR.

Stoehr, J., Marin, J.-M., and Pudlo, P. (2016). Hidden Gibbs random fields model selection using block likelihood information criterion. *Stat*, 5(1):158–172.

Stoehr, J., Pudlo, P., and Cucala, L. (2015). Adaptive ABC model choice and geometric summary statistics for hidden Gibbs random fields. *Statistics and Computing*, 25(1):129–141.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American statistical association*, 85(409):204–212.

Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical review letters*, 58(2):86.

Tallberg, C. (2004). A bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1):1–23.

Traag, V. A. (2015). Faster unfolding of communities: Speeding up the Louvain algorithm. *Physical Review E*, 92(3):032801.

van der Pol, J. (2019). Introduction to network modeling using exponential random graph models (ERGM): Theory and an application using R-project. *Computational Economics*, 54(3):845–875.

Van Duijn, M. A., Gile, K. J., and Handcock, M. S. (2009). A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. *Social Networks*, 31(1):52–62.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Viger, F. and Latapy, M. (2005). Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *International Computing and Combinatorics Conference*, pages 440–449. Springer.

Vignes, M. (2007). *Modèles markoviens graphiques pour la fusion de données individuelles et d'interactions: application à la classification de gènes*. PhD thesis, Université Joseph Fourier Grenoble 1.

Vignes, M. and Forbes, F. (2009). Gene clustering via integrated Markov models combining individual and pairwise features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(2):260–270.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013). Model-based clustering of large networks. *Ann. Appl. Stat.*, 7(2):1010–1039.

Wakita, K. and Tsurumi, T. (2007). Finding community structure in mega-scale social networks. In *Proceedings of the 16th international conference on World Wide Web*, pages 1275–1276. ACM.

Wang, T. and Resnick, S. I. (2019). Consistency of Hill estimators in a linear preferential attachment model. *Extremes*, 22(1):1–28.

Wang, Y., Fang, H., Yang, D., Zhao, H., and Deng, M. (2019). Network clustering analysis using mixture exponential-family random graph models and its application in genetic interaction data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 16(5):1743–1752.

Wasserman, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75(370):280–294.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. an introduction to Markov graphs andp. *Psychometrika*, 61(3):401–425.

Wasserman, S. and Robins, G. (2005). An introduction to random graphs, dependence graphs, and p*. In Carrington, P. J., Scott, J., and Wasserman, S., editors, *Models and Methods in Social Network Analysis*, Structural Analysis in the Social Sciences, page 148–161. Cambridge University Press.

Watts, C. H. and May, R. M. (1992). The influence of concurrent partnerships on the dynamics of HIV/AIDS. *Mathematical Biosciences*, 108(1):89 – 104.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, 393(6684):440.

Wu, C.-h. and Doerschuk, P. C. (1995). Cluster expansions for the deterministic computation of Bayesian estimators based on Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):275–293.

Wyse, J. and Friel, N. (2012). Block clustering with collapsed latent block models. *Statistics and Computing*, 22(2):415–428.

Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):1–35.

Xing, E. P., Fu, W., and Song, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.*, 4(2):535–566.

Xu, K. (2015). Stochastic block transition models for dynamic networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *JMLR: W&CP*, San Diego, CA, USA.

Xu, K. S. and Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562.

Yang, J., Han, C., and Airoldi, E. (2014). Nonparametric estimation and testing of exchangeable graph models. In *Artificial Intelligence and Statistics*, pages 1060–1067.

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.

Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Scientific reports*, 6:30750.

Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Annales de l'I.H.P. Probabilités et statistiques*, 24(2):269–294.

Zhang, A. Y. and Zhou, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.*, 44(5):2252–2280.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583.

Zhang, X., Moore, C., and Newman, M. E. J. (2017a). Random graph models for dynamic networks. *The European Physical Journal B*, 90(10):200.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57.

Zhang, Y., Levina, E., and Zhu, J. (2017b). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783.

Zhao, Y., Levina, E., and Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108(18):7321–7326.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292.

Zreik, R., Latouche, P., and Bouveyron, C. (2016). The dynamic random subgraph model for the clustering of evolving networks. *Computational Statistics*.

Zweig, K. A. et al. (2016). *Network analysis literacy: A Practical Approach to the Analysis of Networks*. Springer.

Appendix A

Supplementary material for Chapter 2

A.1 Proofs of main results for the finite time case

A.1.1 Proof of Corollary 2.3.2

When the number of time steps is fixed and the connection probabilities vary over time, the conditional log-likelihood is

$$\ell_c^T(\theta; Z^{1:T}) = \sum_{t=1}^{T} \sum_{1 \le i < j \le n} X_{ij}^t \log \pi_{Z_i^t Z_j^t}^t + (1 - X_{ij}^t) \log(1 - \pi_{Z_i^t Z_j^t}^t)$$

and the likelihood $\ell^T(\theta)$ is defined as in (2.2.2) with $\ell_c^T(\cdot)$ instead of $\ell_c(\cdot)$. The maximum likelihood estimator is then

$$\hat{\theta} = (\hat{\Gamma}, \hat{\pi}^{1:T}) = \arg\max_{\theta \in \Theta^T} \ell^T(\theta).$$

As before, we denote the normalized log-likelihood $M_{n,T}(\Gamma, \pi^{1:T}) = 2/(n(n-1)T)\ell^T(\theta)$. We introduce the following limiting quantity

$$\mathbb{M}^T(\pi^{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{M}(\pi^t) = \frac{1}{T} \sum_{t=1}^{T} \sup_{A \in \mathcal{A}} \mathbb{M}(\pi^t, A).$$

We follow the lines of the proof of Theorem 2.3.1 in order to prove that we have for any sequence $y_n \to +\infty$, for all $\epsilon > 0$

$$\mathbb{P}_{\theta^*}\left(\sup_{(\Gamma,\pi^{1:T})\in\Theta^T}\left|M_{n,T}(\Gamma,\pi^{1:T}) - \mathbb{M}^T(\pi^{1:T})\right| > \frac{\epsilon y_n}{\sqrt{n}}\right) \underset{n\to+\infty}{\longrightarrow} 0. \tag{A.1.1}$$

Choosing $y_n = r_n^2$, we then use Lemma 2.5.6 to conclude that, as $r_n^2/\sqrt{n} = o(1)$ by assumption, for any $\epsilon > 0$,

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma^1,\dots,\sigma^T\in\mathfrak{S}_Q} \|\hat{\pi}^{1:T}_{\sigma^{1:T}} - \pi^{*1:T}\|_\infty > \epsilon r_n/n^{1/4}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

In particular, for every $t \in [\![1,T]\!]$, $\hat{\pi}^t$ converges in $\mathbb{P}_{\theta^*}$-probability to π^{*t} up to label switching. Then, let us prove that on the event $\{\min_{\sigma^1,\dots,\sigma^T\in\mathfrak{S}_Q} \|\hat{\pi}^{1:T} - \pi^{*1:T}_{\sigma^{1:T}}\|_\infty \leq \epsilon r_n n^{-1/4}\}$ (whose probability converges to 1), for n large enough, the permutation σ^t minimizing the distance between π^{*t} and $\hat{\pi}^t_{\sigma^t}$ is the same for every $t \in [\![1,T]\!]$. We consider n large enough such that $\epsilon r_n n^{-1/4} < \min_{1\leq q\neq l\leq Q} |\pi^*_{qq} - \pi^*_{ll}|/4$. Denoting by $\sigma^1_m,\dots,\sigma^T_m$ the permutations (depending on n) minimizing $\|\hat{\pi}^{1:T} - \pi^{*1:T}_{\sigma^{1:T}}\|_\infty$, we have that, for any $1 \leq t \neq t' \leq T$, if some $q,l \in [\![1,Q]\!]$ are such that $\sigma^t_m(q) = \sigma^{t'}_m(l)$, then

$$\hat{\pi}^t_{\sigma^t_m(q)\sigma^t_m(q)} = \hat{\pi}^t_{\sigma^{t'}_m(l)\sigma^{t'}_m(l)} = \hat{\pi}^{t'}_{\sigma^{t'}_m(l)\sigma^{t'}_m(l)}$$

and on the event we consider

$$\begin{aligned}
|\pi^{*t}_{qq} - \pi^{*t}_{ll}| = |\pi^{*t}_{qq} - \pi^{*t'}_{ll}| &= |\pi^{*t}_{qq} - \hat{\pi}^t_{\sigma^t_m(q)\sigma^t_m(q)} + \hat{\pi}^{t'}_{\sigma^{t'}_m(l)\sigma^{t'}_m(l)} - \pi^{*t'}_{ll}| \\
&\leq |\pi^{*t}_{qq} - \hat{\pi}^t_{\sigma^t_m(q)\sigma^t_m(q)}| + |\hat{\pi}^{t'}_{\sigma^{t'}_m(l)\sigma^{t'}_m(l)} - \pi^{*t'}_{ll}| \\
&\leq 2\epsilon r_n n^{-1/4} < \min_{1\leq q\neq l\leq Q} |\pi^*_{qq} - \pi^*_{ll}|/2,
\end{aligned}$$

implying that $q = l$. This means that on this event, the permutation σ^t_m minimizing the distance between π^{*t} and $\hat{\pi}^t_{\sigma^t}$ is the same for every $t \in [\![1,T]\!]$. We can conclude that

$$\begin{aligned}
\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q} \|\hat{\pi}^{1:T}_\sigma - \pi^{*1:T}\|_\infty > \epsilon r_n/n^{1/4}\right) &= 1 - \mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q} \|\hat{\pi}^{1:T}_\sigma - \pi^{*1:T}\|_\infty \leq \epsilon r_n/n^{1/4}\right) \\
&\underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

□

A.1.2 Proof of Theorem 2.3.4

First, let us introduce some notations, as in the proof of Theorem 2.3.2. For any fixed configuration $z^{*1:T} \in \Omega_\eta$, we define for any configuration $z^{1:T}$ and any parameter θ

$$D_{n,T}(z^{1:T}, \pi^{1:T}) := \left\{(i,j,t) \in I_{n,T}; \pi^t_{z_i^t z_j^t} \neq \pi^t_{z_i^{*t} z_j^{*t}}\right\}$$

and for any $1 \leq t \leq T$

$$D^t_{n,T}(z^t, \pi^t) := \left\{(i,j) \in [\![1,n]\!]^2; i < j \text{ and } \pi^t_{z_i^t z_j^t} \neq \pi^t_{z_i^{*t} z_j^{*t}}\right\},$$

and as before, we abbreviate to D^* (resp. $\breve{D}$), the set $D_{n,T}(z^{1:T}, \pi^{*1:T})$ (resp. the set $D_{n,T}(z^{1:T}, \breve{\pi}^{1:T})$). We also introduce for any $q, l, q', l' \in [\![1, Q]\!]$ the quantities $F_{qlq'l'}$, F_{ql}, $G_{qlq'l'}$ and G_{ql} as before, accordingly to this definition of $D_{n,T}(z^{1:T}, \pi^{1:T})$. Finally, we introduce for any $t \in [\![1,T]\!]$ and $q, l, q', l' \in [\![1, Q]\!]$ the quantities

$$\begin{aligned}
F^t_{qlq'l'} &= F^t_{qlq'l'}(z^t, z^{*t}) := \{(i,j) \in [\![1,n]\!]^2; i < j \text{ and } z_i^t = q, z_j^t = l, z_i^{*t} = q', z_j^{*t} = l'\} \\
F^t_{ql} &= F^t_{ql}(z^t) := \cup_{1 \leq q', l' \leq Q} F^t_{qlq'l'} = \{(i,j) \in [\![1,n]\!]^2; i < j \text{ and } z_i^t = q, z_j^t = l\} \\
G^t_{qlq'l'} &= G^t_{qlq'l'}(z^t, z^{*t}, \pi^{*t}, \breve{\pi}^t) := (D^{*t} \cup \breve{D}^t) \cap F^t_{qlq'l'} \\
&= \{(i,j) \in [\![1,n]\!]^2; i < j \text{ and } z_i^t = q, z_j^t = l, z_i^{*t} = q', z_j^{*t} = l' \\
&\qquad \text{and } (\pi^{*t}_{z_i^t z_j^t} \neq \pi^{*t}_{z_i^{*t} z_j^{*t}} \text{ or } \breve{\pi}^t_{z_i^t z_j^t} \neq \breve{\pi}^t_{z_i^{*t} z_j^{*t}})\} \\
G^t_{ql} &= G^t_{ql}(z^t, z^{*t}, \pi^{*t}, \breve{\pi}^t) := (D^{*t} \cup \breve{D}^t) \cap F^t_{ql} \\
&= \{(i,j) \in [\![1,n]\!]^2; i < j \text{ and } z_i^t = q, z_j^t = l \text{ and } (\pi^{*t}_{z_i^t z_j^t} \neq \pi^{*t}_{z_i^{*t} z_j^{*t}} \text{ or } \breve{\pi}^t_{z_i^t z_j^t} \neq \breve{\pi}^t_{z_i^{*t} z_j^{*t}})\}.
\end{aligned}$$

Note that we can get an equivalent of Lemma 2.5.8 with a similar proof that gives that for any configuration $z^{*1:T}$ in Ω_η, for any configuration $z^{1:T}$ and any $\theta \in \Theta^T$,

$$\left|D_{n,T}(z^{1:T}, \pi^{1:T})\right| \geq \frac{\gamma^2}{4} nr.$$

In the same way, we have an equivalent of Lemma 2.5.9 (with a similar proof) that gives that for any z^t and z^{*t} two configurations at time t such that $\|z^t - z^{*t}\|_0 = r(t)$ and any parameter $\pi^t = (\pi^t_{ql})_{1 \leq q,l \leq Q}$, we have

$$\begin{aligned}
&D^t_{n,T}(z^t, \pi^t) \subset D^t_{n,T}(z^t) := \left\{(i,j) \in [\![1,n]\!]^2 \times [\![1,T]\!]; (z_i^t, z_j^t) \neq (z_i^{*t}, z_j^{*t})\right\} \\
\text{and} \quad &\left|D^t_{n,T}(z^t)\right| \leq 2nr(t). \qquad \text{(A.1.2)}
\end{aligned}$$

