Impact of sampling on structure inference in networks : application to seed exchange networks and to ecology

Timothée Tabouy

NNT : 2019SACLS289

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Établissement d'accueil : AgroParisTech

Laboratoire d'accueil : Mathématiques et informatique appliquées, UMR 518 INRA

Spécialité de doctorat :
Mathématiques appliquées

Timothée TABOUY

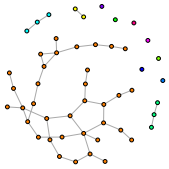
Impact de l'échantillonnage sur l'inférence de structures dans les réseaux : application aux réseaux d'échanges de graines et à l'écologie

Date de soutenance : 30 Septembre 2019

Après avis des rapporteurs : Eric Kolaczyk (Boston University)
Pierre Latouche (Université Paris Descartes)

Jury de soutenance :

M. Pierre BARBILLON	(AgroParisTech) Co-encadrant de thèse
M. Julien CHIQUET	(INRA) Directeur de thèse
Mme Julie JOSSE	(Polytechnique) Présidente du jury
M. Eric KOLACZYK	(Boston University) Rapporteur
M. Pierre LATOUCHE	(Université Paris Descartes) Rapporteur
Mme Tabea REBAFKA	(Sorbonne Université) Examinatrice



Remerciements

Comme tout étudiant lisant une thèse, je me suis souvent arrêté sur les remerciements sans aller plus loin. Parfois court, souvent long, c'est enfin le lieu pour remercier tout ceux qui ont généreusement et passionnément contribué à cette petite fin qu'est la thèse de doctorat. On en a souvent les larmes aux yeux d'ailleurs, car on meurt toujours un peu quand quelque-chose se termine.

J'aimerais ici remercier tout ceux qui de près ou de loin ont contribué à cette thèse et dont la rencontre à fait de moi ce que je suis.

Naturellement je commence par le good cop et le bad cop, alias le codeur fou et le fou codeur, sans qu'un de ces sobriquets ne désigne l'un plutôt que l'autre. Si je dis le petit et le grand, là par contre... Je parle bien entendu de Julien et Pierre. HDR donnant la primeur, mille MERCIS cher directeur sosie du fabuleux Marty McFly, grand maître codeur et pourvoyeur de champagne aux fines bulles. À la fois guide, bienveillant et botteur de fesses tu as tout du parfait directeur de thèse. Évidemment sans toi cette thèse n'aurait jamais eu lieu. D'abord pourvoyeur d'HDR manquante tu m'as par la suite beaucoup appris, sur la forme comme sur le fond. À coder bien sûr, proprement et efficacement. À écrire aussi et surtout à chercher comme tu sais bien le faire. Je n'oublierai pas ce que tu m'as enseigné, conscient de la chance que j'ai je garde ce que tu m'as appris comme un trésor.

Pierre ! La tranquillité en complément de la vivacité. Le brave et le fort, l'ombre et la flamme, chevalier des temps modernes alias Ned Stark. Colosse et virevoltant que tu es, ton jeu physique au foot restera dans les mémoires. Codeur génial à la répartie sanglante, tes "non mais tu rigoles" passeront à la postérité. Un grand MERCI. Pour avoir proposé avec Sophie et Sarah le sujet de stage qui est devenu sujet de thèse ensuite et pour avoir bien voulu co-encadrer ma thèse. Pour ta bienveillance au cours de ces 3 dernières années, le temps passé à discuter et à répondre à mes nombreuses questions. J'ai beaucoup appris en te regardant raisonner et je suis très admiratif de ta compréhension fine des modèles et des données. Souvent j'arrivais avec un problème et je repartais plein d'espoir, à raison, en ce qu'on s'était dit pour le résoudre. Tout comme Julien tu fais un directeur comme beaucoup aimeraient et auraient la chance d'en avoir un. Pour tout ce que je garde de nos échanges et de ce que j'ai appris en travaillant avec vous deux, encore merci.

Merci chaleureusement à Eric Kolaczyk et à Pierre Latouche d'avoir eu la gentillesse d'accepter de rapporter ma thèse et d'avoir été présents le jour de la soutenance. Merci pour vos relectures soigneuses et bienveillantes. J'ai personnellement beaucoup étudié et été très inspiré par vos travaux de recherche respectifs. Merci en particulier à Eric pour les discussions à Eindhoven où j'ai pu profiter en direct de vos talents de chercheur et de votre pédagogie.

Merci à Julie Josse et à Tabea Rebafka d'avoir accepté d'être dans mon jury de thèse. Merci en particulier à Julie de m'avoir invité à parler au séminaire des doctorants du CMAP.

Merci à Sophie et Sarah de m'avoir encadré pendant mon stage de M2, je suis très reconnaissant de toute l'attention et de la gentillesse que vous m'avez montrées pendant ces 3 ans et demi. J'ai beaucoup aimé parler avec chacune et faire de la recherche avec vous.

Merci à Stéphane, pour ta gentillesse et ta bienveillance, pour m'avoir aidé à trouver le stage de M2 qui a débouché sur cette thèse. Tes enseignements et ton travail de recherche ont beaucoup porté et influencé mes travaux. Je suis ébahi par la profondeur de tes questions et remarques, la rapidité et la justesse de tes réflexions. C'est un fait récurrent qu'à chacun de tes exposés, j'ai l'impression de mieux comprendre ce que je fais. Pour tout ça et évidemment un peu plus car tu es Malouin de cœur je te remercie.

Merci à Mahendra d'avoir accepté de collaborer avec moi et de m'avoir accueilli à Jouy régulièrement. Merci de ta disponibilité et de ta grande gentillesse, c'était une vrai chance de travailler avec toi.

Un grand merci au groupe MIRES de m'avoir accueilli au cours de ces 3 dernières années. J'ai été grâce à vous le témoin privilégié de beaux moments de partage et de collaboration scientifique. Un merci particulier à Vanesse, Christian, Mathieu et Matthieu pour vos échanges passionnants et sincères.

Je voudrais ici remercier chaleureusement tout le laboratoire de statistique de l'Agro dans lequel j'ai passé ces 3 dernières années. L'ambiance de travail, vos qualités de chercheurs et vos amitiés ont réellement contribué à me faire grandir à bien des niveaux. Soyez tous assurés de ma gratitude et de mon amitié. Merci Émilie, Maud, Liliane, Christophe, Christophe, Christophe, Marie-Laure, Julie, Éric, Céline, Tristan, Michel, Laure, Jessica, Isabelle, Séverine. Une mention spéciale pour le grand chef Gabriel sans qui les conseils de labo seraient moroses, là où tu passes la folie et les rires suivent ! Un merci particulier à mes co-bureaux successifs et amis. Je pense en premier à Pierre, Marie et Anna qui m'ont accueilli en début de thèse, votre gentillesse et vos bons conseils sans oublier ce tour du monde en bateau virtuel resteront de super souvenirs. Merci Marie, Mathieu, James et Raphaëlle pour avoir partagé avec enthousiasme le BDDSS à tour de rôle. Marie, ta joie permanente et ta gentillesse me resteront à jamais, merci de ton amitié. Mathieu, pour ton amitié et toutes nos discussions je te remercie chaleureusement, j'espère que nos routes se recroiseront à l'avenir :) Et le dernier et non le moindre, merci mon cher Félix pour ta gentillesse et ton amitié sans faille, je te dois beaucoup dans mon admission en médecine et te suis à jamais reconnaissant. Merci Paul, pour ta discrétion et ton exemple, c'est toujours une joie de te croiser. Merci enfin à Saint-Clair mon petit frère de thèse, je te souhaite le meilleur pour la suite, à Yann pour ton amitié et ta gentillesse, à Jade, Rhana, Annarosa, Martina, et Joe pour votre grande gentillesse.

Merci à tous ceux que j'ai rencontrés au cours de séminaires et conférences, qui par leurs idées et questions, m'ont aidés à mieux comprendre et à me poser plus de questions sur mes travaux. Je pense en particulier à Vincent, toujours souriant et gentil, dont les travaux m'ont beaucoup apporté, tu es sûrement la personne dont je connais le mieux la thèse :) Sylvain, grand frère de tous à l'Agro, généreux et souriant, tu inspires beaucoup de monde. J'aurais beaucoup aimé continuer et apprendre avec toi. Jean-Benoist l'irréductible, ta passion et vision profonde des statistiques et de l'informatique forcent le respect ! Merci de ton invitation à Compiègne dont je garde un très bon souvenir :) Merci enfin à Valérie pour ton sourire et ta gentillesse.

Je voudrais remercier maintenant avec beaucoup de gratitude et d'émotion tout ceux à qui je dois d'être là aujourd'hui. Je pense à tous mes professeurs de lycée qui ont cru en moi, et à mes professeurs à l'université. Je pense à Mme Carré qui m'a encouragé et a cru en moi alors que je redoublais et surtout à Mme Berton grâce à qui je suis allé étudier à Orsay. Un grand merci à Dominique Hulin et Jean-Christophe Léger, responsable respectivement des L3 et des L1, à l'époque, à Orsay. La gentillesse et l'énergie que vous avez déployées à croire en chacun de vos étudiants sont pour beaucoup dans ma réussite. J'ai une pensée pour tous les professeurs passionnés qui m'ont enseigné durant mon cursus, soyez tous sincèrement remerciés. Je pense en particulier à Aurélien Galateau qui a été pour moi un exemple et une motivation, merci pour votre gentillesse et votre attention. Un grand merci aussi à toute l'équipe des matheux de l'IUT d'Orsay qui m'ont accueilli avec beaucoup de gentillesse pour faire mon monitorat de thèse.

Je voudrais ici remercier tous mes amis et co-étudiants de mathématiques avec qui j'ai buché et souffert pendant mes années d'études et de préparation de l'agrégation. Je pense à Arnaud, Justine, Victor, Nadège, Yvann, Vincent, Ridwann, Antoine, Antoine, Amandine, Dimitri, Éric, Laurent et Ghislain. Pour vos amitiés qui durent et tous les bons moments partagés je vous dis un grand MERCI. Je n'oublie pas la petite Margaux qui ira loin c'est sûr :) Merci de ton amitié et de ton sourire ;) Enfin, merci à toute la promo du M2 MSV,

c'était un réel plaisir de travailler avec vous et de vous connaître, je pense en particulier à Christophe dont j'espère que les aventures Montpelliéraine se passent bien :) Je n'oublie pas Christophe Giraud, merci de votre attention et de ne jamais compter votre temps pour répondre à nos questions et nous guider.

Merci à Matéo pour ton soutient et ton amitié ;)

Merci à ma belle famille au complet, pour votre soutient sans faille. Pour me suivre dans tout ce que je fais de fou et pour ne jamais cesser de chercher à comprendre. Merci pour votre amour et confiance, c'est très important pour moi.

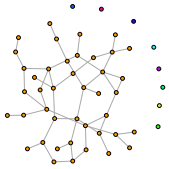
Merci à mes frères et soeurs, à Laure, Emmanuelle, Emmanuel, Tiphaine, Elise et Briac. Pour tout ce qui nous rassemble et tout ce que vous êtes, pour votre amour et parce que je vous aime, MERCI.

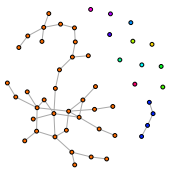
Merci à mes parents, sans qui je ne serai rien. Pour votre exemple en tout, votre soutient et votre amour inconditionnel. Parce que je vous doit tout et bien plus encore, MERCI du fond du coeur.

Merci à celui qui guide nos pas sans jamais nous abandonner et donne sens à toutes choses.

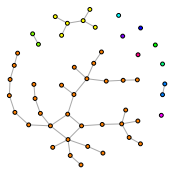
Merci à mon petit Octave, si adorable. T'avoir parmi nous est un vrai bonheur.

Enfin merci à celle qui m'accepte comme je suis, qui met de la lumière chaque jour que Dieu fait et m'aime peu importe le temps qu'il fait en moi. MERCI mon amour, cette thèse t'es dédiée évidemment.



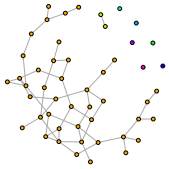


Contents



List of Symbols

- n is the number of nodes in the network,
- $\mathcal{N} = [\![1, n]\!]$ the set of nodes,
- $\mathcal{D} = \mathcal{N} \times \mathcal{N}$ the set of dyads,
- $\mathbf{Y} \in \mathcal{M}_n(\mathbb{R})$ the adjacency matrix of the network,
- $Q \in \mathbb{N}$ the number of blocks (or groups),
- $\mathcal{Q} = [\![1, Q]\!]$ the set of blocks,
- $z \in \mathcal{Q}^n$ the vector of block memberships,
- $\mathbf{Z}_i \in \{0,1\}^Q$ the vector such that $(\mathbf{Z}_i)_q = Z_{iq} = \mathbb{1}_{\{z_i = q\}}$,
- $\mathbf{X}_i \in \mathbb{R}^N$ some covariates of node i such that $(\mathbf{X}_i)_q = X_{iq}$,
- $\mathbf{X}_{ij} \in \mathbb{R}^m$ some covariates of dyad (i, j),
- $\mathbf{X}$ the set of covariates ($\{\mathbf{X}_i\}_i$ on nodes and $\{\mathbf{X}_{ij}\}_{ij}$ on dyads),
- $\mathbf{R} \in \mathcal{M}_n(\mathbb{R})$ the sampling matrix : $R_{ij} = 1$ si Y_{ij} sampled, 0 otherwise,
- $\mathbf{V} \in \{0,1\}^n$ the sampling vector indicating which node is sampled ($V_i = 1$) or not ($V_i = 0$),
- $\mathbf{Y}^{\mathrm{o}} = \{Y_{ij} : R_{ij} = 1\}$ the observed part of the adjacency matrix,
- $\mathbf{Y}^{\mathrm{m}} = \{Y_{ij} : R_{ij} = 0\}$ the non-observed part of the adjacency matrix,
- $\mathcal{D}^o = \{(i, j) : R_{ij} = 1\}$ the observed dyads,
- $\mathcal{D}^m = \{(i, j) : R_{ij} = 0\}$ the non-observed dyads.



Introduction

Contents

1 Motivations : des données manquantes dans les réseaux

Les réseaux sont un moyen naturel de représenter les interactions entre des individus ou plus généralement des entités (protéines, espèces, sites web, etc.). On les rencontre dans de nombreux domaines comme en biologie, en sociologie, en écologie, en industrie, ou Internet. Il existe en statistique de nombreux modèles de réseaux – dont certains seront présentés plus loin – et techniques permettant d'analyser des réseaux (Kolaczyk, 2009 et Matias and Robin, 2014). L'estimation de ces modèles n'est pas en général un problème facile.

Une difficulté supplémentaire souvent rencontrée par le statisticien est de n'avoir accès qu'à un échantillon incomplet du graphe à étudier. En effet, pour des raisons simples telles que le manque de temps, le coût ou la malchance, un observateur peut être conduit à collecter des données incomplètes. On est donc amené à étudier un réseau partiellement observé tout en voulant connaître les propriétés du réseau complet sous-jacent. Dans la suite nous allons motiver notre travail avec deux exemples, le premier concernant des données d'ethnobiologie avec un réseau d'échanges de semences et le second des données génomiques avec un réseau de co-régulation de gènes (ou réseau protéine-protéine).

Le premier exemple que nous étudions est celui d'un réseau d'échanges de semences entre agriculteurs de la région du Mont Kenya. Nous remercions chaleureusement Vanesse Labeyrie d'avoir partagé ces données (voir Labeyrie et al., 2014, 2016) et pris du temps pour en discuter. L'étude de ce réseau a pour but de mieux comprendre l'impact des échanges sur la diversité génétique cultivée. Des études cherchant à lier la diversité génétique cultivée et la diversité culturelle humaine ont montré que les échanges de semences sont intimement liés à l'organisation sociale entre les agriculteurs. C'est un exemple de résultat auquel nous souhaiterions apporter une confirmation (ou non) d'ordre statistique. Ces données sont échantillonnées suivant un processus nœud-centré au sens où certains agriculteurs ont été interrogés pour renseigner à qui ils donnaient et de qui ils recevaient des semences. Ainsi, 155 agriculteurs ont été interrogés, renseignant leurs liens avec un total de 777 agriculteurs (dont les 155 interrogés). Puis, en accord avec les ethnologues, pour simplifier l'étude, nous avons réduit le réseau à 568 des 777 agriculteurs. Pour les 413 non interrogés du réseau

réduit, nous ne sommes pas en mesure de dire quels ont été les échanges ; seuls leurs liens avec les interrogés sont connus.

Le second exemple est celui d'un réseau de co-régulation de gènes. Le réseau extrait de la plateforme `string` (Szklarczyk et al., 2015), accessible à l'adresse http://www.string-db.org, est le réseau de voisinage de la protéine ER (pour Estrogen Receptor) codé par le gène ESR1 (pour EStrogen Receptor 1). Ce réseau est composé de 741 protéines (ou gènes codant ces protéines) et chaque dyade (*i.e.* une paire de nœud) est pondérée par un score appartenant à $[0, 1]$ reflétant le niveau de confiance calculé sur la base d'expériences ou de travaux de recherche. Plus la valeur est proche de 1 plus la probabilité que cette arête existe est grande. Notant ω_{ij} le poids associé à la dyade (i, j), nous définissons la matrice dépendant du seuil γ :

$$\mathbf{A}^\gamma = (A^\gamma)_{i,j} = \begin{cases} 1 & \text{if } \omega_{ij} > 1-\gamma, \\ \texttt{NA} & \text{if } \gamma \leq \omega_{ij} \leq 1-\gamma, \\ 0 & \text{if } \omega_{ij} < \gamma. \end{cases} \tag{1.1}$$

Cette façon d'échantillonner la matrice des poids associés aux arêtes est cette fois-ci centrée sur les dyades.

Remarque 1. *Dans la suite on notera les entrées manquantes dans les dyades par* `NA`.

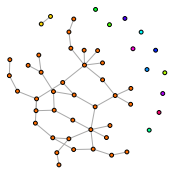
2 État de l'art

Dans cette section, nous allons présenter des modèles de graphes aléatoires, leurs spécificités et quelques résultats théoriques justifiant l'emploi de certains plutôt que d'autres. Ensuite, nous présenterons quelques méthodes d'inférence de ces modèles et leurs limites. Enfin nous introduirons la problématique des données manquantes et l'échantillonnage des réseaux.