Going back to the proof of Theorem 2.3.4, we follow the line of that of Theorem 2.3.2, with a few changes. We get the same decomposition as in equation (2.5.20), replacing π by $\pi^1, \dots, \pi^T$ in the definitions of U_1, U_2 and U_3, and replacing the event $\Omega_{n,T}$ by $\Omega_n = \{\|\hat{\pi}^{1:T} - \pi^{*1:T}\|_\infty \leq v_n\}$. For U_1, the proof does not change. For U_2, we write (instead of (2.5.23))

$$
\begin{aligned}
|U_2| &\leq \left| \sum_{(i,j,t)\in D^*\cup \breve{D}} \sum_{1\leq q,l\leq Q} \frac{\breve{\pi}^t_{ql} - \pi^{*t}_{ql}}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} (X^t_{ij} - \pi^{*t}_{ql}) \mathbb{1}_{z^t_i = q, z^t_j = l} \right| \\
&\leq \sum_{t=1}^T \sum_{1\leq q,l\leq Q} \left| \frac{\breve{\pi}^t_{ql} - \pi^{*t}_{ql}}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} \sum_{(i,j)\in G^t_{ql}} (X^t_{ij} - \pi^{*t}_{ql}) \right| \\
&\leq \sum_{t=1}^T \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}^t_{ql} - \pi^{*t}_{ql}\right|}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} \left| \sum_{q',l'} \sum_{(i,j)\in G^t_{qlq'l'}} (X^t_{ij} - \pi^{*t}_{ql}) \right| \\
&\leq \sum_{t=1}^T \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}^t_{ql} - \pi^{*t}_{ql}\right|}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} \left| \sum_{q',l'} \sum_{(i,j)\in G^t_{qlq'l'}} (X^t_{ij} - \pi^{*t}_{q'l'}) \right| \\
&\quad + \sum_{t=1}^T \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}^t_{ql} - \pi^{*t}_{ql}\right|}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} \left| \sum_{q',l'} (\pi^{*t}_{q'l'} - \pi^{*t}_{ql}) |G^t_{qlq'l'}| \right| .
\end{aligned}
$$

For every $u > 0$, we thus have

$$
\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\{|U_2| > u\} \cap \Omega_n\right) \\
&\leq \sum_{t=1}^T \mathbb{P}^*_{\theta^*}\left(\left\{ \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}^t_{ql} - \pi^{*t}_{ql}\right|}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} \left| \sum_{1\leq q',l'\leq Q} \sum_{(i,j)\in G^t_{qlq'l'}} (X^t_{ij} - \pi^{*t}_{q'l'}) \right| > \frac{u}{2T} \right\} \cap \Omega_n \right) \\
&\quad + \sum_{t=1}^T \mathbb{P}^*_{\theta^*}\left(\left\{ \sum_{1\leq q,l\leq Q} \frac{\left|\breve{\pi}^t_{ql} - \pi^{*t}_{ql}\right|}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})} \left| \sum_{1\leq q',l'\leq Q} (\pi^{*t}_{q'l'} - \pi^{*t}_{ql}) |G^t_{qlq'l'}| \right| > \frac{u}{2T} \right\} \cap \Omega_n \right) . \quad \text{(A.1.3)}
\end{aligned}
$$

We start by dealing with the first term of (A.1.3). Notice that on the event Ω_n, we have $\left|\breve{\pi}^t_{ql} - \pi^{*t}_{ql}\right| / (\pi^{*t}_{ql}(1-\pi^{*t}_{ql})) \leq v_n/\zeta^2$ for every $q, l \in [\![1, Q]\!]$. As the set G^t_{ql} is random

(because $\breve{D}^t$ is random), we write for every $t \in [\![1,T]\!]$, using (A.1.2),

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\left\{\sum_{1\le q,l\le Q}\frac{\left|\breve{\pi}^t_{ql}-\pi^{*t}_{ql}\right|}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})}\left|\sum_{1\le q',l'\le Q}\sum_{(i,j)\in G^t_{ql}}(X^t_{ij}-\pi^{*t}_{q'l'})\right|>\frac{u}{2T}\right\}\cap\Omega_n\right)\\
\le&\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{1\le q',l'\le Q}\sum_{(i,j)\in G^t_{ql}}(X^t_{ij}-\pi^{*t}_{q'l'})\right|>\frac{u\zeta^2}{2Tv_n}\right)\\
\le&\sum_{D\subset D^t_{n,T}(z^t)}\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{1\le q',l'\le Q}\sum_{(i,j)\in F^t_{ql}\cap D}(X^t_{ij}-\pi^{*t}_{q'l'})\right|>\frac{u\zeta^2}{2Tv_n}\right)
\end{aligned}$$

where now D is a deterministic set. By a union bound and Hoeffding's inequality, we have for any $D\subset D^t_{n,T}(z^t)$

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\sum_{1\le q,l\le Q}\left|\sum_{1\le q',l'\le Q}\sum_{(i,j)\in F^t_{ql}\cap D}(X^t_{ij}-\pi^{*t}_{q'l'})\right|>\frac{u\zeta^2}{2Tv_n}\right)\\
\le&Q^2\max_{1\le q,l\le Q}\mathbb{P}^*_{\theta^*}\left(\left|\sum_{1\le q',l'\le Q}\sum_{(i,j)\in F^t_{ql}\cap D}(X^t_{ij}-\pi^{*t}_{q'l'})\right|>\frac{u\zeta^2}{2Tv_nQ^2}\right)\\
\le&2Q^2\exp\left(-\frac{2u^2\zeta^4}{4T^2v_n^2Q^4}\frac{1}{|D|}\right).
\end{aligned}$$

This leads to, for the first term of (A.1.3),

$$\begin{aligned}
&\sum_{t=1}^T\mathbb{P}^*_{\theta^*}\left(\left\{\sum_{1\le q,l\le Q}\left|\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})}\right|\left|\sum_{1\le q',l'\le Q}\sum_{(i,j)\in G^t_{ql}}(X^t_{ij}-\pi^{*t}_{q'l'})\right|>\frac{u}{2T}\right\}\cap\Omega_n\right)\\
\le&\sum_{t=1}^T\sum_{D\subset D^t_{n,T}(z^t)}2Q^2\exp\left(-\frac{2u^2\zeta^4}{4T^2v_n^2Q^4}\frac{1}{|D|}\right)\\
\le&\sum_{t=1}^T\sum_{k=1}^{2nr(t)}\sum_{D\subset D^t_{n,T}(z^t);|D|=k}2Q^2\exp\left(-\frac{2u^2\zeta^4}{4T^2v_n^2Q^4}\frac{1}{k}\right)\\
\le&2Q^2\sum_{t=1}^T\exp\left(-\frac{u^2\zeta^4}{4T^2v_n^2Q^4nr(t)}\right)(2nr(t))^{2nr(t)+1}\\
\le&2Q^2T\exp\left(-\frac{u^2\zeta^4}{4T^2v_n^2Q^4nr}\right)(2nr)^{2nr+1}.
\end{aligned}$$

For the second term of (A.1.3), we get from a union bound and from (A.1.2) that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{P}^*_{\theta^*}\left(\left\{\sum_{1\le q,l\le Q}\left|\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{\pi^{*t}_{ql}(1-\pi^{*t}_{ql})}\right|\left|\sum_{1\le q',l'\le Q}(\pi^{*t}_{q'l'}-\pi^{*t}_{ql})|G^t_{qlq'l'}|\right|>\frac{u}{2T}\right\}\cap\Omega_n\right)\\
\le& Q^2\sum_{t=1}^{T}\max_{1\le q,l\le Q}\mathbb{P}^*_{\theta^*}\left(\left|\sum_{1\le q',l'\le Q}(\pi^{*t}_{q'l'}-\pi^{*t}_{ql})|G^t_{qlq'l'}|\right|>\frac{u\zeta^2}{2Tv_nQ^2}\right)\\
\le& Q^2T\mathbb{P}^*_{\theta^*}\left(2nr>\frac{u\zeta^2}{2v_nTQ^2}\right).
\end{aligned}$$

Finally, we have the following upper bound for U_2

$$\begin{aligned}
\mathbb{P}^*_{\theta^*}\left(\Omega_n\cap\{|U_2|>r\log(nT)\}\right)\le& 2Q^2T\exp\left(-\frac{r\zeta^4(\log(nT))^2}{4Q^4T^2v_n^2n}\right)(2nr)^{2nr+1}\\
&+Q^2T\mathbb{P}^*_{\theta^*}\left(v_n>\frac{\zeta^2\log(nT)}{4Q^2Tn}\right).
\end{aligned}$$

For the third term U_3, denoting $G^{*t}_{ql}=\cup_{1\le q',l'\le Q}G^t_{ql}=\{(i,j)\in D^{*t}\cup\breve{D}^t;z^{*t}_i=q,z^{*t}_j=l\}$, we have

$$\begin{aligned}
U_3=&\sum_{1\le q,l\le Q}\sum_{(i,j,t)\in D^*\cup\breve{D}}\left((\pi^{*t}_{ql}-X^t_{ij})\log\left[1-\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{(1-\pi^{*t}_{ql})}\right]\right.\\
&\left.+(X^t_{ij}-\pi^{*t}_{ql})\log\left[1+\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{\pi^{*t}_{ql}}\right]\right)\mathbb{1}_{z^{*t}_i=q,z^{*t}_j=l}\\
&+\sum_{1\le q,l\le Q}\sum_{(i,j,t)\in D^*\cup\breve{D}}\left((1-\pi^*_{ql})\log\left[1-\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{(1-\pi^{*t}_{ql})}\right]\right.\\
&\left.+\pi^{*t}_{ql}\log\left[1+\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{\pi^{*t}_{ql}}\right]\right)\mathbb{1}_{z^{*t}_i=q,z^{*t}_j=l}\\
=&\sum_{t=1}^{T}\sum_{1\le q,l\le Q}\left(\log\left[1+\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{\pi^{*t}_{ql}}\right]-\log\left[1-\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{(1-\pi^{*t}_{ql})}\right]\right)\sum_{(i,j)\in G^{*t}_{ql}}(X^t_{ij}-\pi^{*t}_{ql})\\
&+\sum_{t=1}^{T}\sum_{1\le q,l\le Q}|G^{*t}_{ql}|\left((1-\pi^{*t}_{ql})\log\left[1+\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{\pi^{*t}_{ql}}\right]+\pi^{*t}_{ql}\log\left[1-\frac{(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})}{(1-\pi^{*t}_{ql})}\right]\right).
\end{aligned}$$

Then, we have on the event Ω_n and for n large enough such that $|(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})/\pi^{*t}_{ql}|\le 1/2$ and $|(\breve{\pi}^t_{ql}-\pi^{*t}_{ql})/(1-\pi^{*t}_{ql})|\le 1/2$ for every q and l, using the fact that $|\log(1+x)|\le 2|x|$

for $x \in [-1/2, 1/2]$,

$$|U_3| \leq \sum_{t=1}^{T} 4\frac{v_n}{\zeta} \sum_{1\leq q,l\leq Q} \left| \sum_{(i,j)\in G_{ql}^{*t}} (X_{ij}^t - \pi_{ql}^{*t}) \right| + \sum_{t=1}^{T} 4\frac{v_n}{\zeta} \sum_{1\leq q,l\leq Q} |G_{ql}^{*t}|.$$

Then, for every $u > 0$,

$$\begin{aligned}\mathbb{P}_{\theta^*}^* \left(\Omega_n \cap \{|U_3| > u\}\right) \leq & \sum_{t=1}^{T} \mathbb{P}_{\theta^*}^* \left(\sum_{1\leq q,l\leq Q} \left| \sum_{(i,j)\in Q_{ql}^{*t}} (X_{ij}^t - \pi_{ql}^{*t}) \right| > \frac{u\zeta}{8v_n T} \right) \\ & + \sum_{t=1}^{T} \mathbb{P}_{\theta^*}^* \left(v_n \sum_{1\leq q,l\leq Q} |G_{ql}^{*t}| > \frac{u\zeta}{8T} \right). \end{aligned} \tag{A.1.4}$$

For the first term of (A.1.4), using Hoeffding's inequality as before,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{P}_{\theta^*}^* \left(\sum_{q,l} \left| \sum_{(i,j)\in G_{ql}^{*t}} (X_{ij}^t - \pi_{ql}^{*t}) \right| > u\zeta/(8v_n T) \right) \\ \leq & \sum_{t=1}^{T} \sum_{k=1}^{2nr(t)} \sum_{D\subset D_{n,T}^t(z^t);|D|=k} \mathbb{P}_{\theta^*}^* \left(\sum_{q,l} \left| \sum_{(i,j)\in F_{ql}^{*t}\cap D} (X_{ij}^t - \pi_{ql}^{*t}) \right| > u\zeta/(8v_n T) \right) \\ \leq & 2Q^2 T \exp\left(-\frac{u^2\zeta^2}{8^2 T^2 Q^4 v_n^2 nr} \right) (2nr)^{2nr+1}, \end{aligned}$$

and for the second term of (A.1.4),

$$\sum_{t=1}^{T} \mathbb{P}_{\theta^*}^* \left(v_n \sum_{q,l} |G_{ql}^{*t}| > \frac{u\zeta}{8T} \right) \leq T\mathbb{P}_{\theta^*}^* \left(v_n > \frac{u\zeta}{16Tnr} \right).$$

Finally, we have the following upper bound for U_3

$$\begin{aligned}\mathbb{P}_{\theta^*}^* \left(\Omega_n \cap \{|U_3| > r\log(nT)\}\right) \leq & 2Q^2 T \exp\left(-\frac{r\zeta^2 (\log(nT))^2}{8^2 T^2 Q^4 v_n^2 n} \right) (2nr)^{2nr+1} \\ & + T\mathbb{P}_{\theta^*}^* \left(v_n > \frac{\zeta \log(nT)}{16Tn} \right). \end{aligned}$$

Now we choose the sequence v_n such that $v_n = o(\sqrt{\log n}/n)$ which is sufficient to imply that the quantities $\mathbb{P}_{\theta^*}^* \left(v_n > \zeta^2 \log(nT)/(4Q^2 Tn)\right)$ and $\mathbb{P}_{\theta^*}^* \left(v_n > \zeta \log(nT)/(16Tn)\right)$ vanish as n increases and we gather the three upper bounds. For large enough values of n and with C_1, C_2, C_3, C_4 and κ positive constants only depending on Q, ζ, K^* and T,

we then have

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\{U_1+U_2-U_3 > -\log(1/(\epsilon y_n)) - 3r\log(nT)\} \cap \Omega_n\right) \\
&\leq \exp\left[(\log(1/(\epsilon y_n)) + 5r\log(nT))\frac{2K^*}{C_\zeta}\right] \exp\left[-nr\frac{(\delta-\eta)^2 K^{*2}}{4C_\zeta}\right] \\
&\quad + 2Q^2T\exp\left(-\frac{r\zeta^4(\log(nT))^2}{4Q^4T^2v_n^2 n}\right)(2nr)^{2nr+1} + 2Q^2T\exp\left(-\frac{r\zeta^2(\log(nT))^2}{8^2T^2Q^4v_n^2 n}\right)(2nr)^{2nr+1} \\
&\leq \exp\left[-(\delta-\eta)^2 C_1 nr + C_2\log(nT)r + C_4\log(1/(\epsilon y_n))\right] \\
&\quad + \kappa\exp\left[5nr\log(nT) - C_3\frac{(\log(nT))^2 r}{nv_n^2}\right].
\end{aligned}$$

Then, introducing

$$\begin{aligned}
u_{nT} &= \exp\left[-(\delta-\eta)^2 C_1 n + C_2\log(nT) + C_4\log(1/(\epsilon y_n))\right] \\
w_{nT} &= \exp\left[-C_3\frac{(\log(nT))^2}{nv_n^2} + 5n\log(nT)\right],
\end{aligned}$$

we conclude as in the proof of Theorem 2.3.2, noticing that nTu_{nT} (resp. nTw_{nT}) converges to 0 as n increases as long as $\log(1/y_n) = o(n)$ (resp. as long as $v_n = o(\sqrt{\log(n)}/n)$). □

A.1.3 Proof of Corollary 2.4.2

As in the proof of Theorem 2.4.1, using the convergence in Equation (A.1.1) and Lemma 2.5.11, we obtain for any $\epsilon > 0$

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\Theta}\left|\frac{2}{n(n-1)T}\mathcal{J}(\hat{\chi}(\theta),\theta) - \mathbb{M}^T(\pi^{1:T})\right| > \frac{\epsilon r_n^2}{\sqrt{n}}\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

We then conclude by using Lemma 2.5.6 applied with $F_{n,T} = \frac{2}{n(n-1)T}\mathcal{J}(\hat{\chi}(\cdot),\cdot)$. □

A.2 Proofs of technical lemmas

A.2.1 Proof of Lemma 2.3.1

As in the proof of Lemma E.2 from Celisse et al. (2012), we use the method of Lagrange multipliers to find the fixed-point equation of the critical point. Recall that $\theta = (\Gamma, \pi)$ and let us denote the likelihood $L(\Gamma,\pi) \coloneqq \exp\ell(\theta) = \mathbb{P}_\theta(X^{1:T})$ and the conditional

likelihood $L_c(z^{1:T}, \pi) = \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T} = z^{1:T})$. Recall the definition of $N_{ql}(z^{1:T})$ in (2.2.1) and that

$$\mathbb{P}_\theta(Z^{1:T} = z^{1:T}) = \prod_{1 \le q,l \le Q} \gamma_{ql}^{N_{ql}(z^{1:T})} \prod_{i=1}^{n} \alpha^1_{z^1_i}.$$

We compute the derivative of the Lagrangian with respect to each parameter γ_{ql}.