2.1 Modèles probabilistes de graphes aléatoires

Nous présentons ici une revue de modèles probabilistes de graphes aléatoires. Les modèles génératifs de graphes comme les modèles petit monde (E J Newman et al., 2002) et d'attachement préférentiel (Barabási and Albert, 1999) ne seront pas présentés ici. Nous commençons par le premier modèle (historiquement parlant) défini par Erdős et Rényi. Puis nous continuons avec les modèles de graphes aléatoires géométriques et exponentiels. Enfin nous aborderons la grande classe des modèles de graphes aléatoires à espaces d'états latents, utilisés pour modéliser l'hétérogénéité des connexions observée dans un réseau. Rappelons qu'en mathématiques, un graphe $\mathbb{G}$ est la donnée du couple $(\{\text{Nœuds}\}, \{\text{Arêtes}\})$. L'ensemble des arêtes est constitué des dyades dont la valeur associée est non nulle. On représente $\mathbb{G}$ par une matrice d'adjacence $\mathbf{Y} = (Y_{ij})_{1\leq i,j\leq n} \in \mathcal{M}_n(\mathbb{R})$ où n est le nombre de nœuds et $Y_{ij} = 1$ s'il y a une arête entre les nœuds i et j, $Y_{ij} = 0$ sinon. Cette matrice peut contenir d'autres valeurs que 0 ou 1. C'est le cas des graphes pondérés qui ne sont pas le sujet principal de cette thèse. D'éventuelles covariables associées aux nœuds seront désignées par la variable $\mathbf{X} = (\mathbf{X}_i)_{1\leq i\leq n}$. Si en revanche elles sont associées aux dyades, on les notera $\mathbf{X} = (X_{ij})_{1\leq i,j\leq n}$. Elles peuvent être catégorielles, quantitatives ou bien un mélange des deux. Rappelons que $\mathcal{N} := [\![1, n]\!]$ désigne l'ensemble des nœuds du graphe et $\mathcal{D} := \mathcal{N} \times \mathcal{N}$ l'ensemble des dyades. Finalement nous noterons $\boldsymbol{\theta}$ et Θ respectivement le paramètre associé à un modèle et l'espace des paramètres.

Par la suite nous nous intéresserons plus particulièrement au cas de graphes binaires (sauf mention du contraire) non-orientés (*i.e.* dont la matrice d'adjacence $\mathbf{Y}$ est symétrique) et sans boucles (*i.e.* la diagonale de $\mathbf{Y}$ est nulle). Ce choix est fait pour plus de simplicité. Cependant tout ce qui suit s'étend naturellement aux cas de graphes pondérés et/ou non-symétriques et/ou avec des boucles.



Modèle d'Erdős-Rényi

Soit $p \in [0,1]$ une probabilité. Ce modèle couramment noté $\mathcal{G}(n,p)$ est défini comme suit :

$$(Y_{ij})_{i,j} \sim^{iid} \mathcal{B}(p)$$

où $\mathcal{B}$ est la loi de Bernoulli. On note que la loi des degrés est binomiale : $D_i := \sum_j Y_{ij} \sim \mathrm{Bin}(n-1,p)$. Citons l'article Gilbert (1959) et le livre plus récent Bollobas (2001) dans lequel ce modèle a été largement étudié.

Modèles de graphes aléatoires géométriques

Les modèles de graphes aléatoires géométriques sont généralement utilisés pour modéliser des réseaux de télécommunications sans fil ou bien des propagations de phénomènes naturels tels que des incendies de forêt ou encore des épidémies. Nous citons le livre de Penrose (2003) qui est une référence incontournable concernant cette classe de modèles. Dans ces modèles, les nœuds sont identifiés à des positions spatiales aléatoires $(N_i)_{1\leq i\leq n}$ dans un espace métrique (E,d) suivant une certaine distribution (uniforme ou gaussienne par exemple). De plus, conditionnellement à la position des nœuds et en se fixant un seuil $r > 0$, la matrice d'adjacence $\mathbf{Y}$ est déterministe. Plus précisément on a

$$(Y_{ij})_{i,j} = \mathbb{1}_{d(N_i,N_j)\leq r}.$$

À noter le lien entre ces modèles et le précédent : on peut montrer que localement, les graphes géométriques contiennent un graphe d'Erdős-Rényi (voir Channarond, 2013).

Modèle Expected Degree (EDD)

Adapté des modèles définis dans Chung and Lu (2002) et Newman (2003), le modèle EDD permet de contrôler en espérance les degrés de chaque nœud. Dans la suite on notera $(K_i)_i$ les « degrés attendus » (non nécessairement entiers) associés à chaque nœud et G une densité de probabilité. Alors,

$$(K_i)_i \sim^{iid} G$$
$$(Y_{ij})_{i,j} \mid (K_i),(K_j) \sim^{ind} \mathcal{B}(K_iK_j/\kappa),\ \kappa \in \mathbb{R}.$$

De cette façon, $\mathbb{E}[D_i|K_i] \propto K_i$, $\forall i \in [\![1,n]\!]$.

Modèle de graphes aléatoires exponentiels (ERGM)

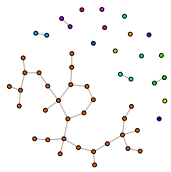
La famille exponentielle permet à elle seule de définir un grand nombre de lois classiques telles que les lois de Bernoulli, binomiale, Poisson et gaussienne par exemple. L'idée des ERGM (Exponential Random Graph Models) est d'utiliser la caractère générique de l'écriture de la famille exponentielle pour modéliser un graphe. Ce modèle exprime la loi d'un graphe en fonction de statistiques exhaustives, telles que le nombre d'arêtes dans le graphe, le nombre de triangles pour décrire tout type de dépendance entre les arêtes. Leur loi est de la forme

$$\mathbb{P}(\mathbf{Y}=\mathbf{y}) = \frac{1}{\Lambda_\theta}\exp\left(\sum_{C\in\mathcal{D}}\theta_C g_C(y)\right)$$

où $\mathbf{y}$ est une réalisation de $\mathbf{Y}$ et

(i) chaque C est appelé une configuration, *i.e.* un sous-ensemble d'arêtes ou de dyades parmi un sous-ensemble de nœuds du graphe ;

(ii) $g_C(y) = \prod_{(i,j)\in C} y_{ij}$ indique si la configuration C apparaît dans Y ;



(iii) les paramètres θ_C sont associés à la configuration C. Si $\theta_C = 0$ alors les arêtes de la configuration sont indépendantes conditionnellement aux arêtes restantes du graphe. Une valeur non nulle de θ_C induit de la dépendance entre les arêtes de C conditionnellement aux restes des arêtes. Enfin, quand une valeur positive de θ_C encourage la configuration, une valeur négative la pénalise ;

(iv) Λ_θ est une constante de normalisation dépendant de $\boldsymbol{\theta}$,

$$\Lambda_\theta = \sum_y \exp\left(\sum_{C\in\mathcal{D}} \theta_C g_C(y)\right). \tag{1.2}$$

On remarquera que toute collection de paramètres $\{\theta_C, C \in \mathcal{D}\}$ ne conduit pas automatiquement à un modèle où la distribution jointe de $\mathbf{Y}$ est bien définie. Les conditions de validité sont décrites dans le théorème de Hammersley-Clifford (voir Besag, 1974). On notera certaines limites théoriques et pratiques importantes évoquées dans Chatterjee and Diaconis (2013) à propos des ERGM (ce travail concerne exclusivement le cas des graphes denses et s'inscrit dans le cadre de la théorie limite des graphes développée dans Lovász, 2012) comme la dégénérescence de modèles (beaucoup de réalisations sont en fait soit des graphes vides d'arêtes soit complets), ou l'inutilité des statistiques exhaustives prises en comptes dans le modèle (Bhamidi et al., 2011), ou encore le fait que beaucoup de réalisations des ERGM ne se distinguent pas d'un graphe simulé suivant le modèle d'Erdős-Rényi.

Modèles de graphes aléatoires à espaces d'états latents

Dans cette partie nous allons présenter en détail la classe des modèles à espaces d'états latents. Ces modèles ont pour principal objectif la modélisation de l'hétérogénéité des profils des individus dans un graphe. Ils permettent aussi de faire du clustering en regroupant les nœuds dans des groupes homogènes d'un point de vue de leurs connectivité. Nous noterons $\mathbf{Z} := \{\mathbf{Z}_i, i \in \mathcal{N}\}$ les variables latentes associées aux nœuds. Dans la suite deux cas se dissocieront : quand l'espace latent est fini et quand il est continu. Quand il est fini, nous noterons Q le nombre d'états latents (ou groupes) et $\mathcal{Q} := [\![1, Q]\!]$. Les variables latentes prennent alors la forme de Q-uplets composés de 1 et de 0. La valeur 1 à la coordonnée k signifie que le nœud est dans l'état latent k. Un nœud peut appartenir à 0, 1 ou plusieurs états latents suivant les modèles. Quand l'espace latent est continu, nous considérerons que les variables latentes appartiennent à $\mathbb{R}^d$.

Stochastic Block Model (SBM). Dans le SBM, les variables latentes associées aux nœuds sont catégorielles. Plus précisément elles codent l'information du groupe auquel appartient chaque nœud. On décrit la structure de dépendance entre $\mathbf{Z}$ et $\mathbf{Y}$ dans le cas où $n = 3$ dans la Figure 1.1

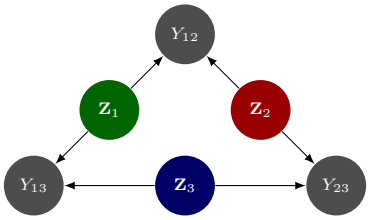
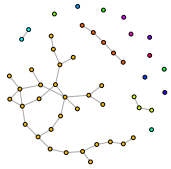


FIGURE 1.1 – Modèle graphique réduit associé au SBM.

D'après la Figure 1.1, on a $\mathbb{P}(\mathbf{Y}\,|\mathbf{Z}) = \otimes_{(i,j)\in\mathcal{D}} \mathbb{P}(Y_{ij}\,|\mathbf{Z}_i, \mathbf{Z}_j)$. Ainsi on définit le Stochastic



Block Model dans le cas général comme suit :

$$(\mathbf{Z}_i)_i \sim^{iid} \mathcal{M}(1, \boldsymbol{\alpha})$$
$$(Y_{ij})_{i,j} \mid (Z_{iq} = 1), (Z_{j\ell} = 1) \sim^{ind} f(\cdot; \delta_{q\ell})$$

où $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_Q)$ est le paramètre de mélange et $\boldsymbol{\delta} = (\delta_{q\ell})_{(q,\ell)\in\mathcal{Q}^2}$ la matrice des paramètres de connexion intra et inter-groupes. Cette matrice est aussi appelée matrice de connectivité dans le cas de données binaires. Le cas du SBM binaire (ou Bernoulli) a été étudié dans Snijders and Nowicki (1997) et Daudin et al. (2008). Il correspond au cas où la loi de Y_{ij} sachant $\mathbf{Z}_i, \mathbf{Z}_j$ est la loi de Bernoulli (voir Table 1.1) et

$$\forall y \in \{0, 1\} \text{ et } \forall \pi \in [0, 1], \quad f(y; \pi) = \pi^y (1 - \pi)^{1-y}.$$

C'est ce cas qui concentrera l'essentiel de notre attention dans la thèse.

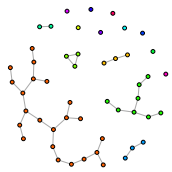
Le cas des SBM pondérés a été largement développé dans Mariadassou et al. (2010). On y trouve par exemple les cas détaillés (modèle et estimation) des lois d'émissions Poisson, régression poissonnienne avec effets homogène (PRMH) ou inhomogène (PRMI), multinomiale, gaussienne (univariée, bivariée ou encore la régression linaire). Nous rappelons quelques lois classiques et les paramètres associés dans la Table 1.1. Nous citerons aussi les modèles Multiplex SBM (Barbillon et al., 2015) et Multipartite SBM (Bar-Hen et al., 2018) dans lesquels des liens de différentes natures entre des individus sont modélisés ; le Stochastic Topic Block Model (Bouveyron et al., 2016 et Corneli et al., 2018) dans lequel les arêtes représentent un échange de texte et enfin les modèles de SBM dynamiques (Matias and Miele, 2017, Matias et al., 2018 et Ludkin et al., 2018) tous directement inspirés du SBM.

Modèle	**Loi de probabilité**	**Paramètre** $\delta_{q\ell}$
Bernoulli	$\mathcal{B}(\pi_{q\ell})$	$\pi_{q\ell}$
Bernoulli avec covariables	$\mathcal{B}(g(\gamma_{q\ell} + X_{ij}^T\boldsymbol{\beta}))$	$(\gamma_{q\ell}, \boldsymbol{\beta})$
Poisson	$\mathcal{P}(\lambda_{q\ell})$	$\lambda_{q\ell}$
PRMH	$\mathcal{P}(\lambda_{q\ell} e^{\boldsymbol{\beta}^T X_{ij}})$	$(\lambda_{q\ell}, \boldsymbol{\beta})$
PRMI	$\mathcal{P}(\lambda_{q\ell} e^{\boldsymbol{\beta}_{q\ell}^T X_{ij}})$	$(\lambda_{q\ell}, \boldsymbol{\beta}_{q\ell})$
Multinomial	$\mathcal{M}(1; \mathbf{p}_{q\ell})$	$\mathbf{p}_{q\ell}$
Gaussien univarié	$\mathcal{N}(\mu_{q\ell}, \sigma_{q\ell}^2)$	$(\mu_{q\ell}, \sigma_{q\ell}^2)$
Gaussien bivarié	$\mathcal{N}(\boldsymbol{\mu}_{q\ell}, \boldsymbol{\Sigma}_{q\ell})$	$(\boldsymbol{\mu}_{q\ell}, \boldsymbol{\Sigma}_{q\ell})$
Régression linéaire simple	$\mathcal{N}(a_{q\ell} + bX_{ij}, \sigma^2)$	$(a_{q\ell}, b, \sigma^2)$
Régression linéaire	$\mathcal{N}(\boldsymbol{\beta}_{q\ell}^T X_{ij}, \sigma_{q\ell}^2)$	$(\boldsymbol{\beta}_{q\ell}, \sigma_{q\ell}^2)$

TABLE 1.1 – Modèles de SBM et lois de probabilités de : $Y_{ij}|Z_{iq} = 1, Z_{j\ell} = 1$, sauf pour le modèle gaussien bivarié dont la loi est celle de : $(Y_{ij}, Y_{ji})|Z_{iq} = 1, Z_{j\ell} = 1$.

Le SBM est identifiable (voir Celisse et al., 2012) sauf sur un ensemble de mesure de Lebesgue nulle et à permutation des labels près.

Degree Corrected Stochastic Block Model (DCSBM). Une limite du SBM tel que défini précédemment concerne le fait que tous les nœuds d'un même groupe sont stochastiquement équivalents. En particulier $\mathbb{E}[Y_{ij}|Z_{iq} = 1, Z_{j\ell} = 1] = \pi_{q\ell}$ ne dépend que des classes des nœuds i et j et en espérance 2 nœuds d'un même groupe ont les mêmes degrés. Ainsi, dans le but de prendre en compte la propension de chaque nœud à se connecter aux autres et introduire plus d'hétérogénéité dans les degrés des nœuds du graphe, les auteurs de Karrer and Newman (2011) ont proposé le DCSBM dans lequel chaque nœud i a un paramètre γ_i qui lui est propre et influe sur son degré de sorte que $\mathbb{E}[Y_{ij}|Z_{iq} = 1, Z_{j\ell} = 1] = \gamma_i\gamma_j\pi_{q\ell}$. Ces paramètres de degré sont soumis à la contrainte $\sum_i \gamma_i \mathbb{1}_{\{Z_{iq}=1\}} = 1$, $\forall q \in \mathcal{Q}$ pour que le modèle soit identifiable.



Ce modèle a été étudié dans le cas d'une loi d'émission de Bernoulli ou Poisson. La loi de Poisson permet une mise à jour explicite des paramètres du modèle *a contrario* de la loi de Bernoulli. La loi de Poisson est donc utilisée pour approcher la loi de Bernoulli dans le cas d'une faible espérance correspondant à des données parcimonieuses. Une étude théorique de ces modèles est menée dans Zhao et al. (2012b).

Popularity-Adjusted Block Model (PABM). La « popularité » des nœuds est fortement liée à la notion de structure en communauté (ou affiliation). La popularité observée d'un nœud i au sein de la communauté (ou groupe) q est donnée par $\mathrm{P}_{iq} = \sum_{j,\ \mathbf{z}_j=q} Y_{ij}$ et son espérance $\mu_{iq} = \mathbb{E}[\mathrm{P}_{iq}]$. On remarque que, dans le cas du SBM, $\mu_{iq} = \#\{j,\ Z_{jq} = 1\}\pi_{c_i q}$ et que pour le DCSBM, $\mu_{iq} = \gamma_i \pi_{c_i q}$ où $c_i = \sum_{q=1}^{Q} q \mathbb{1}_{\{Z_{iq}=1\}}$. Donc, dans le SBM, les nœuds d'une même communauté ont tous la même popularité et dans le DCSBM, la popularité des nœuds est modulée par leurs degrés. Ainsi, deux nœuds de forts degrés ont plus de chance de se connecter entre eux que deux nœuds de faibles degrés. Ce point de modélisation n'est pas très réaliste si par exemple les communautés représentent des clans opposés. En effet, dans ce cas les nœuds de forts degrés représentent les « chefs » opposés et il est peu probable qu'ils se connectent entre eux. Au contraire, les nœuds de faible degré représentant des gens dont la position sociale est moins extrême se connecteront plus volontiers. Le PABM (Sengupta and Chen, 2018 et Noroozi et al., 2019) prend en compte cet aspect de modélisation dans sa définition.
Soit $(\lambda_{iq})_{i\in\mathcal{N}, q\in\mathcal{Q}}$ les paramètres de popularité associés à chaque nœud et vis-à-vis de chaque classe. Alors,

$$(\mathbf{Z}_i)_i \sim^{iid} \mathcal{M}(1, \boldsymbol{\alpha})$$
$$(Y_{ij})_{i,j} \mid (Z_{iq} = 1), (Z_{j\ell} = 1) \sim^{ind} \mathcal{B}(\lambda_{i\ell}\lambda_{jq}).$$

Le modèle est identifiable sous la contrainte que $\Lambda_{q\ell} = \Lambda_{\ell q}$ où $\Lambda_{q\ell} := \sum_{j,\ \mathbf{z}_j=q} \lambda_{j\ell}$. De plus, sous le PABM, $\mu_{iq} = \lambda_{iq}\Lambda_{qc_i}$.

Overlapping Stochastic Block Model (OSBM). Une hypothèse du SBM est que chaque nœud appartient à un et un seul groupe. L'overlapping SBM (Latouche et al., 2011) est une adaptation du SBM qui permet à chaque nœud d'appartenir à plusieurs groupes. Plus précisément, la loi de $\mathbf{Z}$ est donnée par

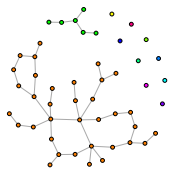
$$(\mathbf{Z}_i)_i \sim^{iid} \otimes_{q=1}^{Q} \mathcal{B}(\alpha_q).$$

On remarque que cette définition autorise toutes les composantes de $\mathbf{Z}_i$ à être à zéro, ce qui permet de ne pas avoir de classe composée de nœuds de faible degré comme on peut souvent observer dans les problèmes de classification. En effet, ces « outliers » ne sont pas classés par OSBM, *i.e.* $\mathbf{Z}_i = 0$. On peut maintenant donner la loi conditionnelle de Y_{ij} sachant $\mathbf{Z}_i$ et $\mathbf{Z}_j$:

$$(Y_{ij})_{i,j} \ \mid \ (Z_{iq} = 1), (Z_{j\ell} = 1) \sim^{ind} \mathcal{B}(g(a_{q\ell})) \tag{1.3}$$

où $g(x) = (1 + e^x)^{-1}$. De plus $a_{\mathbf{Z}_i\mathbf{Z}_j} = \mathbf{Z}_i^T W \mathbf{Z}_j + \mathbf{Z}_i^T U + \mathbf{Z}_j^T V + W^*$ avec $W \in \mathcal{M}_Q(\mathbb{R})$ et $U, V \in \mathcal{M}_{Q1}(\mathbb{R})$. Le premier terme de (1.3) décrit les interactions entre les nœuds i et j alors que les second et troisième termes modélisent respectivement les capacités des nœuds i et j à se connecter aux autres nœuds. Le dernier terme est scalaire et représente un biais pour modéliser la parcimonie. Enfin, nous noterons que l'OSBM est identifiable (Latouche et al., 2011), à échange des numéros des groupes près et sauf sur un ensemble de mesure de Lebesgue nulle.