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_{ql}} \left[L(\Gamma, \pi) + \sum_{m=1}^{Q} \lambda_m \left(\sum_{k=1}^{Q} \gamma_{mk} - 1 \right) \right] \\
=&\frac{\partial}{\partial \gamma_{ql}} \left(\sum_{z^{1:T}} L_c(z^{1:T}, \pi) \mathbb{P}_\theta(Z^{1:T} = z^{1:T}) \right) + \lambda_q \\
=&\sum_{z^{1:T}} L_c(z^{1:T}, \pi) \frac{N_{ql}(z^{1:T})}{\gamma_{ql}} \mathbb{P}_\theta(Z^{1:T} = z^{1:T}) + \lambda_q \\
=&\frac{1}{\gamma_{ql}} \left(\sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{z^{1:T}} L_c(z^{1:T}, \pi) \mathbb{P}_\theta(Z^{1:T} = z^{1:T}) \mathbb{1}_{z_i^t = q, z_i^{t+1} = l} + \lambda_q \gamma_{ql} \right) \\
=&\frac{1}{\gamma_{ql}} \left(\sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_\theta(X^{1:T}, Z_i^t = q, Z_i^{t+1} = l) + \lambda_q \gamma_{ql} \right).
\end{aligned}$$

At the critical point $\breve{\theta} = (\breve{\gamma}, \breve{\pi})$, we obtain that for each $(q, l) \in [\![1, Q]\!]^2$ we have

$$\breve{\gamma}_{ql} \propto \sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}(X^{1:T}, Z_i^t = q, Z_i^{t+1} = l)$$

where $\propto$ means 'proportional to'. The constraint $\sum_l \gamma_{ql} = 1$ gives the normalizing term and we obtain

$$\breve{\gamma}_{ql} = \frac{\sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}(X^{1:T}, Z_i^t = q, Z_i^{t+1} = l)}{\sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}(X^{1:T}, Z_i^t = q)} = \frac{\sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}(Z_i^t = q, Z_i^{t+1} = l \mid X^{1:T})}{\sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_{\breve{\theta}}(Z_i^t = q \mid X^{1:T})}.$$

□

A.2.2 Proof of Lemma 2.4.1

We can write the quantity to optimize

$$\begin{aligned}
\mathcal{J}(\chi,\theta) =& \mathbb{E}_{\mathbb{Q}_\chi}\left[\log \mathbb{P}_\theta(X^{1:T}, Z^{1:T})\right] + \mathcal{H}(\mathbb{Q}_\chi) \\
=& \mathbb{E}_{\mathbb{Q}_\chi}\left[\log \mathbb{P}_\theta(X^{1:T} \mid Z^{1:T})\right] + \mathbb{E}_{\mathbb{Q}_\chi}\left[\log \mathbb{P}_\theta(Z^{1:T})\right] - \mathbb{E}_{\mathbb{Q}_\chi}\left[\log \mathbb{Q}_\chi(Z^{1:T})\right] \\
=& \mathbb{E}_{\mathbb{Q}_\chi}\left[\sum_{t=1}^{T}\sum_{i<j} X_{ij}^t \log \pi_{Z_i^t Z_j^t} + (1 - X_{ij}^t)\log(1 - \pi_{Z_i^t Z_j^t})\right] \\
&+ \mathbb{E}_{\mathbb{Q}_\chi}\left[\sum_{i=1}^{n} \log \alpha_{Z_i^1} + \sum_{i=1}^{n}\sum_{t=1}^{T-1} \log \gamma_{Z_i^t Z_i^{t+1}}\right] \\
&- \mathbb{E}_{\mathbb{Q}_\chi}\left[\sum_{i=1}^{n} \log \mathbb{Q}_\chi(Z_i^1) + \sum_{i=1}^{n}\sum_{t=1}^{T-1} \log \mathbb{Q}_\chi(Z_i^{t+1} \mid Z_i^t)\right] \\
=& \sum_{t=1}^{T}\sum_{i<j}\sum_{q,l} \tau_{iq}^t \tau_{jl}^t \left[X_{ij}^t \log \pi_{ql} + (1 - X_{ij}^t)\log(1 - \pi_{ql})\right] \\
&+ \sum_{i=1}^{n}\sum_{q=1}^{Q} \tau_{iq}^1 \log \alpha_q + \sum_{i=1}^{n}\sum_{q,l}\sum_{t=1}^{T-1} \eta_{iql}^t \log \gamma_{ql} \\
&- \sum_{i=1}^{n}\sum_{q=1}^{Q} \tau_{iq}^1 \log \tau_{iq}^1 - \sum_{i=1}^{n}\sum_{t=1}^{T-1}\sum_{q,l} \eta_{iql}^t \log \frac{\eta_{iql}^t}{\tau_{iq}^t}.
\end{aligned} \tag{A.2.1}$$

Using this expression, we can obtain directly the expected fixed-point equation for the variational estimator of the transition probability from q to l. □

A.2.3 Proof of Lemma 2.5.1

We rely on the notation introduced in the proof of Theorem 2.3.1. For any $t \in [\![1, T]\!]$, using classical dependency rules in directed acyclic graphs and the expression (2.5.3) of $\hat{z}^t$, we write

$$\begin{aligned}
\log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) &= \log \sum_{z^t} \mathbb{P}_\theta(X^t \mid Z^t = z^t)\mathbb{P}_\theta(Z^t = z^t \mid X^{1:t-1}) \\
&\leq \log\left[\mathbb{P}_\theta(X^t \mid Z^t = \hat{z}^t)\sum_{z^t} \mathbb{P}_\theta(Z^t = z^t \mid X^{1:t-1})\right] \\
&\leq \log \mathbb{P}_\theta(X^t \mid Z^t = \hat{z}^t)
\end{aligned}$$

and thus

$$\log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) - \log \mathbb{P}_\theta(X^t \mid Z^t = \hat{z}^t) \leq 0. \tag{A.2.2}$$

Using Bayes' rule, we have

$$\log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) = \log \mathbb{P}_\theta(X^t, Z^t \mid X^{1:t-1}) - \log \mathbb{P}_\theta(Z^t \mid X^{1:t}).$$

Taking the expectation of this quantity with respect to any distribution $\mathbb{Q}$ on Z^t, we obtain

$$\begin{aligned}\log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) &= \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X^t, Z^t \mid X^{1:t-1})\right] + \mathrm{KL}\left(\mathbb{Q}; \mathbb{P}_\theta(Z^t \mid X^{1:t})\right) + \mathcal{H}(\mathbb{Q}) \\ &\geq \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X^t, Z^t \mid X^{1:t-1})\right] + \mathcal{H}(\mathbb{Q}) \\ &\geq \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(X^t \mid Z^t)\right] + \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{P}_\theta(Z^t \mid X^{1:t-1})\right] + \mathcal{H}(\mathbb{Q}),\end{aligned}$$

where $\mathrm{KL}\left(\mathbb{Q}; \mathbb{P}_\theta(Z^t \mid X^{1:t})\right) = \mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{Q}(Z^t) - \log \mathbb{P}_\theta(Z^t \mid X^{1:t})\right]$ is a Kullback-Leibler divergence (thus non negative) and $\mathcal{H}(\mathbb{Q}) = -\mathbb{E}_{\mathbb{Q}}\left[\log \mathbb{Q}(Z^t)\right]$ is the entropy of $\mathbb{Q}$.

Taking now $\mathbb{Q}$ as the Dirac distribution located on $\hat{z}^t$, we have $\mathcal{H}(\mathbb{Q}) = 0$ and

$$\log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) \geq \log \mathbb{P}_\theta(X^t \mid Z^t = \hat{z}^t) + \log \mathbb{P}_\theta(Z^t = \hat{z}^t \mid X^{1:t-1}). \tag{A.2.3}$$

Now, combining Inequalities (A.2.2) and (A.2.3), we obtain

$$\log \mathbb{P}_\theta(Z^t = \hat{z}^t \mid X^{1:t-1}) \leq \log \mathbb{P}_\theta(X^t \mid X^{1:t-1}) - \log \mathbb{P}_\theta(X^t \mid Z^t = \hat{z}^t) \leq 0,$$

giving the expected result. □

A.2.4 Proof of Lemma 2.5.2

To prove this lemma, we first establish a control of the expectation of the random variable appearing in the statement.

Lemma A.2.1. *We have the following inequality for $z^{*1:T}$ and $z^{1:T}$ any configurations and any $\theta \in \Theta$*

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[\sup_{\substack{(z^{1:T},\pi)\in \\ [\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}} \left|\frac{2}{n(n-1)T}\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^t z_j^t}}{1-\pi_{z_i^t z_j^t}}\right)\right| \,\middle|\, Z^{1:T} = z^{*1:T}\right] \\ \leq &\sqrt{\frac{2}{n(n-1)T}}\Lambda\end{aligned}$$

with $\Lambda = 2\log[(1-\zeta)/\zeta]$.

We now turn to the proof of Lemma 2.5.2. Let us first recall Talagrand's inequality (see for e.g. Massart, 2007, page 170, Equation (5.50)).

Theorem (Talagrand's inequality). Let $\{Y_{ij}^t\}_{1\leq i<j\leq n, 1\leq t\leq T}$ denote independent and centered random variables. Define

$$\forall g \coloneqq \{g_{ij}^t\}_{1\leq i<j\leq n,1\leq t\leq T} \in \mathcal{G}, \quad S_{n,T}(g) = \sum_{1\leq i<j\leq n}\sum_{t=1}^{T} Y_{ij}^t g_{ij}^t,$$

where $\mathcal{G} \subset \mathbb{R}^{n(n-1)T/2}$. Let us further assume that there exist $b > 0$ and $\sigma^2 > 0$ such that $|Y_{ij}^t g_{ij}^t| \leq b$ for every $(i,j,t) \in [\![1,n]\!]^2 \times [\![1,T]\!]$ and any $g \in \mathcal{G}$ and $\sup_{g\in\mathcal{G}} \sum_{i<j}\sum_t \mathrm{Var}(Y_{ij}^t g_{ij}^t) \leq \sigma^2$. Then, for every $\beta > 0$ and $x > 0$, for any finite set $\{g_1, \ldots, g_{2^{n(n-1)T/2}}\}$ of elements of $\mathcal{G}$, we have

$$\mathbb{P}\Bigg(\max_{g\in\{g_1,\ldots,g_{2^{n(n-1)T/2}}\}} S_{n,T}(g) \geq \mathbb{E}\left[\max_{g\in\{g_1,\ldots,g_{2^{n(n-1)T/2}}\}} S_{n,T}(g)\right](1+\beta) + \sqrt{2\sigma^2 x} + b(\beta^{-1}+3^{-1})x\Bigg) \leq e^{-x}.$$

First, notice that $\operatorname{argmin}_{\varpi\in[\zeta,1-\zeta]} \log(\varpi/(1-\varpi)) = \zeta$ and $\arg\max_{\varpi\in[\zeta,1-\zeta]} \log(\varpi/(1-\varpi)) = 1-\zeta$ so that we have

$$\begin{aligned}\mathbb{P}^*_{\theta^*}\left(\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^t z_j^t}}{1-\pi_{z_i^t z_j^t}}\right)\right| > \epsilon\right) \\ \leq \mathbb{P}^*_{\theta^*}\left(\max_{\varpi\in\{\zeta,1-\zeta\}^{n(n-1)T/2}} \frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\varpi_{i,j}^t}{1-\varpi_{i,j}^t}\right)\right| > \epsilon\right)\end{aligned}$$

with $\varpi \coloneqq \{\varpi_{i,j}^t\}_{1\leq i<j\leq n,1\leq t\leq T}$. The set $\{\zeta, 1-\zeta\}^{n(n-1)T/2}$ is finite, of size $2^{n(n-1)T/2}$. Let us now apply Talagrand's inequality to our setup. Note that for every $(i,j,t) \in [\![1,n]\!]^2 \times [\![1,T]\!]$, for any $\pi \in [\zeta, 1-\zeta]^{Q^2}$, we have

$$\left|(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^t z_j^t}}{1-\pi_{z_i^t z_j^t}}\right)\right| \leq \log[(1-\zeta)/\zeta] = \frac{\Lambda}{2}$$

almost surely thanks to Assumption 3, and with Λ as defined in Lemma A.2.1. Combining this result with Lemma A.2.1 and writing

$$\Omega = (1+\beta)\Lambda\sqrt{n(n-1)T/2} + \sqrt{n(n-1)T(\Lambda/2)^2 x_{n,T}} + (1/\beta + 1/3)(\Lambda/2)x_{n,T},$$

we have for any $\epsilon > 0$, for any $\beta > 0$, applying Talagrand's inequality with $b = \Lambda/2$ and $\sigma^2 = n(n-1)T/2(\Lambda/2)^2$,

$$\begin{aligned}
&\mathbb{P}^*_{\theta^*}\left(\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^t z_j^t}}{1-\pi_{z_i^t z_j^t}}\right)\right| > \epsilon\right)\\
\leq&\mathbb{P}^*_{\theta^*}\left(\max_{\varpi\in\{\zeta,1-\zeta\}^{n(n-1)T/2}} \frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\varpi_{i,j}^t}{1-\varpi_{i,j}^t}\right)\right| > \epsilon\right)\\
\leq&\mathbb{P}^*_{\theta^*}\left(\epsilon < \max_{\substack{\varpi\in\\ \{\zeta,1-\zeta\}^{n(n-1)T/2}}} \frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\varpi_{i,j}^t}{1-\varpi_{i,j}^t}\right)\right| \leq \frac{2}{n(n-1)T}\Omega\right)\\
&+\mathbb{P}^*_{\theta^*}\left(\max_{\substack{\varpi\in\\ \{\zeta,1-\zeta\}^{n(n-1)T/2}}} \frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t - \pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\varpi_{i,j}^t}{1-\varpi_{i,j}^t}\right)\right| > \frac{2}{n(n-1)T}\Omega\right)\\
\leq&\mathbb{P}^*_{\theta^*}\left(\frac{2}{n(n-1)T}\Omega > \epsilon\right) + 2e^{-x_{n,T}} \leq \mathbb{1}_{\epsilon<2\Omega/(n(n-1)T)} + 2e^{-x_{n,T}}.
\end{aligned}$$

□

A.2.5 Proof of Lemma 2.5.3

For any $\eta \in (0, \delta)$, Hoeffding's inequality (see for example Theorem 2.8 from Boucheron et al., 2013) gives that

$$\begin{aligned}
&\mathbb{P}_\theta\left(\forall t \in [\![1,T]\!], \forall q \in [\![1,Q]\!], \frac{N_q(Z^t)}{n} \geq \alpha_q - \eta\right)\\
=&1 - \mathbb{P}_\theta\left(\exists t \in [\![1,T]\!], \exists q \in [\![1,Q]\!]; \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Z_i^t=q} < \alpha_q - \eta\right)\\
\geq&1 - \sum_{q=1}^{Q}\sum_{t=1}^{T}\exp\left(-2\eta^2 n\right) \geq 1 - QT\exp\left(-2\eta^2 n\right),
\end{aligned}$$

which concludes the proof. □

A.2.6 Proof of Lemma 2.5.4

First notice that $\arg\max_{A\in\mathcal{A}}\mathbb{M}(\pi, A)$ may not be unique, it is in fact a closed subset of $\mathcal{A}$. However, we choose a fixed element $\bar{A}_\pi$ in this subset in the following. Letting $\epsilon > 0$

and $\eta \in (0, \delta)$ and using Lemma 2.5.3, we can split the probability as

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|>\frac{\epsilon r_n}{6\sqrt{n}}\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\left\{\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|>\frac{\epsilon r_n}{6\sqrt{n}}\right\}\cap\Omega_\eta(\theta^*)\right)\\
&+QT\exp\left(-2\eta^2 n\right),
\end{aligned}$$

recalling that

$$\Omega_\eta(\theta) := \left\{z^{1:T}\in[\![1,Q]\!]^{nT};\forall t\in[\![1,T]\!],\forall q\in[\![1,Q]\!],\frac{N_q(z^t)}{n}\geq\alpha_q-\eta\right\}.$$

We thus want to bound the quantity

$$\mathbb{P}_{\theta^*}\left(T^{-1}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|>\epsilon r_n/(6\sqrt{n})\right)$$

on the event $\left\{Z^{1:T}\in\Omega_\eta(\theta^*)\right\}$, which means bounding

$$\mathbb{P}_{\theta^*}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|>\frac{\epsilon r_n}{6\sqrt{n}}\,\middle|\,Z^{1:T}\in\Omega_\eta(\theta^*)\right).$$

Let us denote for any matrix P of size $m\times n$ the norm $\|P\|_\infty=\max_{(i,j)\in[\![1,m]\!]\times[\![1,n]\!]}|P_{ij}|$. Then note that, for any matrix $\breve{A}$ with coefficients in $[0,1]$, for any $\pi\in[\zeta,1-\zeta]^{Q^2}$, using Assumption 2 and 3,

$$\begin{aligned}
&\left(\mathbb{M}(\pi,\bar{A}_\pi)-\mathbb{M}(\pi,\breve{A})\right)\\
\leq&\sum_{q,l}\alpha_q^*\alpha_l^*\sum_{q',l'}|\bar{a}_{qq'}\bar{a}_{ll'}-\breve{a}_{qq'}\breve{a}_{ll'}|\sup_{\pi\in[\zeta,1-\zeta]^{Q^2}}|\pi_{ql}^*\log\pi_{q'l'}+(1-\pi_{ql}^*)\log(1-\pi_{q'l'})|\\
\leq&2(1-\delta)^2(1-\zeta)\log(1/\zeta)\sum_{q,l}\sum_{q',l'}|\bar{a}_{qq'}\bar{a}_{ll'}-\breve{a}_{qq'}\breve{a}_{ll'}|\\
\leq&2(1-\delta)^2(1-\zeta)\log(1/\zeta)Q^4 2\|\breve{A}-\bar{A}_\pi\|_\infty := c\|\breve{A}-\bar{A}_\pi\|_\infty
\end{aligned}$$

with $c = 4(1-\delta)^2(1-\zeta)\log(1/\zeta)Q^4$. On the event $\Omega_\eta(\theta^*)$ we then have

$$
\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|>\frac{\epsilon r_n}{6\sqrt{n}}\right)\\
=&1-\mathbb{P}_{\theta^*}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2}}\left|\mathbb{M}(\pi,\bar{A}_\pi^t)-\mathbb{M}(\pi,\bar{A}_\pi)\right|\leq\frac{\epsilon r_n}{6\sqrt{n}}\right)\\
\leq&1-\mathbb{P}_{\theta^*}\left(\forall t\in[\![1,T]\!],\sup_{\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2}}\left(\mathbb{M}(\pi,\bar{A}_\pi)-\mathbb{M}(\pi,\bar{A}_\pi^t)\right)\leq\frac{\epsilon r_n}{6\sqrt{n}}\right)\\
\leq&1-\mathbb{P}_{\theta^*}\left(\forall t\in[\![1,T]\!],\forall\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2},\left(\mathbb{M}(\pi,\bar{A}_\pi)-\mathbb{M}(\pi,\bar{A}_\pi^t)\right)\leq\frac{\epsilon r_n}{6\sqrt{n}}\right)\\
\leq&1-\mathbb{P}_{\theta^*}\left(\forall t\in[\![1,T]\!],\forall\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2},\exists\breve{A}\in\mathcal{A}^t(Z^{1:T});\left(\mathbb{M}(\pi,\bar{A}_\pi)-\mathbb{M}(\pi,\breve{A})\right)\leq\frac{\epsilon r_n}{6\sqrt{n}}\right)\\
\leq&1-\mathbb{P}_{\theta^*}\left(\forall t\in[\![1,T]\!],\forall\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2},\exists\breve{A}\in\mathcal{A}^t(Z^{1:T});\|\breve{A}-\bar{A}_\pi\|_\infty<\frac{\epsilon r_n}{6c\sqrt{n}}\right).
\end{aligned}
$$

We then show that for any $\epsilon>0$, for every $t\in[\![1,T]\!]$ and every $\pi\in[\zeta,1-\zeta]^{\mathcal{Q}^2}$, for any n such that $n>6c\sqrt{n}/[\epsilon r_n(\delta-\eta)]$, there exists some $\breve{A}\in\mathcal{A}^t(Z^{1:T})$ such that $\|\breve{A}-\bar{A}_\pi\|_\infty<\epsilon r_n/(6c\sqrt{n})$, i.e. such that for every q,l, $|\breve{a}_{ql}-\bar{a}_{ql}|<\epsilon r_n/(6c\sqrt{n})$. For every $1\leq q\leq Q$, we can construct $\breve{A}_{q\cdot}=(\breve{a}_{q1},\ldots,\breve{a}_{qQ})$ as follows. On the event $\Omega_\eta(\theta^*)$, for every $q\in[\![1,Q]\!]$, for any n such that $n>6c\sqrt{n}/[\epsilon r_n(\delta-\eta)]$, we have $N_q(Z^t)\epsilon r_n/(6c\sqrt{n})>1$ for every $t\in[\![1,T]\!]$. We then construct $(\breve{n}_{ql})_{1\leq l\leq Q}$ as follows and take $\breve{a}_{ql}=\breve{n}_{ql}/N_q(Z^{1:T})$ for every $l\in[\![1,Q]\!]$.

- for $l=1$ choose $\breve{n}_{q1}$ as the closest integer to $N_q(Z^t)\bar{a}_{q1}$. It is in the interval $(N_q(Z^t)\bar{a}_{q1}-1,N_q(Z^t)\bar{a}_{q1}+1)$ so we have $|\bar{a}_{q1}-\breve{n}_{q1}/N_q(Z^t)|<1/N_q(Z^t)<\epsilon r_n/(6c\sqrt{n})$. Moreover, note that $0\leq\breve{n}_{q1}\leq N_q(Z^t)$ because $0\leq N_q(Z^t)\bar{a}_{q1}\leq N_q(Z^t)$.
- Repeat for $l=2,\ldots,Q$
  - if $\sum_{l'=1}^{l-1}(N_q(Z^t)\bar{a}_{ql'}-\breve{n}_{ql'})\geq 0$ choose $\breve{n}_{ql}$ as the closest bigger (or equal) integer to $N_q(Z^t)\bar{a}_{ql}$.
  - if $\sum_{l'=1}^{l-1}(N_q(Z^t)\bar{a}_{ql'}-\breve{n}_{ql'})<0$ choose $\breve{n}_{ql}$ as the closest smaller (or equal) integer to $N_q(Z^t)\bar{a}_{ql}$.

  As before, $\breve{n}_{ql}$ is in the interval $(N_q(Z^t)\bar{a}_{ql}-1,N_q(Z^t)\bar{a}_{ql}+1)$ so we have $|\bar{a}_{ql}-\breve{n}_{ql}/N_q(Z^t)|<1/N_q(Z^{1:T})<\epsilon r_n/(6c\sqrt{n})$. Moreover $0\leq\breve{n}_{ql}\leq N_q(Z^t)$ because

$0 \leq N_q(Z^t)\bar{a}_{ql} \leq N_q(Z^t)$. We also have (by induction)

$$\left|\sum_{l'=1}^{l}(N_q(Z^t)\bar{a}_{ql'} - \breve{n}_{ql'})\right| = \left|\left(\sum_{l'=1}^{l-1} N_q(Z^t)\bar{a}_{ql'} - \breve{n}_{ql'}\right) + N_q(Z^t)\bar{a}_{ql} - \breve{n}_{ql}\right| < 1.$$

In the end, we have $|\sum_{l=1}^{Q}(N_q(Z^t)\bar{a}_{ql} - \breve{n}_{ql})| < 1$ i.e. $|N_q(Z^t) - \sum_{l=1}^{Q}\breve{n}_{ql}| < 1$, meaning that $\sum_{l=1}^{Q}\breve{n}_{ql} = N_q(Z^t)$, both $N_q(Z^t)$ and $\sum_{l=1}^{Q}\breve{n}_{ql}$ being integers. Then, if $n > 6c\sqrt{n}/[\epsilon r_n(\delta - \eta)]$, there exists $\breve{A} \in \mathcal{A}^t(Z^{1:T})$ such that $\|\breve{A} - \bar{A}_\pi\|_\infty < \epsilon r_n/(6c\sqrt{n})$. This leads to

$$\mathbb{P}_{\theta^*}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\mathbb{M}(\pi, \bar{A}_\pi^t) - \mathbb{M}(\pi, \bar{A}_\pi)\right| > \frac{\epsilon r_n}{6\sqrt{n}}\right) \leq QT\exp(-2\eta^2 n) + 1 - \mathbb{1}_{n>6c\sqrt{n}/[\epsilon r_n(\delta-\eta)]}$$

which concludes the proof. □

A.2.7 Proof of Lemma 2.5.5

We can upper bound the expectation as follows

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left[\left|\frac{N_q(Z^1)N_l(Z^1)}{n(n-1)} - \alpha_q^*\alpha_l^*\right|\right] =& \mathbb{E}_{\theta^*}\left[\left|\left(\frac{N_q(Z^1)}{n} - \alpha_q^*\right)\frac{N_l(Z^1)}{n-1} + \alpha_q^*\left(\frac{N_l(Z^1)}{n-1} - \alpha_l^*\right)\right|\right] \\
\leq& \mathbb{E}_{\theta^*}\left[\left|\frac{N_q(Z^1)}{n} - \alpha_q^*\right|\frac{N_l(Z^1)}{n-1}\right] + \alpha_q^*\mathbb{E}_{\theta^*}\left[\left|\frac{N_l(Z^1)}{n-1} - \alpha_l^*\right|\right] \\
\leq& \sqrt{\mathbb{E}_{\theta^*}\left[\left(\frac{N_q(Z^1)}{n} - \alpha_q^*\right)^2\right]\mathbb{E}_{\theta^*}\left[\frac{N_l(Z^1)^2}{(n-1)^2}\right]} \\
&+ \alpha_q^*\sqrt{\mathbb{E}_{\theta^*}\left[\left(\frac{N_l(Z^1)}{n-1} - \alpha_l^*\right)^2\right]}.
\end{aligned}$$

We have for any $q \in [\![1, Q]\!]$

$$\mathbb{E}_{\theta^*}\left[N_q(Z^1)^2\right] = \sum_{i,j}\mathbb{E}_{\theta^*}\left[\mathbb{1}_{Z_i^1=q}\mathbb{1}_{Z_j^1=q}\right] = \sum_i \alpha_q^* + \sum_{i\neq j}\alpha_q^{*2} = n\alpha_q^* + n(n-1)\alpha_q^{*2}.$$

This implies that

$$\begin{aligned}
\mathbb{E}_{\theta^*}\left[\left(\frac{N_q(Z^1)}{n} - \alpha_q^*\right)^2\right] &= \mathbb{E}_{\theta^*}\left[\frac{N_q(Z^1)^2}{n^2}\right] - \alpha_q^{*2} \\
&= \frac{1}{n}\alpha_q^* + \frac{n-1}{n}\alpha_q^{*2} - \alpha_q^{*2} = \frac{1}{n}\alpha_q^*(1-\alpha_q^*),
\end{aligned}$$

and identically

$$\mathbb{E}_{\theta^*}\left[\left(\frac{N_l(Z^1)}{n-1}-\alpha_l^*\right)^2\right]=\mathbb{E}_{\theta^*}\left[\frac{N_l(Z^1)^2}{(n-1)^2}\right]+\alpha_l^{*2}-2\frac{n}{n-1}\alpha_l^{*2}$$
$$=\frac{n}{(n-1)^2}\alpha_l^*-\frac{1}{n-1}\alpha_l^{*2}=\frac{1}{n-1}\alpha_l^*\left(\frac{n}{n-1}-\alpha_l^*\right).$$

This leads to

$$\mathbb{E}_{\theta^*}\left[\left|\frac{N_q(Z^1)N_l(Z^1)}{n(n-1)}-\alpha_q^*\alpha_l^*\right|\right]\leq\sqrt{\frac{1}{n}\alpha_q^*(1-\alpha_q^*)\left(\frac{n}{(n-1)^2}\alpha_q^*+\frac{n}{n-1}\alpha_q^{*2}\right)}$$
$$+\alpha_q^*\sqrt{\frac{1}{n-1}\alpha_l^*\left(\frac{n}{n-1}-\alpha_l^*\right)}$$
$$\leq\sqrt{\frac{1}{(n-1)^2}+\frac{1}{n-1}}+\sqrt{\frac{n}{(n-1)^2}}\leq 2\frac{\sqrt{n}}{n-1},\quad\text{(A.2.4)}$$

using the fact that $0\leq\alpha_q^*\leq 1$ for every $q\in[\![1,Q]\!]$. □

A.2.8 Proof of Lemma 2.5.6

We first consider the case when $T\to\infty$, and π is constant over time. We use the following lemma.