Mixed Membership Stochastic Block Model (MMSBM). Ce modèle est défini pour la première fois dans Airoldi et al. (2008) avec pour but de prendre en compte dans chaque relation possible la variabilité propre à chaque nœud de se connecter aux autres. Dans le



modèle OSBM, chaque nœud appartient à plusieurs classes et la probabilité que deux nœuds se connectent dépend de toutes les classes auxquels appartiennent les deux nœuds. Dans MMSBM, pour chaque dyade, la probabilité que les deux nœuds en question se connectent ne dépend que d'un seul des groupes auxquels appartient chaque nœud. Plus précisément,

$$\begin{aligned}(\boldsymbol{\gamma}_i)_i &\sim^{iid} \text{Dirichlet}(\boldsymbol{\alpha})\\ (\mathbf{Z}_{i\to j})_{i,j} &\sim^{ind} \mathcal{M}(1,\boldsymbol{\gamma}_i)\\ (\mathbf{Z}_{j\to i})_{i,j} &\sim^{ind} \mathcal{M}(1,\boldsymbol{\gamma}_j)\\ (Y_{ij})_{i,j} \mid (\mathbf{Z}_i),(\mathbf{Z}_j) &\sim^{ind} \mathcal{B}(\mathbf{Z}_{i\to j}^T\boldsymbol{\pi}\mathbf{Z}_{j\to i}).\end{aligned}$$

Le paramètre de mélange est donc $\boldsymbol{\alpha} \in \mathcal{Q}$, et la matrice $Q \times Q$ de connectivité est $\boldsymbol{\pi}$, comme dans le SBM. Notons que l'identifiabilité de ce modèle n'a pas encore été démontrée.

Latent Block Model (LBM). Le LBM est un modèle fréquemment utilisé pour faire du clustering de ligne et colonne de matrice *a priori* non carrée (ou bi-clustering), pas nécessairement une matrice d'adjacence au sens classique. Il a été introduit par Govaert and Nadif (2003). Nous noterons $\mathbf{M} \in \mathcal{M}_{nd}(\mathbb{R})$ une telle matrice. Ce modèle est très lié au SBM qui correspond au cas où les groupes en colonnes sont identiques aux groupes en lignes. Il permet en particulier de modéliser un graphe bipartite. On définit $\mathbf{Z}$ et $\mathbf{W}$ les variables latentes respectivement associées aux lignes et aux colonnes, leur loi étant donnée par

$$\begin{aligned}(\mathbf{Z}_i)_i &\sim^{iid} \mathcal{M}(1,\boldsymbol{\alpha})\\ (\mathbf{W}_j)_j &\sim^{iid} \mathcal{M}(1,\boldsymbol{\rho}).\end{aligned}$$

On a donc

$$(M_{ij})_{i,j} \mid (\mathbf{Z}_i),(\mathbf{W}_j) \sim^{ind} f(\cdot;\mathbf{Z}_i^T\boldsymbol{\pi}\mathbf{W}_j).$$

Les paramètres $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{Q_1})$ et $\boldsymbol{\rho} = (\rho_1, ..., \rho_{Q_2})$ sont les paramètres de mélange associés respectivement aux lignes et aux colonnes et $\boldsymbol{\pi} \in \mathcal{M}_{Q_1Q_2}(\mathbb{R})$ la matrice des paramètres de connexions inter et intra-groupes. Comme dans le cas du SBM la densité f peut par exemple appartenir à la famille exponentielle ou être une densité quelconque.

Le LBM est identifiable sous certaines conditions décrites dans Keribin et al. (2015) et à permutation près des labels en lignes et en colonnes.

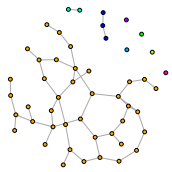
Latent Position Cluster (LPCM). Introduit par Handcock et al. (2007) et inspiré du SBM, le LPCM a pour but de modéliser un réseau basé sur les distances sociales entre individus. Ceci implique que deux personnes ayant un profil social proche auront une plus grande probabilité de se connecter. Cette tendance à s'affilier à ses semblables s'appelle l'homophilie. On associe (comme dans le SBM) à chaque nœud du graphe une variable latente continue à valeurs dans $\mathbb{R}^d$ correspondant aux caractéristiques sociales des nœuds :

$$(\mathbf{Z}_i)_i \sim^{iid} \sum_{q\in\mathcal{Q}} \alpha_q \mathcal{N}(\mu_q, \sigma_q^2 I_d). \tag{1.4}$$

Les nœuds sont donc d'abord placés dans des groupes à la façon du SBM avec probabilités $\boldsymbol{\alpha}$, puis leurs caractéristiques sociales – qui dépendent du groupe auxquels ils appartiennent – suivent une certaine loi normale multi-dimensionnelle. Les variables latentes sont ainsi indépendantes conditionnellement aux groupes auxquels chaque nœud appartient (correspondant à la première étape du processus). Finalement,

$$(Y_{ij})_{i,j} \mid (\mathbf{Z}_i),(\mathbf{Z}_j) \sim^{ind} \mathcal{B}(g(\beta_0 c_{ij} - \beta_1\|\mathbf{Z}_i - \mathbf{Z}_j\|)) \tag{1.5}$$

avec $g = (1+e^x)^{-1}$ et $\beta_1 \geq 0$ pour que la probabilité soit faible quand la distance sociale est grande entre i et j et élevée quand ils sont proches socialement. Les $(c_{ij})_{(i,j)\in\mathcal{D}}$ jouent le rôle de covariables observées sur les arêtes.



Latent Positions Model (LPM). Très proche dans sa définition du LPCM, le LPM (voir Channarond, 2013, pour une définition et une étude théorique du modèle) modélise aussi la notion d'homophilie. Dans ce modèle, les variables latentes sont continues et appartiennent à l'espace $\mathbb{R}^d$. Soit f une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, on considère le modèle suivant :

$$(\mathbf{Z}_i)_i \sim^{iid} f$$
$$(Y_{ij})_{i,j} \mid (\mathbf{Z}_i), (\mathbf{Z}_j) \sim^{ind} \mathcal{B}\left(k\left(\frac{\mathbf{Z}_i - \mathbf{Z}_j}{h_n}\right)\right)$$

avec $h_n > 0$ et $k : \mathbb{R} \mapsto [0,1]$ une fonction isotrope (*i.e.* ne dépendant pas de la direction) et décroissante par rapport à la norme euclidienne.

Graphon (ou W-graph). Ce modèle de graphe introduit par Lovász and Szegedy (2006) est très populaire car très général dans sa formulation. De plus il peut être vu comme une limite pour les graphes denses. Il est très activement étudié. On cite par exemple le livre de Lovász (2012) qui est une référence incontournable concernant ce modèle.

Un graphon est une fonction symétrique et mesurable $f : (I^2, \mathcal{B}(I^2)) \mapsto (I, \mathcal{B}(I))$ avec $I = [0,1]$. Comme dans les modèles LPCM et LPM, les variables latentes sont continues :

$$(\mathbf{Z}_i)_i \sim^{iid} \mathcal{U}(I)$$
$$(Y_{ij})_{i,j} \mid (\mathbf{Z}_i), (\mathbf{Z}_j) \sim^{ind} f(\mathbf{Z}_i, \mathbf{Z}_j). \tag{1.6}$$

Le modèle défini ainsi n'est pas identifiable. En effet, pour toute mesure σ préservant les fonctions définies et à image dans $[0,1]$, composer la fonction f avec σ donne exactement le même modèle qu'avec la fonction f. Pour pallier ce problème, on peut par exemple supposer que $x \mapsto \int f(x,y)\mathrm{d}y$ est croissante (Latouche et al., 2018).
Enfin, pour relier le SBM au graphon, on remarquera que l'équation (1.6) permet de définir le SBM comme un graphon. En effet, si la fonction f est constante sur chaque bloc rectangulaire de taille $\alpha_q \times \alpha_\ell$ et vaut $\pi_{q\ell}$, alors ceci correspond au SBM de paramètres $\boldsymbol{\alpha}$ et $\boldsymbol{\pi}$ tel que défini précédemment. La variable latente $\mathbf{Z}_i$ associée au nœud i est alors simplement le numéro du sous-intervalle dans lequel tombe la variable aléatoire $\mathbf{Z}_i$.

2.2 Garanties théoriques : théorème de représentation et loi limite

Théorème de représentation par un modèle de Graphon

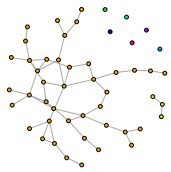
Le résultat suivant justifie l'étude des modèles à variables latentes et particulièrement du SBM. Il établit le modèle de graphon – qui est une version générale et continue d'un modèle à variable latente – comme une représentation possible de tous ces modèles.

Théorème 1. *(Lovász and Szegedy, 2006, Théorème 2.7) Pour toute suite de modèles $(\mathbb{G}_n)$ de graphes aléatoires de taille n vérifiant les conditions suivantes, il existe une fonction symétrique $\kappa : [0,1]^2 \mapsto [0,1]$ telle que le modèle a la même loi que $\mathcal{H}_{n,\kappa}$:*

(i) La loi de $\mathbb{G}_n$ est invariante par permutation des numéros des nœuds (on dit que c'est un modèle de graphe échangeable).

(ii) La loi du sous-graphe induit pas l'ensemble de nœuds allant de 1 à $n-1$ est la même que la loi de $\mathbb{G}_{n-1}$

(iii) Pour tout $1 < k < n$, les sous-graphes de $\mathbb{G}_n$ induits par les ensembles de nœuds $[\![1,i]\!]$ et $[\![i+1,n]\!]$ sont indépendants.



Limite de graphes aléatoires

Le chapitre 11 du livre Lovász (2012) est consacré à l'étude de la convergence des suites de graphes aléatoires denses et à l'identification de la loi (ou graphe) limite. Dans le cadre théorique qui y est développé il est montré que le graphon est le bon modèle limite pour des suites de graphes aléatoires tels que ceux satisfaisant les conditions du Théorème 1.

2.3 Inférence des modèles de graphe

Il existe globalement 2 grandes familles de méthode pour estimer les modèles décrits à la section 2.1 : (i) les méthodes particulaires de type `MCMC` (`Markov Chain Monte Carlo`) et (ii) les méthodes dérivées de l'algorithme `EM` (`Expectation - Maximization`, Dempster et al., 1977). Concernant les méthodes `MCMC`, les algorithmes de Metropolis-Hastings ou l'échantillonneur de Gibbs (voir Gilks et al., 1995) sont les plus couramment utilisés. Leur but est d'estimer une loi *a posteriori*. Concernant les dérivées de l'`EM`, il existe de nombreux algorithmes tels que le `CEM` (Celeux and Govaert, 1991), le `VEM` (Jordan et al., 1998 et Jaakkola, 2000), ou encore le `VBEM` (Latouche et al., 2012 et Brault, 2014), dont l'objectif est d'estimer le maximum de vraisemblance. Ces deux familles de méthodes sont parfois couplées, comme dans les algorithmes `SEM` (Celeux and Diebolt, 1985), `MCEM` (Wei and Tanner, 1990), `SAEM` (Delyon et al., 1999) ou encore `SEM-Gibbs` et `SEM+VEM` (voir Brault, 2014).

Certains algorithmes de clustering comme les k-means (Steinhaus (1956) et MacQueen (1967)) et le Spectral Clustering (Donath and Hoffman (1973) et Rohe et al. (2010) appliqué au SBM) sont fréquemment utilisés pour initialiser les algorithmes `MCMC` ou les dérivées de l'`EM`. L'algorithme des k-means divise les données en k groupes en minimisant la distance d'un point à la moyenne des points de son groupe. Le spectral clustering quant à lui applique une méthode de clustering comme les k-means aux vecteurs propres de la matrice Laplacienne associée à un graphe.

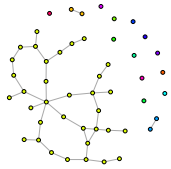
Concernant les modèles de graphes aléatoires à variables latentes, il est impossible de factoriser la loi de $\mathbf{Z}$ conditionnellement à $\mathbf{Y}$ (voir Lauritzen, 1996). Ces modèles ne peuvent donc pas être estimés avec l'algorithme `EM`. Les algorithmes `MCMC` ont d'abord été utilisés pour estimer ces modèles on citera Celeux and Govaert (1991), Nowicki and Snijders (2001) pour le SBM et Govaert and Nadif (2003) pour le LBM. Ils souffrent cependant d'un manque de rapidité computationnelle et ne s'appliquent qu'à des réseaux composés de plusieurs centaines de nœuds. Les algorithmes avec approximation variationnelle (Jordan et al. (1998), Jaakkola (2000)), basés sur l'`EM` et dans lesquels la loi de $\mathbf{Z}|\mathbf{Y}$ est approchée ont ensuite été utilisés : voir Daudin et al. (2008) pour le SBM, Keribin et al. (2015) pour le LBM, Latouche et al. (2012) pour le OSBM, Airoldi et al. (2008) pour le MMSBM, Latouche and Robin (2016) pour le graphon. Ceci est vrai aussi pour toutes les versions du SBM (dynamiques ou pas).

Pour l'estimation de très grands graphes, on citera le Largest Gaps Algorithm (LGA) (Channarond (2013) pour le SBM et Brault (2014) pour le LBM) qui estime les paramètres du modèle ainsi que la classification des nœuds à partir de la densité empirique des degrés des nœuds en évaluant les plus grands « gaps » dans la densité. Cependant cet algorithme s'applique principalement à de gros graphes composés d'au moins plusieurs milliers de nœuds.

L'estimation des modèles ERGM quant à eux se fait essentiellement sur la base d'algorithmes `MCMC`, voir Snijders (2002), Hunter and Handcock (2006a) et Kolaczyk (2009) pour plus de détails. On remarquera que la constante de normalisation Λ_θ dépend de θ, c'est un problème pour utiliser des algorithmes `MCMC` comme le Metropolis-Hastings.

Dans la suite nous présentons en détail l'algorithme `VEM` avec comme fil rouge son application au SBM.

Variational `EM`. Cet algorithme a pour but de maximiser la vraisemblance des données observées. On décompose pour cela la log-vraisemblance des données $\log p_\theta(\mathbf{Y})$ de la façon



suivante :

$$\log p_\theta(\mathbf{Y}) = \log p_\theta(\mathbf{Y}, \mathbf{Z}) - \log p_\theta(\mathbf{Z}|\mathbf{Y}). \tag{1.7}$$

Comme dans le raisonnement amenant à l'algorithme EM décrit dans Dempster et al. (1977), on intègre (1.7) de chaque coté par rapport à une loi $\mathbb{Q}$ portant seulement sur $\mathbf{Z}$. On obtient :

$$\log p_\theta(\mathbf{Y}) = \mathbb{E}_\mathbb{Q}[\log p_\theta(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}(\mathbb{Q}) + \mathcal{KL}[\mathbb{Q}\|p_\theta(\mathbf{Z}|\mathbf{Y})] \tag{1.8}$$

où $\mathcal{H}$ est l'entropie de Shannon (positive, maximale pour la loi uniforme) et $\mathcal{KL}(\mathbb{Q}, \mathbb{P})$ la divergence de Kullback-Leibler (positive et nulle si et seulement si $\mathbb{P} \overset{loi}{=} \mathbb{Q}$).

Si $\mathbb{Q}$ était égale à $p_\theta(\mathbf{Z}|\mathbf{Y})$, alors maximiser (1.8) en $\boldsymbol{\theta}$ reviendrait à utiliser exactement l'algorithme EM. Cependant, à cause de la structure de dépendance existante dans les modèles de graphes aléatoires à variables latentes entre les variables $\mathbf{Y}$ et $\mathbf{Z}$ (Figure 1.1), le calcul de l'étape E de l'algorithme EM, autrement dit le calcul de $\mathbb{E}_{\mathbf{Y}|\mathbf{Z}}[\log p_\theta(\mathbf{Y}, \mathbf{Z})]$, se trouve compromis car la distribution de $\mathbb{P}(\mathbf{Z}|\mathbf{Y})$ ne se factorise d'aucune façon. Finalement, en utilisant la décomposition (1.8) et la positivité de la divergence de Kullback-Leibler, on obtient :

$$\begin{aligned} \log p_\theta(\mathbf{Y}) \geq \mathcal{J}_{\boldsymbol{\theta},\mathbb{Q}}(\mathbf{Y}) \quad &:= \quad \mathbb{E}_\mathbb{Q}[\log p_\theta(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}(\mathbb{Q}) && (1.9) \\ &= \quad \log p_\theta(\mathbf{Y}) - \mathcal{KL}[\mathbb{Q}\|p_\theta(\mathbf{Z}|\mathbf{Y})]. && (1.10) \end{aligned}$$

Améliorer la borne inférieure (1.10) équivaut à minimiser $\mathcal{KL}[\mathbb{Q}\|p_\theta(\mathbf{Z}|\mathbf{Y})]$ en $\mathbb{Q}$ dans l'ensemble des distributions, c'est-à-dire trouver $\mathbb{Q} = p_\theta(\mathbf{Z}|\mathbf{Y})$ qui est optimale. Or ceci reviendrait à estimer le maximum de vraisemblance de $p_\theta(\mathbf{Y})$ avec l'algorithme EM, ce qui n'est pas possible car $p_\theta(\mathbf{Z}|\mathbf{Y})$ n'est pas calculable.

Une alternative est de restreindre l'ensemble des distributions sur lequel on veut minimiser $\mathcal{KL}[\mathbb{Q}\|p_\theta(\mathbf{Z}|\mathbf{Y})]$. Il est proposé par exemple dans le cas du SBM (voir Jaakkola, 2000 et Daudin et al., 2008) de prendre la classe des distributions factorisables, *i.e.* qui s'écrivent sous la forme

$$\mathbb{Q}(\mathbf{Z}, \boldsymbol{\tau}) = \prod_{i \in \mathcal{N}} \mathbb{Q}(\mathbf{Z}_i, \boldsymbol{\tau}_i). \tag{1.11}$$

Cette approximation est aussi appelée « champs moyens ». Dans le cadre précis du SBM on prendra $\mathbb{Q}(\mathbf{Z}, \boldsymbol{\tau}) = \otimes_{i \in \mathcal{N}} \mathcal{M}(\mathbf{Z}_i; \boldsymbol{\tau}_i)$ avec $\boldsymbol{\tau}_i = (\tau_{i1}, ..., \tau_{iQ}) \in [0,1]^Q$. On insiste sur le fait que cette approximation n'est pas toujours la plus adéquate car elle dépend du modèle en question. De plus, d'autres algorithmes tel que Belief Propagation (Pearl, 1982) permettent une estimation (exacte pour des arbres) de la loi de $\mathbf{Z}$ sachant $\mathbf{Y}$. Ce n'est pas le cas pour le SBM car conditionnellement à $\mathbf{Y}$, le graphe de dépendance des $\mathbf{Z}_i$ est une clique (*i.e.* tous les nœuds sont reliés par une arête). Ceci est dû au principe de moralisation dans les modèles graphiques (voir Lauritzen, 1996).