Lemma A.2.2. *For any $\theta\in\Theta$, we have for ϵ small enough (precisely $0<\epsilon<\min_{1\leq q\neq q'\leq Q}\max_{1\leq l\leq Q}|\pi_{ql}^*-\pi_{q'l}^*|/2$)*

$$\min_{\sigma\in\mathfrak{S}_Q}\|\pi_\sigma-\pi^*\|_\infty>\epsilon\implies\mathbb{M}(\pi^*)-\mathbb{M}(\pi)>\frac{2\delta^2}{Q^2}\epsilon^2.$$

This gives an upper bound on the probability of interest

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma-\pi^*\|_\infty>\epsilon\sqrt{v_{n,T}}\right)\leq\mathbb{P}_{\theta^*}\left(\mathbb{M}(\pi^*)-\mathbb{M}(\hat{\pi})>\frac{2\delta^2}{Q^2}\epsilon^2 v_{n,T}\right).$$

By definition of $\hat{\theta}=(\hat{\Gamma},\hat{\pi})$, we write

$$\mathbb{M}(\pi^*)=F_{n,T}(\hat{\Gamma},\pi^*)+\mathbb{M}(\pi^*)-F_{n,T}(\hat{\Gamma},\pi^*)\leq F_{n,T}(\hat{\Gamma},\hat{\pi})+\mathbb{M}(\pi^*)-F_{n,T}(\hat{\Gamma},\pi^*),$$

implying that

$$\mathbb{M}(\pi^*)-\mathbb{M}(\hat{\pi})\leq\left[F_{n,T}(\hat{\Gamma},\hat{\pi})-\mathbb{M}(\hat{\pi})\right]+\left[\mathbb{M}(\pi^*)-F_{n,T}(\hat{\Gamma},\pi^*)\right].$$

We then obtain the following upper bound, that converges to 0 as n and T increase by assumption,

$$\mathbb{P}_{\theta^*}\left(\min_{\sigma\in\mathfrak{S}_Q}\|\hat{\pi}_\sigma-\pi^*\|_\infty>\epsilon\sqrt{v_{n,T}}\right)\leq\mathbb{P}_{\theta^*}\left(F_{n,T}(\hat{\Gamma},\hat{\pi})-\mathbb{M}(\hat{\pi})>\frac{\delta^2}{Q^2}\epsilon^2 v_{n,T}\right)+\mathbb{P}_{\theta^*}\left(\mathbb{M}(\pi^*)-F_{n,T}(\hat{\Gamma},\pi^*)>\frac{\delta^2}{Q^2}\epsilon^2 v_{n,T}\right).$$

When the number of time steps T is fixed and π is allowed to vary over time, the proof is almost the same. Indeed, $\min_{\sigma^1,\dots,\sigma^T\in\mathfrak{S}_Q}\|\hat{\pi}^{1:T}_{\sigma^{1:T}}-\pi^{*1:T}\|_\infty>\epsilon\sqrt{v_n}$ means that there exists $t\in[\![1,T]\!]$ such that $\min_{\sigma^t\in\mathfrak{S}_Q}\|\hat{\pi}^t_{\sigma^t}-\pi^{*t}\|_\infty>\epsilon\sqrt{v_n}$ and we can apply Lemma A.2.2 to this $\hat{\pi}^t$ to obtain that $\mathbb{M}(\pi^{*t})-\mathbb{M}(\hat{\pi}^t)>2\epsilon^2\delta^2 v_n/Q^2$. This implies that $\mathbb{M}^T(\pi^{*1:T})-\mathbb{M}^T(\hat{\pi}^{1:T})>2\epsilon^2\delta^2 v_n/(TQ^2)$, which allows to conclude in the same way as before. □

A.2.9 Proof of Lemma 2.5.7

We have

$$\begin{aligned}\log\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T}=z^{1:T}\mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T}=z^{*1:T}\mid X^{1:T})}&=\log\frac{\mathbb{P}_{\breve{\theta}}(X^{1:T}\mid Z^{1:T}=z^{1:T})}{\mathbb{P}_{\breve{\theta}}(X^{1:T}\mid Z^{1:T}=z^{*1:T})}+\log\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T}=z^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T}=z^{*1:T})}\\&=\sum_{t=1}^{T}\sum_{1\leq i<j\leq n}\left(X^t_{ij}\log\frac{\breve{\pi}_{z^t_iz^t_j}}{\breve{\pi}_{z^{*t}_iz^{*t}_j}}+(1-X^t_{ij})\log\frac{1-\breve{\pi}_{z^t_iz^t_j}}{1-\breve{\pi}_{z^{*t}_iz^{*t}_j}}\right)\\&\quad+\sum_{i=1}^{n}\log\frac{\breve{\alpha}_{z^1_i}}{\breve{\alpha}_{z^{*1}_i}}+\sum_{t=1}^{T-1}\sum_{i=1}^{n}\log\frac{\breve{\gamma}_{z^t_iz^{t+1}_i}}{\breve{\gamma}_{z^{*t}_iz^{*t+1}_i}}.\end{aligned}$$

We decompose this sum as

$$\begin{aligned}&\log\frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T}=z^{1:T}\mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T}=z^{*1:T}\mid X^{1:T})}\\&=\sum_{t=1}^{T}\sum_{1\leq i<j\leq n}\left(X^t_{ij}\log\frac{\pi^*_{z^t_iz^t_j}}{\pi^*_{z^{*t}_iz^{*t}_j}}+(1-X^t_{ij})\log\frac{1-\pi^*_{z^t_iz^t_j}}{1-\pi^*_{z^{*t}_iz^{*t}_j}}\right)\\&\quad+\sum_{i=1}^{n}\log\frac{\breve{\alpha}_{z^1_i}}{\breve{\alpha}_{z^{*1}_i}}+\sum_{t=1}^{T-1}\sum_{i=1}^{n}\log\frac{\breve{\gamma}_{z^t_iz^{t+1}_i}}{\breve{\gamma}_{z^{*t}_iz^{*t+1}_i}}\\&\quad+\sum_{t=1}^{T}\sum_{1\leq i<j\leq n}\left(X^t_{ij}\log\frac{\breve{\pi}_{z^t_iz^t_j}}{\pi^*_{z^t_iz^t_j}}\frac{\pi^*_{z^{*t}_iz^{*t}_j}}{\breve{\pi}_{z^{*t}_iz^{*t}_j}}+(1-X^t_{ij})\log\frac{1-\breve{\pi}_{z^t_iz^t_j}}{1-\pi^*_{z^t_iz^t_j}}\frac{1-\pi^*_{z^{*t}_iz^{*t}_j}}{1-\breve{\pi}_{z^{*t}_iz^{*t}_j}}\right).\end{aligned}\tag{A.2.5}$$

In the first sum of the right-hand side of (A.2.5), the terms are different from zero only for triplets (i,j,t) in D^*. Similarly in the last sum, the terms are different from zero for triplets (i,j,t) in $D^* \cup \breve{D}$. As a consequence, we obtain

$$\begin{aligned}
&\log \frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} \\
&= \sum_{(i,j,t)\in D^*} \left(X_{ij}^t \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} + (1 - X_{ij}^t) \log \frac{1 - \pi^*_{z_i^t z_j^t}}{1 - \pi^*_{z_i^{*t} z_j^{*t}}} \right) + \sum_{i=1}^n \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \\
&\quad + \sum_{(i,j,t)\in D^*\cup\breve{D}} \left(X_{ij}^t \log \frac{\breve{\pi}_{z_i^t z_j^t}}{\pi^*_{z_i^t z_j^t}} \frac{\pi^*_{z_i^{*t} z_j^{*t}}}{\breve{\pi}_{z_i^{*t} z_j^{*t}}} + (1 - X_{ij}^t) \log \frac{1 - \breve{\pi}_{z_i^t z_j^t}}{1 - \pi^*_{z_i^t z_j^t}} \frac{1 - \pi^*_{z_i^{*t} z_j^{*t}}}{1 - \breve{\pi}_{z_i^{*t} z_j^*}} \right).
\end{aligned}$$

We now write the last sum in the right-hand side as

$$\begin{aligned}
&\sum_{(i,j,t)\in D^*\cup\breve{D}} \left(X_{ij}^t \log \frac{\breve{\pi}_{z_i^t z_j^t}}{\pi^*_{z_i^t z_j^t}} \frac{\pi^*_{z_i^{*t} z_j^{*t}}}{\breve{\pi}_{z_i^{*t} z_j^{*t}}} + (1 - X_{ij}^t) \log \frac{1 - \breve{\pi}_{z_i^t z_j^t}}{1 - \pi^*_{z_i^t z_j^t}} \frac{1 - \pi^*_{z_i^{*t} z_j^{*t}}}{1 - \breve{\pi}_{z_i^{*t} z_j^{*t}}} \right) \\
&= \sum_{(i,j,t)\in D^*\cup\breve{D}} \left\{ X_{ij}^t \left[\log \left(1 + \frac{\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^t z_j^t}} \right) + \log \frac{\pi^*_{z_i^{*t} z_j^{*t}}}{\breve{\pi}_{z_i^{*t} z_j^{*t}}} \right] \right. \\
&\qquad \left. + (1 - X_{ij}^t) \left[\log \left(1 - \frac{\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t}}{1 - \pi^*_{z_i^t z_j^t}} \right) + \log \frac{1 - \pi^*_{z_i^{*t} z_j^{*t}}}{1 - \breve{\pi}_{z_i^{*t} z_j^{*t}}} \right] \right\}.
\end{aligned}$$

Distinguishing between the cases where $X_{ij}^t = 1$ and $X_{ij}^t = 0$, we obtain

$$\begin{aligned}
&\sum_{(i,j,t)\in D^*\cup\breve{D}} \left(X_{ij}^t \log \frac{\breve{\pi}_{z_i^t z_j^t}}{\pi^*_{z_i^t z_j^t}} \frac{\pi^*_{z_i^{*t} z_j^{*t}}}{\breve{\pi}_{z_i^{*t} z_j^{*t}}} + (1 - X_{ij}^t) \log \frac{1 - \breve{\pi}_{z_i^t z_j^t}}{1 - \pi^*_{z_i^t z_j^t}} \frac{1 - \pi^*_{z_i^{*t} z_j^{*t}}}{1 - \breve{\pi}_{z_i^{*t} z_j^{*t}}} \right) \\
&= \sum_{(i,j,t)\in D^*\cup\breve{D}} \log \left[1 + \frac{(\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t})(X_{ij}^t - \pi^*_{z_i^t z_j^t})}{\pi^*_{z_i^t z_j^t}(1 - \pi^*_{z_i^t z_j^t})} \right] \\
&\quad - \sum_{(i,j,t)\in D^*\cup\breve{D}} \log \left[1 + \frac{(\breve{\pi}_{z_i^{*t} z_j^{*t}} - \pi^*_{z_i^{*t} z_j^{*t}})(X_{ij}^t - \pi^*_{z_i^{*t} z_j^{*t}})}{\pi^*_{z_i^{*t} z_j^{*t}}(1 - \pi^*_{z_i^{*t} z_j^{*t}})} \right].
\end{aligned}$$

In the end, we decompose

$$\begin{aligned}\log \frac{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{1:T} \mid X^{1:T})}{\mathbb{P}_{\breve{\theta}}(Z^{1:T} = z^{*1:T} \mid X^{1:T})} = &\sum_{(i,j,t)\in D^*} \left(X_{ij}^t \log \frac{\pi^*_{z_i^t z_j^t}}{\pi^*_{z_i^{*t} z_j^{*t}}} + (1 - X_{ij}^t) \log \frac{1 - \pi^*_{z_i^t z_j^t}}{1 - \pi^*_{z_i^{*t} z_j^{*t}}} \right) \\ &+ \sum_{i=1}^n \log \frac{\breve{\alpha}_{z_i^1}}{\breve{\alpha}_{z_i^{*1}}} + \sum_{t=1}^{T-1} \sum_{i=1}^n \log \frac{\breve{\gamma}_{z_i^t z_i^{t+1}}}{\breve{\gamma}_{z_i^{*t} z_i^{*t+1}}} \\ &+ \sum_{(i,j,t)\in D^*\cup\breve{D}} \log \left[1 + \frac{(\breve{\pi}_{z_i^t z_j^t} - \pi^*_{z_i^t z_j^t})(X_{ij}^t - \pi^*_{z_i^t z_j^t})}{\pi^*_{z_i^t z_j^t}(1 - \pi^*_{z_i^t z_j^t})} \right] \\ &- \sum_{(i,j,t)\in D^*\cup\breve{D}} \log \left[1 + \frac{(\breve{\pi}_{z_i^{*t} z_j^{*t}} - \pi^*_{z_i^{*t} z_j^{*t}})(X_{ij}^t - \pi^*_{z_i^{*t} z_j^{*t}})}{\pi^*_{z_i^{*t} z_j^{*t}}(1 - \pi^*_{z_i^{*t} z_j^{*t}})} \right],\end{aligned}$$

which gives the result.

A.2.10 Proof of Lemma 2.5.8

We first notice that

$$\begin{aligned}\left| D_{n,T}(z^{1:T}, \pi) \right| =& \frac{1}{2} \left| \left\{ (i,j,t) \in [\![1,n]\!]^2 \times [\![1,T]\!]; \pi_{z_i^t z_j^t} \neq \pi_{z_i^{*t} z_j^{*t}} \right\} \right| \\ =& \frac{1}{2} \sum_{t=1}^T \left| \left\{ (i,j) \in [\![1,n]\!]^2; \pi_{z_i^t z_j^t} \neq \pi_{z_i^{*t} z_j^{*t}} \right\} \right|.\end{aligned}$$

For every $t \in [\![1,T]\!]$, we can apply Proposition B.4. from Celisse et al. (2012), as their Assumption (A4) is required to hold only for z^{*t} (see proof) and is valid on $\Omega_\eta(\theta)$ with the constant $\delta - \eta$. We obtain

$$\left| \left\{ (i,j) \in [\![1,n]\!]^2; \pi_{z_i^t z_j^t} \neq \pi_{z_i^{*t} z_j^{*t}} \right\} \right| \geq \frac{(\delta - \eta)^2}{2} n r(t).$$

We conclude by noticing that $\sum_{t=1}^T r(t) = r$.

A.2.11 Proof of Lemma 2.5.9

The inclusion of the sets is straightforward. Now we have

$$\begin{aligned}
&\left|\left\{(i,j,t)\in[\![1,n]\!]^2\times[\![1,T]\!];\pi_{z_i^t z_j^t}\neq\pi_{z_i^{*t}z_j^{*t}}\right\}\right|\\
\leq&\left|\left\{(i,j,t)\in[\![1,n]\!]^2\times[\![1,T]\!];(z_i^t,z_j^t)\neq(z_i^{*t},z_j^{*t})\right\}\right|\\
\leq&\left|\left\{(i,j,t)\in[\![1,n]\!]^2\times[\![1,T]\!];z_i^t\neq z_i^{*t}\right\}\right|+\left|\left\{(i,j,t)\in[\![1,n]\!]^2\times[\![1,T]\!];z_j^t\neq z_j^{*t}\right\}\right|\\
\leq&2\sum_{t=1}^{T}nr(t)\leq 2nr.
\end{aligned}$$

A.2.12 Proof of Lemma 2.5.10

First, let us decompose the quantity at stake as follows

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat\theta_\sigma}\left(Z_i^t=q,Z_i^{t+1}=l\mid X^{1:T}\right)-\alpha_q^*\gamma_{ql}^*\right|>\epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat\theta_\sigma}\left(Z_i^t=q,Z_i^{t+1}=l\mid X^{1:T}\right)-\frac{N_{ql}(Z^{1:T})}{n(T-1)}\right|>\frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
&+\mathbb{P}_{\theta^*}\left(\left|\frac{N_{ql}(Z^{1:T})}{n(T-1)}-\alpha_q^*\gamma_{ql}^*\right|>\frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right),
\end{aligned}\tag{A.2.6}$$

and upper bound the two terms in the right-hand side of (A.2.6). For the first one we will follow the proof of Theorem 3.9 from Celisse et al. (2012). Let $z^{1:T}$ denote a fixed configuration. We work on the set $\{Z^{1:T}=z^{1:T}\}$ and write

$$\begin{aligned}
V_1(z^{1:T}):=&\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat\theta_\sigma}\left(Z_i^t=q,Z_i^{t+1}=l\mid X^{1:T}\right)-\frac{N_{ql}(z^{1:T})}{n(T-1)}\right|\\
\leq&\left|\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat\theta_\sigma}\left(Z_i^t=q,Z_i^{t+1}=l\mid X^{1:T}\right)\mathbb{1}_{(z_i^t,z_i^{t+1})=(q,l)}-\frac{N_{ql}(z^{1:T})}{n(T-1)}\right|\\
&+\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat\theta_\sigma}\left(Z_i^t=q,Z_i^{t+1}=l\mid X^{1:T}\right)\mathbb{1}_{(z_i^t,z_i^{t+1})\neq(q,l)}\\
\leq&\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\left(1-\mathbb{P}_{\hat\theta_\sigma}\left((Z_i^t,Z_i^{t+1})=(z_i^t,z_i^{t+1})\mid X^{1:T}\right)\right)\mathbb{1}_{(z_i^t,z_i^{t+1})=(q,l)}\\
&+\frac{1}{n(T-1)}\sum_{t=1}^{T-1}\sum_{i=1}^{n}\mathbb{P}_{\hat\theta_\sigma}\left((Z_i^t,Z_i^{t+1})\neq(z_i^t,z_i^{t+1})\mid X^{1:T}\right)\mathbb{1}_{(z_i^t,z_i^{t+1})\neq(q,l)}\\
\leq&\,2\mathbb{P}_{\hat\theta_\sigma}\left(Z^{1:T}\neq z^{1:T}\mid X^{1:T}\right).
\end{aligned}$$

Then

$$
\begin{aligned}
&\mathbb{P}_{\theta^*}\left(V_1(Z^{1:T}) > \frac{\epsilon}{2} r_{n,T} \frac{\sqrt{\log n}}{\sqrt{nT}}\right) \\
=&\mathbb{E}_{\theta^*}\left[\mathbb{P}_{\theta^*}\left(V_1(Z^{1:T}) > \frac{\epsilon}{2} r_{n,T} \frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T}\right)\right] \\
\leq& \sum_{z^{1:T}} \mathbb{P}_{\theta^*}\left(\mathbb{P}_{\hat{\theta}_\sigma}\left(Z^{1:T} \neq z^{1:T} \mid X^{1:T}\right) > \frac{\epsilon}{4} r_{n,T} \frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \mathbb{P}_{\theta^*}\left(Z^{1:T} = z^{1:T}\right) \\
\leq& \sum_{z^{1:T}} \mathbb{P}_{\theta^*}\left(\frac{\mathbb{P}_{\hat{\theta}_\sigma}\left(Z^{1:T} \neq z^{1:T} \mid X^{1:T}\right)}{\mathbb{P}_{\hat{\theta}_\sigma}\left(Z^{1:T} = z^{1:T} \mid X^{1:T}\right)} > \frac{\epsilon}{4} r_{n,T} \frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \mathbb{P}_{\theta^*}\left(Z^{1:T} = z^{1:T}\right) \\
\leq& QT \exp(-2\eta^2 n) + \mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) \\
&+ CnT \exp\left[-(\delta-\eta)^2 C_1 n + C_2 \log(nT) + C_4 \log\left(\frac{4\sqrt{nT}}{\epsilon r_{n,T}\sqrt{\log n}}\right)\right] \\
&+ CnT \exp\left[-C_3 \frac{(\log(nT))^2}{n v_{n,T}^2} + 3n \log(nT)\right],
\end{aligned}
\tag{A.2.7}
$$

where the last inequality comes from Theorem 2.3.2 where the bound is uniform with respect to $z^{1:T}$.

Now, for the second term of (A.2.6), we use the following lemma.

Lemma A.2.3. *There exist $c_1, c_2 > 0$ such that for any $\epsilon > 0$, for any sequence $\{r_{n,T}\}_{n,T\geq 1}$, we have, as long as $\epsilon r_{n,T}\sqrt{\log n}/(2\alpha_q^*\gamma_{ql}^*\sqrt{nT}) < 1$,*

$$
\mathbb{P}_{\theta^*}\left(\left|\frac{N_{ql}(Z^{1:T})}{n(T-1)} - \alpha_q^*\gamma_{ql}^*\right| > \frac{\epsilon}{2} r_{n,T} \frac{\sqrt{\log n}}{\sqrt{nT}}\right) \leq c_1 \exp\left(-c_2 \epsilon^2 r_{n,T}^2\right). \tag{A.2.8}
$$

We then combine the two upper bounds obtained in (A.2.7) and (A.2.8) in order to conclude, the assumption $\epsilon r_{n,T}\sqrt{\log n}/(2\alpha_q^*\gamma_{ql}^*\sqrt{nT}) < 1$ being satisfied for n and T large enough because $r_{n,T} = o(\sqrt{nT/\log n})$. We obtain the expected result, using the fact that $\log(T) = o(n)$, that $r_{n,T}$ increases to infinity and that $v_{n,T} = o\left(\sqrt{\log(nT)/n}\right)$,

$$
\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)} \sum_{t=1}^{T-1} \sum_{i=1}^{n} \mathbb{P}_{\hat{\theta}_\sigma}\left(Z_i^t = q, Z_i^{t+1} = l \mid X^{1:T}\right) - \alpha_q^*\gamma_{ql}^*\right| > \epsilon y_{n,T}\right) \\
\leq& \mathbb{P}_{\theta^*}\left(\|\hat{\pi}_\sigma - \pi^*\|_\infty > v_{n,T}\right) + o(1).
\end{aligned}
$$

□

A.2.13 Proof of Lemma 2.5.11

We have the following inequalities by definition of $\hat{z}^{1:T}$, $\mathcal{J}(\chi, \theta)$ and $\hat{\chi}(\theta)$ and because the Kullback-Leibler divergence is non-negative

$$\mathcal{J}(\hat{z}^{1:T}, \theta) \leq \mathcal{J}(\hat{\chi}(\theta), \theta) \leq \ell(\theta) \leq \ell_c(\theta, \hat{z}^{1:T}), \tag{A.2.9}$$

with $\mathcal{J}(\hat{z}^{1:T}, \theta) = \ell(\theta) - \mathrm{KL}(\delta_{\hat{z}^{1:T}}, \mathbb{P}_\theta(\cdot|X^{1:T}))$. We write this Kullback-Leibler divergence (from $\mathbb{P}_\theta(\cdot|X^{1:T})$ to $\mathbb{Q}_\chi = \delta_{\hat{z}^{1:T}}$, with $\chi = (\tau, \eta)$ such that $\tau^t_{iq} = \hat{z}^t_{iq}$ and $\eta^t_{iql} = \hat{z}^t_{iq}\hat{z}^{t+1}_{il}$) as follows

$$\mathrm{KL}(\delta_{\hat{z}^{1:T}}, \mathbb{P}_\theta(\cdot|X^{1:T})) = -\log \mathbb{P}_\theta(\hat{z}^{1:T}|X^{1:T}).$$

We then obtain

$$\begin{aligned}\mathcal{J}(\hat{z}^{1:T}, \theta) =& \log \mathbb{P}_\theta(X^{1:T}) + \log \mathbb{P}_\theta(\hat{z}^{1:T}|X^{1:T}) = \mathbb{P}_\theta(X^{1:T}|\hat{z}^{1:T}) + \log \mathbb{P}_\theta(\hat{z}^{1:T}) \\ =& \ell_c(\theta; \hat{z}^{1:T}) + \sum_{i=1}^n \log \alpha_{\hat{z}^1_i} + \sum_{i=1}^n \sum_{t=2}^T \log \gamma_{\hat{z}^{t-1}_i \hat{z}^t_i}.\end{aligned}$$

Combined with (A.2.9), this leads to the following inequality for any parameter $\theta \in \Theta$

$$\begin{aligned}|\mathcal{J}(\hat{\chi}(\theta), \theta) - \ell(\theta)| \leq \left|\mathcal{J}(\hat{z}^{1:T}, \theta) - \ell_c(\theta, \hat{z}^{1:T})\right| &\leq -\sum_{i=1}^n \log \alpha_{\hat{z}^1_i} - \sum_{i=1}^n \sum_{t=2}^T \log \gamma_{\hat{z}^{t-1}_i \hat{z}^t_i} \\ &\leq nT \log(1/\delta).\end{aligned}$$

We can conclude that

$$\sup_{\theta \in \Theta} \left| \frac{2}{n(n-1)T} \mathcal{J}(\hat{\chi}(\theta), \theta) - \frac{2}{n(n-1)T} \ell(\theta) \right| \leq \frac{2 \log(1/\delta)}{n-1}.$$

□

A.2.14 Proof of Lemma 2.5.12

This proof is quite similar to that of Lemma 2.5.10. For any $\epsilon > 0$, let us write

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)-\alpha_q^*\gamma_{ql}^*\right|>\epsilon r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)-\frac{N_{ql}(Z^{1:T})}{n(T-1)}\right|>\frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
&+\mathbb{P}_{\theta^*}\left(\left|\frac{N_{ql}(Z^{1:T})}{n(T-1)}-\alpha_q^*\gamma_{ql}^*\right|>\frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)
\end{aligned}$$

and upper bound the two probabilities in the right-hand side of this inequality. We already proved in Lemma 2.5.10 that the second term converges to 0 thanks to the assumptions on the sequence $\{r_{n,T}\}_{n,T\geq 1}$. For the first term, let $z^{1:T}$ denote a fixed configuration. Let us work on the set $\{Z^{1:T}=z^{1:T}\}$ and use the same method as in the proof of Lemma 2.5.10,

$$\begin{aligned}
&\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)\\
=&\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)\mathbb{1}_{z_i^t=q,z_i^{t+1}=l}\\
&+\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)\mathbb{1}_{(z_i^t,z_i^{t+1})\neq(q,l)},
\end{aligned}$$

leading to

$$\left|\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)-\frac{N_{ql}(z^{1:T})}{n(T-1)}\right|\leq 2\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z^{1:T}\neq z^{1:T}).$$

Then we obtain

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T-1}\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z_i^t=q,Z_i^{t+1}=l)-\frac{N_{ql}(Z^{1:T})}{n(T-1)}\right|>\frac{\epsilon}{2}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right)\\
\leq&\sum_{z^{1:T}}\mathbb{P}_{\theta^*}\left(\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z^{1:T}\neq z^{1:T})>\frac{\epsilon}{4}r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\,\middle|\,Z^{1:T}=z^{1:T}\right)\mathbb{P}_{\theta^*}\left(Z^{1:T}=z^{1:T}\right).
\end{aligned}$$

For each $z^{1:T}$, we use the following lemma.

Lemma A.2.4. *Denoting* $\tilde{\mathbb{P}}_\sigma(\cdot) = \mathbb{P}_{\tilde{\theta}_\sigma}(Z^{1:T} = \cdot \mid X^{1:T})$*, we have the following inequality for any configuration* $z^{1:T}$

$$\left|\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(z^{1:T}) - \tilde{\mathbb{P}}_\sigma(z^{1:T})\right| \leq \sqrt{-\frac{1}{2}\log\left(\tilde{\mathbb{P}}_\sigma(z^{1:T})\right)}.$$

This gives us

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z^{1:T} \neq z^{1:T}) > \frac{\epsilon}{4} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \\
\leq &\mathbb{P}_{\theta^*}\left(\left|\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(Z^{1:T} \neq z^{1:T}) - \tilde{\mathbb{P}}_\sigma(Z^{1:T} \neq z^{1:T})\right| > \frac{\epsilon}{8} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \\
&+ \mathbb{P}_{\theta^*}\left(\tilde{\mathbb{P}}_\sigma(Z^{1:T} \neq z^{1:T}) > \frac{\epsilon}{8} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \\
\leq &\mathbb{P}_{\theta^*}\left(\sqrt{-\frac{1}{2}\log\left(\tilde{\mathbb{P}}_\sigma(z^{1:T})\right)} > \frac{\epsilon}{8} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \\
&+ \mathbb{P}_{\theta^*}\left(\tilde{\mathbb{P}}_\sigma(Z^{1:T} \neq z^{1:T}) > \frac{\epsilon}{8} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right) \\
\leq &\mathbb{P}_{\theta^*}\left(\tilde{\mathbb{P}}_\sigma(Z^{1:T} \neq z^{1:T}) > 1 - \exp\left(-\frac{\epsilon^2 r_{n,T}^2 \log n}{32nT}\right) \;\middle|\; Z^{1:T} = z^{1:T}\right) \\
&+ \mathbb{P}_{\theta^*}\left(\tilde{\mathbb{P}}_\sigma(Z^{1:T} \neq z^{1:T}) > \frac{\epsilon}{8} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}} \;\middle|\; Z^{1:T} = z^{1:T}\right). \qquad \text{(A.2.10)}
\end{aligned}$$

Noticing that the assumptions on $\{r_{n,T}\}_{n,T\geq 1}$ imply that

$$-\log\left[1 - \exp\left(-\frac{\epsilon^2 r_{n,T}^2 \log n}{32nT}\right)\right] = o(n) \quad \text{and} \quad -\log\left[r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right] = o(n),$$

we can conclude by applying the result of Theorem 2.3.2 with the estimator $\tilde{\theta}_\sigma = (\tilde{\Gamma}_\sigma, \tilde{\pi}_\sigma)$ for both terms of the right-hand side of (A.2.10). □