Finalement en utilisant les équations (1.9) et (1.10), en remplaçant l'étape E de l'algorithme EM par une étape dite variationnelle E (ou VE) on obtient l'algorithme VEM qui consiste à itérer les 2 étapes suivantes jusqu'à convergence vers un maximum local :

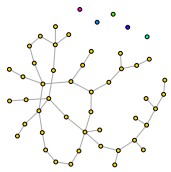
- **VE-step :** avec l'estimation courante $\boldsymbol{\theta}^h$ of $\boldsymbol{\theta}$, calculer

$$\boldsymbol{\tau}^{h+1} = \arg\min_{\boldsymbol{\tau} \in [0,1]^Q} \mathcal{KL}[\mathbb{Q}(\mathbf{Z}; \boldsymbol{\tau})||p_{\boldsymbol{\theta}^h}(\mathbf{Z}|\mathbf{Y})].$$

- **M-step :** mettre à jour l'estimation courante $\boldsymbol{\theta}^h$ de $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^{h+1} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbb{Q}(\mathbf{Z}; \boldsymbol{\tau}_{h+1})}[\log(p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}))].$$

Cet algorithme produit une séquence $\{\boldsymbol{\tau}^h, \boldsymbol{\theta}^h, h \geq 0\}$ croissante pour la fonction $\mathcal{J}_{\boldsymbol{\theta},\mathbb{Q}}$.



Sélection de modèle. L'estimation du nombre de groupes dans les modèles de mélanges avec variables latentes se fait généralement en utilisant le critère ICL (Integrated Completed Likelihood) proposé dans Biernacki et al. (2000). Ce critère est un dérivé du critère BIC (Bayesian Information Criterion) de Schwarz (1978) et pénalise la vraisemblance complète classifiante des données.

2.4 Théorie des données manquantes

La problématique des données manquantes est centrale en statistique. Nous rappelons dans cette partie les rudiments de la théorie introduite dans Rubin (1976) adaptée aux données « réseaux ».

Définition 1 (Matrice et vecteur d'échantillonnage)**.** *Soit* $\mathbf{R} = (R_{ij})_{(i,j)\in\mathcal{D}}$*, alors pour tout* $(i,j) \in \mathcal{D}$*,* $R_{ij} = 1$ *si la dyade* (i,j) *est échantillonnée au cours du processus d'échantillonnage et* $R_{ij} = 0$ *dans le cas contraire. Pour les échantillonnages centrés sur les nœuds, nous utiliserons plus spécifiquement la variable* $\mathbf{V} = (V_i)_{i\in\mathcal{N}} = (\mathbb{1}_{\{i \text{ est échantillonné}\}})_{i\in\mathcal{N}}$ *indiquant quel nœud a été échantillonné.*

Rappelons les notations suivantes : $\mathcal{D}^o := \{(i,j) : R_{ij} = 1\}$, $\mathcal{D}^m := \{(i,j) : R_{ij} = 0\}$, $\mathbf{Y}^o := \{Y_{ij} : (i,j) \in \mathcal{D}^o\}$ et $\mathbf{Y}^m := \{Y_{ij} : (i,j) \in \mathcal{D}^m\}$. Dans la suite nous noterons $\boldsymbol{\psi} \in \Psi$ respectivement le paramètre d'échantillonnage et l'espace des paramètres associé. Nous supposerons tout le temps que les paramètres $\boldsymbol{\theta}$ et $\boldsymbol{\psi}$ vivent dans un espace produit. On considérera aussi que les données précèdent toujours l'échantillonnage, c'est pourquoi nous nous intéressons à la loi de $\mathbf{R}$ sachant $\mathbf{Y}$. Ainsi, en reprenant les travaux de D.B. Rubin nous dirons que les données manquantes sont

(i) *Missing Completely At Random* (MCAR), si l'échantillonnage ne dépend ni des valeurs des données observées ni des valeurs des données non-observées. Mathématiquement cela revient à dire que
$$\mathbf{R} \perp\!\!\!\perp \mathbf{Y}.$$

(ii) *Missing At Random* (MAR), si l'échantillonnage ne dépend que des valeurs des données observées mais pas des valeurs des données non-observées. En termes mathématiques,
$$\mathbf{R} \perp\!\!\!\perp \mathbf{Y}^m \mid \mathbf{Y}^o.$$

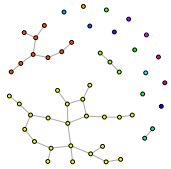
(iii) *Not Missing At Random* (NMAR), si l'échantillonnage dépend des valeurs de toutes les données. C'est le cas par exemple de la censure.

On remarque que (i) est un cas particulier de (ii). L'intérêt de la dichotomie entre le cas MAR et NMAR réside dans la proposition suivante.

Proposition 1. *Si la loi de* $\mathbf{R}$ *satisfait (i) ou (ii), alors pour tout* $\boldsymbol{\psi} \in \Psi$ *tel que* $p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^o, \mathbf{R}) \neq 0$ *:*
$$\arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^o, \mathbf{R}) = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{Y}^o),$$
où p *désigne la vraisemblance des données. Ceci revient à dire que la loi du processus d'échantillonnage ne perturbe pas l'inférence des paramètres du modèle génératif des données.*

2.5 Échantillonner un réseau

Dans le domaine de l'échantillonnage de réseau, nous allons aborder deux problématiques partant de postulats différents. D'un coté la théorie des données manquantes développée par D. Rubin rappelée dans la section 2.4, qui s'inscrit dans un cadre « model-based » au sens où $\mathbf{Y}$ est une variable aléatoire. On s'intéresse dans ce cas explicitement à ce qui caractérise sa loi. D'un autre coté une approche dite « design-based » (voir Kolaczyk, 2009) dans laquelle



$\mathbf{Y}$ n'est pas une variable aléatoire, *i.e.* l'aléa est dû à l'échantillonnage seul. Dans ce second cas, les quantités d'intérêts sont des caractéristiques structurelles du réseau telles que le nombre d'arêtes, la moyenne des degrés des nœuds, la loi des scores de centralité associés aux nœuds ou encore des covariables associés aux nœuds. Dans Handcock and Gile (2010) les auteurs traitent du cas « model-based » pour des échantillonnages MAR dans le cadre des ERGM et du cas « design-based » pour les mêmes échantillonnages.

Cadre « design-based » : estimateurs d'Horvitz-Thompson

Imaginons que l'on veuille estimer la somme de quantités d'intérêts associées aux nœuds d'un réseau (âge, taille, sexe, etc.) notée $s = \sum_{i \in \mathcal{N}} t_i$. Soit $E = \{i_1, ...i_{n_E}\}$ un échantillon de taille n_E de l'ensemble des nœuds tel que pour chaque i dans E la quantité t_i est observée. Supposons de plus que l'échantillonnage utilisé consiste en un tirage uniforme avec remise. Alors $\hat{s} = \frac{n}{n_E} \sum_{i \in E} t_i$ est un estimateur sans biais de s (*i.e.* $\mathbb{E}[\hat{s}] = s$). Si désormais l'échantillonnage n'est pas un simple tirage uniforme avec remise, on se demande si l'estimateur $\hat{s}$ de s est encore sans biais, si oui comment le débiaiser. La réponse est apportée dans Horvitz and Thompson (1952). Supposons que pour tout nœud i dans $\mathcal{N}$ la probabilité qu'il soit dans E est donnée par μ_i. Alors, l'estimateur d'Horvitz-Thompson de s est

$$\hat{s}_\mu = \sum_{i \in E} \frac{t_i}{\mu_i}.$$

C'est un estimateur sans biais de s. En effet :

$$\mathbb{E}[\hat{s}_\mu] = \mathbb{E}\left[\sum_{i \in E} \frac{t_i}{\mu_i}\right] = \mathbb{E}\left[\sum_{i \in \mathcal{N}} \frac{t_i}{\mu_i} \mathbb{1}_{i \in E}\right] = \sum_{i \in \mathcal{N}} \frac{t_i}{\mu_i} \mathbb{E}\left[\mathbb{1}_{i \in E}\right] = s.$$

Pour la suite, nous noterons $\mu_{ij} = \mathbb{P}(i \in E, j \in E)$.

Nous allons maintenant donner deux exemples de stratégies d'échantillonnage ainsi que les probabilités associées aux estimateurs d'Horvitz-Thompson.