A.2.15 Proof of Lemma A.2.1

The proof follows the lines of the proof of Lemma C.3. from Celisse et al. (2012). Let $\mathbb{E}^*_{\theta^*}[\cdot]$ denote the expectation given $Z^{1:T} = z^{*1:T}$, i.e. $\mathbb{E}^*_{\theta^*}[\cdot] = \mathbb{E}_{\theta^*}[\cdot \mid Z^{1:T} = z^{*1:T}]$. Introducing a ghost sample $\{\tilde{X}^t_{ij}\}_{i,j,t}$ that is independent of $\{X^t_{ij}\}_{i,j,t}$ and has the same

distribution, we write

$$
\begin{aligned}
E :=& \mathbb{E}^*_{\theta^*}\left[\sup_{\substack{(z^{1:T},\pi)\in \\ [\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}}\left|\frac{2}{n(n-1)T}\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t-\pi^*_{z_i^{*t}z_j^{*t}})\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right] \\
=& \mathbb{E}^*_{\theta^*}\left\{\sup_{\substack{(z^{1:T},\pi)\in \\ [\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}}\left|\frac{2}{n(n-1)T}\mathbb{E}^*_{\theta^*}\left[\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\;\middle|\;\{X_{ij}^t\}_{i,j,t}\right]\right|\right\} \\
\leq& \mathbb{E}^*_{\theta^*}\left\{\mathbb{E}^*_{\theta^*}\left[\sup_{\substack{(z^{1:T},\pi)\in \\ [\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}}\frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\;\middle|\;\{X_{ij}^t\}_{i,j,t}\right]\right\} \\
\leq& \mathbb{E}^*_{\theta^*,X,\tilde{X}}\left[\sup_{\substack{(z^{1:T},\pi)\in \\ [\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}}\frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right],
\end{aligned}
$$

where $\mathbb{E}^*_{\theta^*,X,\tilde{X}}[\cdot]$ denotes the expectation with respect to $\{X,\tilde{X}\}=\{X_{ij}^t,\tilde{X}_{ij}^t\}_{i,j,t}$ under the true parameter θ^* and given $Z^{1:T}=z^{*1:T}$. At this point, we notice that, if $\{\epsilon_{ij}^t\}_{i,j,t}\coloneqq\epsilon$ are n^2T independent Rademacher variables, then the random variables

$$
\mathbb{E}_\epsilon\left|\sum_{t=1}^{T}\sum_{i<j}\epsilon_{ij}^t(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right| \quad\text{and}\quad \left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|
$$

follow the same distribution, which implies that

$$
\begin{aligned}
&\mathbb{E}^*_{\theta^*,X,\tilde{X}}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}\frac{2}{n(n-1)T}\mathbb{E}_\epsilon\left|\sum_{t=1}^{T}\sum_{i<j}\epsilon_{ij}^t(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right] \\
&\quad=\mathbb{E}^*_{\theta^*,X,\tilde{X}}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}}\frac{2}{n(n-1)T}\left|\sum_{t=1}^{T}\sum_{i<j}(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right].
\end{aligned}
$$

As a consequence, we have

$$
\begin{aligned}
E \leq & \mathbb{E}^*_{\theta^*,X,\tilde{X}}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\mathbb{E}_\epsilon\left|\sum_{t=1}^T\sum_{i<j}\epsilon_{ij}^t(X_{ij}^t-\tilde{X}_{ij}^t)\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right] \\
\leq & \mathbb{E}^*_{\theta^*}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\mathbb{E}_\epsilon\left|\sum_{t=1}^T\sum_{i<j}\epsilon_{ij}^t X_{ij}^t\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right] \\
& + \mathbb{E}^*_{\theta^*}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\mathbb{E}_\epsilon\left|\sum_{t=1}^T\sum_{i<j}\epsilon_{ij}^t \tilde{X}_{ij}^t\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right] \\
\leq & 2\mathbb{E}^*_{\theta^*}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\mathbb{E}_\epsilon\left|\sum_{t=1}^T\sum_{i<j}\epsilon_{ij}^t X_{ij}^t\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right|\right].
\end{aligned}
$$

Then using Jensen's inequality, Assumption 3 and the bound $\mathrm{Var}_\epsilon(\epsilon_{ij}^t X_{ij}^t)\leq 1$, we get

$$
\begin{aligned}
E \leq & 2\mathbb{E}^*_{\theta^*}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\sqrt{\mathbb{E}_\epsilon\left[\left(\sum_{t=1}^T\sum_{i<j}\epsilon_{ij}^t X_{ij}^t\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right)^2\right]}\right] \\
\leq & 2\mathbb{E}^*_{\theta^*}\left[\sup_{(z^{1:T},\pi)\in[\![1,Q]\!]^{nT}\times[\zeta,1-\zeta]^{Q^2}} \frac{2}{n(n-1)T}\sqrt{\mathrm{Var}_\epsilon\left[\sum_{t=1}^T\sum_{i<j}\epsilon_{ij}^t X_{ij}^t\log\left(\frac{\pi_{z_i^tz_j^t}}{1-\pi_{z_i^tz_j^t}}\right)\right]}\right] \\
\leq & 2\mathbb{E}^*_{\theta^*}\left[\frac{2}{n(n-1)T}\sup_{\pi\in[\zeta,1-\zeta]}\log\left(\frac{\pi}{1-\pi}\right)\sqrt{\frac{n(n-1)T}{2}}\right] \leq \sqrt{\frac{2}{n(n-1)T}}\Lambda,
\end{aligned}
$$

where $\Lambda = 2\log[(1-\zeta)/\zeta]$, concluding the proof. □

A.2.16 Proof of Lemma A.2.2

We assume that $\min_{\sigma\in\mathfrak{S}_Q}\|\pi_\sigma-\pi^*\|_\infty>\epsilon$. Without loss of generality, assume that the permutation (or one of the permutations) minimizing this distance is the identity. Let us write, using the fact that I_Q the identity matrix of size Q maximizes in A (over the set of $Q\times Q$ stochastic matrices) the quantity $\mathbb{M}(\pi^*,A)$ (see the proof of Theorem 3.6 in Celisse et al. (2012)) and denoting $(\bar{a}_{qq'})_{q,q'\in[\![1,Q]\!]}$ the coefficients of $\bar{A}_\pi$ (thus depending on π),

$$
\begin{aligned}
\mathbb{M}(\pi^*)-\mathbb{M}(\pi) &= \sum_{q,l}\alpha_q^*\alpha_l^*\sum_{q',l'}\bar{a}_{qq'}\bar{a}_{ll'}\left[\pi_{ql}^*\log\frac{\pi_{ql}^*}{\pi_{q'l'}}+(1-\pi_{ql}^*)\log\frac{1-\pi_{ql}^*}{1-\pi_{q'l'}}\right] \\
&= \sum_{q,l}\alpha_q^*\alpha_l^*\sum_{q',l'}\bar{a}_{qq'}\bar{a}_{ll'}K(\pi_{ql}^*,\pi_{q'l'})
\end{aligned}
$$

denoting $K(p_1, p_2) = p_1 \log(p_1/p_2) + (1-p_1)\log[(1-p_1)/(1-p_2)] > 0$ the Kullback-Leibler divergence from a Bernoulli distribution with parameter p_2 to a Bernoulli distribution with parameter p_1. For every q, there exists $q' \coloneqq f(q)$ such that $\bar{a}_{qq'} \geq 1/Q$ because $\bar{A}_\pi$ is a stochastic matrix. Using Assumption 2, we obtain

$$\mathbb{M}(\pi^*) - \mathbb{M}(\pi) \geq \frac{\delta^2}{Q^2} \sum_{q,l} K(\pi^*_{ql}, \pi_{f(q)f(l)}) \geq \frac{\delta^2}{Q^2} \sum_{q,l} 2(\pi^*_{ql} - \pi_{f(q)f(l)})^2$$

thanks to a result on Kullback-Leibler divergence for Bernoulli distributions (see for instance Bubeck (2010), Chapter 10, Section 2, Lemma 10.3). We then want to show that there exist q, l such that $|\pi^*_{ql} - \pi_{f(q)f(l)}| > \epsilon$.

- If f is a permutation, the assumption $\min_{\sigma \in \mathfrak{S}_Q} \|\pi_\sigma - \pi^*\|_\infty > \epsilon$ gives the expected result.
- If f is not a permutation, it is not injective and there exist $q_1 \neq q_2$ such that $f(q_1) = f(q_2)$. Thanks to Assumption 1, take $l_0 \in [\![1, Q]\!]$ such that $|\pi_{q_1 l_0} - \pi_{q_2 l_0}| = \max_{l \in [\![1,Q]\!]} |\pi_{q_1 l} - \pi_{q_2 l}| > 0$. Then

$$\begin{aligned} |\pi^*_{q_1 l_0} - \pi_{f(q_1)f(l_0)}| + |\pi_{f(q_2)f(l_0)} - \pi^*_{q_2 l_0}| &\geq |\pi^*_{q_1 l_0} - \pi_{f(q_1)f(l_0)} + \pi_{f(q_2)f(l_0)} - \pi^*_{q_2 l_0}| \\ &\geq |\pi^*_{q_1 l_0} - \pi^*_{q_2 l_0}| > 0 \end{aligned}$$

  leading to either $|\pi^*_{q_1 l_0} - \pi_{f(q_1)f(l_0)}| \geq |\pi^*_{q_1 l_0} - \pi^*_{q_2 l_0}|/2 > \epsilon$ or $|\pi^*_{q_2 l_0} - \pi_{f(q_2)f(l_0)}| \geq |\pi^*_{q_1 l_0} - \pi^*_{q_2 l_0}|/2 > \epsilon$, using the fact that $\epsilon < \min_{1 \leq q \neq q' \leq Q} \max_{1 \leq l \leq Q} |\pi^*_{ql} - \pi^*_{q'l}|/2$.

So, as there exist q, l such that $|\pi^*_{ql} - \pi_{f(q)f(l)}| > \epsilon$, we have

$$\mathbb{M}(\pi^*) - \mathbb{M}(\pi) > \frac{2\delta^2}{Q^2}\epsilon^2.$$

□

A.2.17 Proof of Lemma A.2.3

For any node $i \in [\![1, n]\!]$, the Markov chain $\{Z_i^t\}_{t \geq 1}$ is geometrically ergodic because its transition matrix Γ satisfies Doeblin's condition thanks to Assumption 2. For any $z \in [\![1, Q]\!]$, let us denote δ_z the Dirac mass at z. There exists a positive constant A and some $r \in (0, 1)$ such that $\forall q \in [\![1, Q]\!]$ and $\forall t \geq 1$, we have

$$\left\| \delta_q \Gamma^t - \alpha \right\|_{TV} \leq A r^t,$$

where $\|\cdot\|_{TV}$ is the total variation norm. This leads to

$$\left\|\delta_q\Gamma^t-\alpha\right\|_{TV}=\frac{1}{2}\left\|\delta_q\Gamma^t-\alpha\right\|_1=\frac{1}{2}\sum_{l\in[\![1,Q]\!]}|\Gamma^t(q,l)-\alpha_l|\leq Ar^t.$$

We now consider the Markov chain $\{Z^t=(Z_1^t,\dots,Z_n^t)\}_{t\geq 1}$ of the n nodes evolving through time. Note that it is irreducible and aperiodic. Moreover, its transition matrix is given by $P_n=\Gamma^{\otimes n}$, the n-th Kronecker power of Γ and its stationary distribution is $\alpha^{\otimes n}$. For any $z=(z_1,\dots,z_n)\in[\![1,Q]\!]^n$, let us denote $\mu_{n,z}=\otimes_{i=1}^n\delta_{z_i}$. For every $t\geq 1$, we can decompose

$$\begin{aligned}\left\|\mu_{n,z}P_n^t-\alpha^{\otimes n}\right\|_{TV}&=\left\|\left(\bigotimes_{i=1}^n\delta_{z_i}\right)(\Gamma^{\otimes n})^t-\alpha^{\otimes n}\right\|_{TV}=\left\|\left(\bigotimes_{i=1}^n\delta_{z_i}\right)(\Gamma^t)^{\otimes n}-\alpha^{\otimes n}\right\|_{TV}\\&=\left\|\bigotimes_{i=1}^n\left(\delta_{z_i}\Gamma^t\right)-\alpha^{\otimes n}\right\|_{TV}=\frac{1}{2}\left\|\bigotimes_{i=1}^n\left(\delta_{z_i}\Gamma^t\right)-\alpha^{\otimes n}\right\|_1\\&=\frac{1}{2}\sum_{(z_1',\dots,z_n')\in[\![1,Q]\!]^n}\left|\prod_{i=1}^n\Gamma^t(z_i,z_i')-\prod_{i=1}^n\alpha_{z_i'}\right|.\end{aligned}$$

We use

$$\prod_{i=1}^n\Gamma^t(z_i,z_i')-\prod_{i=1}^n\alpha_{z_i'}=\sum_{i=1}^n\left\{\left(\prod_{j=1}^{i-1}\alpha_{z_j'}\right)\left[\Gamma^t(z_i,z_i')-\alpha_{z_i'}\right]\prod_{k=i+1}^n(\mu_{z_k}\Gamma^t)_{z_k'}\right\}.$$

So, reorganizing the terms, we write

$$\begin{aligned}&\left\|\mu_{n,z}P_n^t-\alpha^{\otimes n}\right\|_{TV}\\\leq&\frac{1}{2}\sum_{(z_1',\dots,z_n')\in[\![1,Q]\!]^n}\sum_{i=1}^n\left\{\left(\prod_{j=1}^{i-1}\alpha_{z_j'}\right)\left|\Gamma^t(z_i,z_i')-\alpha_{z_i'}\right|\prod_{k=i+1}^n(\mu_{z_k}\Gamma^t)_{z_k'}\right\}\\\leq&\frac{1}{2}\sum_{i=1}^n\sum_{z_1'}\alpha_{z_1'}\dots\sum_{z_{i-1}'}\alpha_{z_{i-1}'}\sum_{z_i'}\left|\Gamma^t(z_i,z_i')-\alpha_{z_i'}\right|\sum_{z_{i+1}'}\Gamma^t(z_{i+1},z_{i+1}')\dots\sum_{z_n'}\Gamma^t(z_n,z_n')\\\leq&\frac{1}{2}\sum_{i=1}^n\sum_{z_i'}\left|\Gamma^t(z_i,z_i')-\alpha_{z_i'}\right|\leq nAr^t.\end{aligned}$$

Let us recall the definition of an ϵ-mixing time. For any Markov transition matrix M over the set $\mathcal{X}$ with stationary distribution α, for any $\epsilon>0$, the ϵ-mixing time of the Markov chain is defined as

$$\tau(\epsilon)=\min\{t\geq 1;\max_{x\in\mathcal{X}}\|\delta_xM^t-\alpha\|_{TV}\leq\epsilon\}.$$

Denoting by $\tau_n(\epsilon)$ the ϵ-mixing time of the Markov chain $\{Z^t\}_{t\geq 1}$, we thus obtain

$$\tau_n(\epsilon) \leq \frac{\log(nA/\epsilon)}{\log(1/r)}.$$

Now, we introduce a new Markov chain $Y = \{Y^t\}_{t\geq 1}$, that is defined by

$$Y^t = (Z^t, Z^{t+1}) \quad \forall t \geq 1.$$

Notice that it is irreducible and aperiodic, with stationary distribution ρ defined for every state $(q_1^t, \dots, q_n^t, q_1^{t+1}, \dots, q_n^{t+1})$ by

$$\rho_{(q_1^t,\dots,q_n^t,q_1^{t+1},\dots,q_n^{t+1})} = \alpha_{q_1^t} \dots \alpha_{q_n^t} \gamma_{q_1^t q_1^{t+1}} \dots \gamma_{q_n^t q_n^{t+1}}.$$

It is easily seen that for any $\epsilon > 0$, its ϵ-mixing time $\tau_{Y,n}(\epsilon)$ equals $\tau_n(\epsilon) + 1$. We apply Theorem 3 from Chung et al. (2012), for any $\eta \leq 1/8$, considering the weight function $f(Y^t) = \sum_{i=1}^n \mathbb{1}_{Z_i^t = q, Z_i^{t+1} = l}$ for every $t \geq 1$ (of expectation $n\alpha_q^*\gamma_{ql}^*$ under the stationary distribution). Then $N_{ql}(Z^{1:T}) = \sum_{t=1}^{T-1} f(Y^t)$, and denoting $\epsilon_{n,T} = \epsilon r_{n,T}\sqrt{\log n}/(2\alpha_q^*\gamma_{ql}^*\sqrt{nT})$, we obtain that there exist $c_1, c_2 > 0$ such that for any $\epsilon > 0$, as long as $\epsilon_{n,T} \leq 1$

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left|\frac{N_{ql}(Z^{1:T})}{n(T-1)} - \alpha_q^*\gamma_{ql}^*\right| > \frac{\epsilon}{2} r_{n,T}\frac{\sqrt{\log n}}{\sqrt{nT}}\right) \\
=&\mathbb{P}_{\theta^*}\left(N_{ql}(Z^{1:T}) > (1+\epsilon_{n,T})n\alpha_q^*\gamma_{ql}^*(T-1)\right) \\
&+ \mathbb{P}_{\theta^*}\left(N_{ql}(Z^{1:T}) < (1-\epsilon_{n,T})n\alpha_q^*\gamma_{ql}^*(T-1)\right) \\
\leq& c_1 \exp\left(-\frac{\epsilon_{n,T}^2 n\alpha_q^*\gamma_{ql}^*(T-1)}{72\tau_{Y,n}(\eta)}\right) \leq c_1 \exp\left(-c_2\epsilon^2 r_{n,T}^2\right).
\end{aligned}$$

□

A.2.18 Proof of Lemma A.2.4

For any configuration $z^{1:T}$,

$$\begin{aligned}
\left|\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}(z^{1:T}) - \tilde{\mathbb{P}}_\sigma(z^{1:T})\right| \leq \left\|\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)} - \tilde{\mathbb{P}}_\sigma\right\|_{TV} &\leq \sqrt{\frac{1}{2}KL(\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}, \tilde{\mathbb{P}}_\sigma)} \leq \sqrt{\frac{1}{2}KL(\delta_{z^{1:T}}, \tilde{\mathbb{P}}_\sigma)} \\
&\leq \sqrt{-\frac{1}{2}\log\left(\tilde{\mathbb{P}}_\sigma(z^{1:T})\right)},
\end{aligned}$$

the third inequality being true because by definition $\mathbb{Q}_{\hat{\chi}(\tilde{\theta}_\sigma)}$ minimizes $KL(\cdot, \tilde{\mathbb{P}}_\sigma)$ over the set of variational distributions. □

Appendix B

Supplementary material for Chapter 3

B.1 Identifiability of the Potts model

In the following, we assume that i is fixed and omit the index i in the notation (in the parameter, random variable, location graph and normalising constant).

Proof of Lemma 3.3.1. Assume that there exists (α, β) and (α', β') (satisfying the constraint of Assumption 1 on α) inducing the same distribution on the random variable $Z^{1:L}$, i.e. such that $\mathbb{P}_{(\alpha,\beta)}(\cdot) = \mathbb{P}_{(\alpha',\beta')}(\cdot)$. Then we have for any configuration $z^{1:L} \in Q^L$ the equality

$$\begin{aligned} &- \log S(\alpha, \beta) + \sum_{q=1}^{Q} \alpha_q \sum_{l=1}^{L} \mathbb{1}_{z^l=q} + \beta \sum_{(l,l')\in E} \mathbb{1}_{z^l=z^{l'}} \\ &= - \log S(\alpha', \beta') + \sum_{q=1}^{Q} \alpha'_q \sum_{l=1}^{L} \mathbb{1}_{z^l=q} + \beta' \sum_{(l,l')\in E} \mathbb{1}_{z^l=z^{l'}}. \end{aligned} \tag{B.1.1}$$

In particular, for any $q \in [\![1, Q]\!]$, for the configuration $z^{1:L} = (q, \dots, q)$ in which every location belongs to group q, we have

$$- \log S(\alpha, \beta) + L\alpha_q + |E|\beta = - \log S(\alpha', \beta') + L\alpha'_q + |E|\beta' \tag{B.1.2}$$

where $|E|$ denotes the cardinality of E, and summing this quantity for every $q \in [\![1, Q]\!]$ and dividing by Q gives us, thanks to the constraint on α (Assumption 1)

$$- \log S(\alpha, \beta) + |E|\beta = - \log S(\alpha', \beta') + |E|\beta'. \tag{B.1.3}$$

Substracting (B.1.3) from (B.1.2) leads to $\alpha_q = \alpha'_q$ for every $q \in [\![1, Q]\!]$. Now, for the identification of β, we get from (B.1.3) that

$$-\log S(\alpha', \beta') = -\log S(\alpha, \beta) + |E|\beta - |E|\beta',$$

and using the fact that $\alpha_q = \alpha'_q$, Equation (B.1.1) becomes

$$\beta \left(\sum_{(l,l') \in E} \mathbb{1}_{z^l = z^{l'}} - |E| \right) = \beta' \left(\sum_{(l,l') \in E} \mathbb{1}_{z^l = z^{l'}} - |E| \right).$$

In particular, for any configuration $z^{1:L}$ such that at least one pair of neighbour locations in the graph $\mathcal{G}$ belong to two different groups (so that the terms between brackets are not equal to zero), we obtain that $\beta = \beta'$. □

B.2 Mean field approximation

We describe here the mean field approximation in our setup and in particular the computation of the means used in this approximation and its use in the mean field EM algorithm. Indeed, instead of using in the algorithm an approximation based on a simulated configuration of $Z^{1:L}$ as in Section 3.4.4, we could also consider at each step t of the algorithm a mean field approximation. It consists in replacing the distributions (of the Markov fields and of the Markov fields given the observations) by distributions factorising over the locations and nodes such that for each pair of location and node, we assume that the neighbours of the location and the other nodes at the same location are fixed to their expectation given the observations. For details about the mean field approximation, see Section 1.7.5.1.

Note that the mean field approximation is the one that minimises the Kullback-Leibler divergence from the true conditional distribution (see for example Blei et al. (2017)). However, the simulated EM has shown better performances than the mean field EM on experiments (Celeux et al., 2003). In particular, the mean field EM seems to be very sensitive to its initialisation (Forbes and Fort, 2007).

B.2.1 Approximation

First, let us recall that we denote $\mathbf{Z}_i^l = (Z_{iq}^l)_{1 \leq q \leq Q} \in \{0,1\}^Q$ with $Z_{iq}^l = \mathbb{1}_{Z_i^l = q}$ for every l, i and q. For this type of approximation, as we will consider the expectations $\tau_i^l \in [0,1]^Q$

of the latent variables $\mathbf{Z}_i^l$, we will need the quantities appearing in (3.4.10)[1] to be defined for $\tau_i^l = (\tau_{iq}^l)_{1\leq q\leq Q} \in [0,1]^Q$ instead of $\{0,1\}^Q$, for every i and l. We will then define the probability of the observation X^l given $Z_i^l = q$ and given the expectation value $\tau_{jq'}^l$ of $Z_{jq'}^l$ for every q' and for $j \neq i$ as follows, for any parameter θ

$$\mathbb{P}_\pi(X_{i\cdot}^l \mid Z_i^l = q, Z_{-i}^l = \tau_{-i}^l) \propto \prod_{j\neq i}\prod_{q'=1}^{Q}\left((\pi_{qq'}^l)^{X_{ij}^l}(1-\pi_{qq'}^l)^{1-X_{ij}^l}\right)^{\tau_{jq'}^l}, \tag{B.2.1}$$

and the probability of Z_i^l given the expectations of the latent variables at every neighbour location $\tau_i^{\mathcal{N}_i(l)} = (\tau_{iq}^{l'})_{1\leq q\leq Q, l'\in\mathcal{N}_i(l)}$

$$\mathbb{P}_\psi(Z_i^l \mid \mathbf{Z}_i^{\mathcal{N}_i(l)} = \tau_i^{\mathcal{N}_i(l)}) = \frac{\exp\left(\alpha_{iZ_i^l} + \beta_i \sum_{l'\in\mathcal{N}_i(l)} \tau_{iZ_i^l}^{l'}\right)}{\sum_{q=1}^{Q}\exp\left(\alpha_{iq} + \beta_i \sum_{l'\in\mathcal{N}_i(l)} \tau_{iq}^{l'}\right)}, \tag{B.2.2}$$

these definitions coinciding with the previous ones for $\tau_i^l \in \{0,1\}^Q$.

As mentioned in Celeux et al. (2003), to obtain an approximation of these means, we rely on the self consistency condition (see Section 1.7.5.1), stating that the mean obtained based on the mean field approximation must be equal to the mean used to define this approximation, i.e. for any (i,l), when fixing the other values at their mean τ, the expectation of $\mathbf{Z}_i^l = (Z_{iq}^l)_{1\leq q\leq Q} \in [0,1]^Q$ under the approximation must be equal to τ_i^l. Then at each step t, using the definitions (B.2.1) and (B.2.2), we deduce that the expectation values for the parameter $\theta^{(t-1)}$ satisfy the fixed point equation

$$\tau_{iq}^l \propto \exp\left(\alpha_{iq}^{(t-1)} + \beta_i^{(t-1)} \sum_{l'\in\mathcal{N}_i(l)} \tau_{iq}^{l'}\right) \prod_{j\neq i}\prod_{q'}\left((\pi_{qq'}^{(t-1)l})^{X_{ij}^l}(1-\pi_{qq'}^{(t-1)l})^{1-X_{ij}^l}\right)^{\tau_{jq'}^l}. \tag{B.2.3}$$

In practice, the solution of this fixed-point equation is computed iteratively. We then replace the distributions of the latent variables and of the latent variables given the observations by their respective approximations, given in (3.4.8) and (3.4.10), replacing $\tilde{z}$ by τ.

B.2.2 Estimation of π

The estimation of π in the mean field EM is identical to its estimation in the simulated EM, using the mean field approximation instead of the mean field like approximation of the conditional distributions in (3.4.14) and (3.4.15).

[1] i.e. for any (i,l) the probability of $\mathbf{Z}_i^l$ given that the neighbours are equal to $\mathbf{z}_i^{\mathcal{N}_i(l)} = (\mathbf{z}_i^{l'})_{l'\in\mathcal{N}_i(l)}$ and the probability of X^l given $\mathbf{Z}^l = (\mathbf{z}_1^l, \ldots, \mathbf{z}_{i-1}^l, q, \mathbf{z}_{i+1}^l, \ldots, \mathbf{z}_n^l)$

B.2.3 Estimation of (α, β)

In this context, the approximation of Q_2 under the mean field approximation is as follows

$$
\begin{aligned}
&\tilde{Q}_2(\alpha, \beta | \theta^{(t-1)}) \\
&= \sum_{i=1}^{n} \sum_{l=1}^{L} \mathbb{E}_{\tilde{z}, \theta^{(t-1)}} \left[\log \mathbb{P}_{\psi} \left(Z_i^l \mid \mathbf{Z}_i^{\mathcal{N}_i(l)} = \tau_i^{\mathcal{N}_i(l)} \right) \middle| X^{1:L} \right] \\
&= \sum_{i=1}^{n} \sum_{l=1}^{L} \mathbb{E}_{\tilde{z}, \theta^{(t-1)}} \left[\log \frac{\exp\left(\alpha_{iZ_i^l} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iZ_i^l}^{l'} \right)}{\sum_{q=1}^{Q} \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iq}^{l'} \right)} \middle| X^{1:L} \right] \\
&= \sum_{i=1}^{n} \sum_{l=1}^{L} \left\{ \left(\sum_{q=1}^{Q} \alpha_{iq} \mathbb{P}^{\tau}_{\theta^{(t-1)}} \left(Z_i^l = q \middle| X^l \right) \right) + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \sum_{q=1}^{Q} \tau_{iq}^{l'} \mathbb{P}^{\tau}_{\theta^{(t-1)}} \left(Z_i^l = q \middle| X^l \right) \right. \\
&\qquad \left. - \log \left[\sum_{q=1}^{Q} \exp \left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iq}^{l'} \right) \right] \right\}.
\end{aligned}
$$

The derivative of that quantity with respect to α_{iq} (i.e. the quantity we want to set equal to zero) is

$$
\frac{\partial \tilde{Q}_2(\alpha, \beta | \theta^{(t-1)})}{\partial \alpha_{iq}} = \sum_{l=1}^{L} \mathbb{P}^{\tau}_{\theta^{(t-1)}}(Z_i^l = q \mid X^l) - \sum_{l=1}^{L} \frac{\exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iq}^{l'} \right)}{\sum_{q'=1}^{Q} \exp\left(\alpha_{iq'} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iq'}^{l'} \right)},
$$

and the one with respect to β_i is

$$
\begin{aligned}
\frac{\partial \tilde{Q}_2(\alpha, \beta | \theta^{(t-1)})}{\partial \beta_i} = &\sum_{l=1}^{L} \sum_{l' \in \mathcal{N}_i(l)} \sum_{q=1}^{Q} \tau_{iq}^{l'} \mathbb{P}^{\tau}_{\theta^{(t-1)}} \left(Z_i^l = q \middle| X^l \right) \\
&- \sum_{l=1}^{L} \frac{\sum_{q=1}^{Q} \left(\sum_{l' \in \mathcal{N}_i(l)} \tau_{iq}^{l'} \right) \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iq}^{l'} \right)}{\sum_{q=1}^{Q} \exp\left(\alpha_{iq} + \beta_i \sum_{l' \in \mathcal{N}_i(l)} \tau_{iq}^{l'} \right)},
\end{aligned}
$$

where $\mathbb{P}^{\tau}_{\theta^{(t-1)}}(\cdot \mid X^{1:L})$ is the mean field approximation of $\mathbb{P}_{\theta^{(t-1)}}(\cdot \mid X^{1:L})$ and is defined in the same way as the mean field like approximation (3.4.10), but based on the definitions in (B.2.1) and (B.2.2).