Sous-graphe induit. On sélectionne aléatoirement et uniformément un sous-ensemble de nœuds d'un graphe $\mathbb{G} = (N, A)$ et on observe tout le sous graphe induit par ces nœuds (*i.e.* le sous-graphe $\mathbb{G}^\star = (N^\star, A^\star)$ de $\mathbb{G}$ composé des nœuds échantillonnés ($N^\star$ de cardinal $n^\star$) ainsi que les arêtes de $\mathbb{G}$ entre eux ($A^\star$)). On a donc

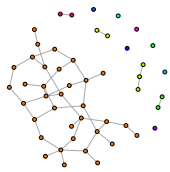
$$\mu_i = \frac{n^\star}{n} \quad \text{et} \quad \mu_{ij} = \frac{n^\star(n^\star - 1)}{n(n-1)}.$$

Sous-graphe incident. On sélectionne par un tirage sans remise un sous-ensemble d'arêtes $A^\star$ (de cardinal $N_{A^\star}$) d'un graphe $\mathbb{G}$ et on observe tous les nœuds incident à ces arêtes : $N^\star$. On a donc

$$\mu_i = \begin{cases} 1 - \frac{\binom{N_A - D_i}{n^\star}}{\binom{N_A}{n^\star}}, & \text{si } n^\star \leq N_A - D_i \\ 1 & , \text{sinon} \end{cases} \quad \text{et} \quad \mu_{ij} = \frac{n^\star}{N_A}$$

ou D_i est le degré du nœud i.

On trouvera dans Kolaczyk (2009) d'autres exemples d'échantillonnages centrés sur les nœuds ainsi que les probabilités d'être échantillonné μ_i et μ_{ij} respectivement associées aux nœuds et aux arêtes.



Cadre « model-based »

Dans cette section, les données $\mathbf{Y}$ suivent un modèle probabiliste. Le but est d'inférer les paramètres de la loi de $\mathbf{Y}$ à partir d'un échantillon incomplet de $\mathbf{Y}$. Suivant le travail de Handcock and Gile (2010), nous allons définir quelques stratégies d'échantillonnage Missing at Random de graphe et nous donnerons leurs lois. Rappelons que dans le cas de données manquantes MAR, l'inférence des paramètres de la loi de $\mathbf{Y}$ se fait sur la partie observée des données uniquement et sans biais induit par la stratégie d'échantillonnage. Nous supposerons par la suite les graphes symétriques et sans boucles, cependant tout ce qui suit se transpose directement au cas des graphes non-symétriques avec ou sans boucles.

Ego-centric design. Cet échantillonnage consiste à sélectionner indépendamment des nœuds du graphe avec probabilité ψ et d'observer toutes les valeurs des dyades entre les nœuds sélectionnés. Ainsi :

$$\mathbb{P}(R_{ij} = 1|\mathbf{Y}, \psi) = 1 - (1-\psi)^2 \ , \ \forall i < j.$$

De plus, en notant 1_n le vecteur colonne de dimension n composé de 1, le vecteur $\mathbf{V}$ définie dans la définition 1 est donnée pas la relation logique : $\mathbf{V} = [\mathbf{R}1_n = (n-1)1_n]$. Réciproquement on peut retrouver $\mathbf{R}$ à partir de $\mathbf{V}$ avec la relation $\mathbf{R} = 1_n \circ \mathbf{V} + \mathbf{V} \circ 1_n - \mathbf{V} \circ \mathbf{V}$, où $\circ$ est le produit de Kronecker entre deux matrices. Donc, en se donnant r et v des réalisations respectivement de $\mathbf{R}$ et $\mathbf{V}$ liés par la relation précédente, on a

$$\mathbb{P}(\mathbf{R} = r|\mathbf{Y}, \psi) = \mathbb{P}(\mathbf{V} = v|\mathbf{Y}, \psi) = \psi^{v^T 1_n}(1-\psi)^{n - v^T 1_n}.$$

Cet échantillonnage est MCAR car la probabilité d'observer une dyade est complètement indépendante de sa valeur et de la valeur des autres dyades du graphe. Ce qui n'est pas le cas des deux stratégies d'échantillonnage que nous allons définir dans la suite (qui sont MAR mais pas MCAR). Dans ces exemples, les dyades seront observées par vagues successives et la probabilité d'observer une dyade dépend des dyades vues aux vagues précédentes.

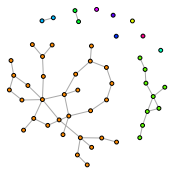
One-wave link-tracing design. Pour cet échantillonnage, on sélectionne indépendamment des nœuds du graphe avec probabilité ψ. Ceci constitue un premier groupe de nœuds $\mathbf{V}_0$. Puis on observe toutes les dyades dont au moins l'une des extrémités inclue un des nœuds sélectionnés auparavant. Les voisins des nœuds du premier groupe (*i.e.* les nœuds partageant une arête avec les nœuds du premier groupe) constituent alors un deuxième groupe de nœuds que l'on notera $\mathbf{V}_1$. Finalement, on observe toutes les dyades incluant les nœuds du deuxième groupe.

On a donc la relation $\mathbf{V} = \mathbf{V}_0 + \mathbf{V}_1$ et $\mathbf{V}_1$ est relié à $\mathbf{V}_0$ et $\mathbf{Y}$ par la relation logique $\mathbf{V}_1 = [\mathbf{Y}\mathbf{V}_0 \odot (1 - \mathbf{V}_0) > 0]$ où $\odot$ est le produit de Hadamard de deux matrices. Par ailleurs les variables $\mathbf{R}$ et $\mathbf{V}$ sont reliées de la même façon que pour l'*ego-centric design*. En conséquence, la loi du *one-wave link-tracing design* est donnée par

$$\mathbb{P}(\mathbf{R} = r|\mathbf{Y}, \psi) = \sum_{v_0,\ v_0 + [\mathbf{Y}v_0 \odot (1-v_0) > 0] = v} \psi^{v_0^T 1_n}(1-\psi)^{n - v_0^T 1_n}.$$

Multi-wave link-tracing design. Cet échantillonnage consiste en l'application successive du *one-wave link-tracing design*. Nous noterons k le nombre de « vagues » successives d'échantillonnages. En utilisant les notations précédemment définies, pour $m \in [\![1, k]\!]$ on définit l'ensemble des nœuds échantillonnés à la k-ième vague sachant toutes les vagues précédentes, par la relation logique $\mathbf{V}_m = [\mathbf{Y}\mathbf{V}_{m-1} \odot (1 - \sum_{t=0}^{m-1} \mathbf{V}_t) > 0]$. Finalement, la loi du *k-waves link tracing design* est donnée par

$$\mathbb{P}(\mathbf{R} = r|\mathbf{Y}, \psi) = \sum_{v_0,\ v_0 + v_1 + \cdots + v_k = k} \psi^{v_0^T 1_n}(1-\psi)^{n - v_0^T 1_n}.$$



3 Contributions de la thèse

Nous décrivons dans cette section ce que nous avons apporté d'original dans l'étude du SBM avec données manquantes.

3.1 Données manquantes dans le SBM

Le SBM est un modèle à variables latentes correspondant aux groupes auxquels appartiennent les nœuds. Cela signifie que sans la connaissance des groupes, la vraisemblance de $\mathbf{Y}$ n'a pas de forme explicite. La notion de « Missing At Random » telle que définie dans Rubin (1976) n'est donc pas applicable directement et nécessite une adaptation. Considérons la vraisemblance des données observées :

$$p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^o, \mathbf{R}) = \int\int p_{\boldsymbol{\theta}}(\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{Z}) p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{Z}) \mathrm{d}\mathbf{Z}\mathrm{d}\mathbf{Y}^m. \tag{1.12}$$

Trouver une condition type MAR pour le cas de modèles à variables latentes revient à chercher quels liens de dépendances entre les variables $\mathbf{Y}$, $\mathbf{R}$ et $\mathbf{Z}$ permettrait d'aboutir à une forme factorisée de (1.12) pour avoir $p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^o, \mathbf{R}) = p_{\boldsymbol{\theta}}(\mathbf{Y}^o) p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^o)$. Ces liens sont explicités dans la figure 1.2 dans laquelle $\mathbf{R}$ n'est pas un nœud parent car nous supposons que les données existent avant l'échantillonnage. Par ailleurs, le lien entre $\mathbf{Z}$ et $\mathbf{Y}$ est donné par la structure de dépendance du SBM.

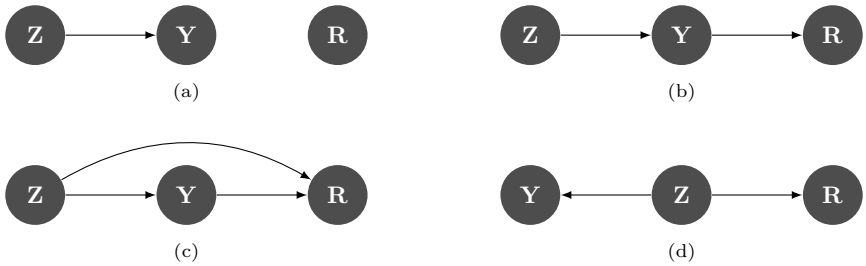


FIGURE 1.2 – Graphes acycliques dirigés de dépendances conditionnelles entre $\mathbf{Y}$, $\mathbf{Z}$ et $\mathbf{R}$ dans un cadre de données manquantes dans le SBM.

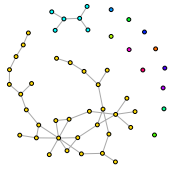
Nous sommes désormais en mesure d'établir une condition nécessaire et suffisante pour que les données manquantes d'un modèle à variables latentes soient Missing At Random. Il faut et il suffit que $\mathbf{R}$ soit indépendante de $(\mathbf{Y}^m, \mathbf{Z})$ conditionnellement à $\mathbf{Y}^o$. On peut montrer que cette condition n'est compatible qu'avec les DAG (a) et (b), le DAG (a) correspondant au cas Missing Completely At Random.

3.2 Stratégies d'échantillonnages

Nous avons vu dans la partie 1 des exemples de réseaux échantillonnés par des stratégies respectivement nœuds-centrées et dyades-centrées. Nous donnons maintenant des exemples de stratégies d'échantillonnages d'un réseau binaire inspirées des données décrites dans la section 1.

Stratégies dyades-centrées :

- ⋆ **Random-dyad sampling :** chaque dyade $(i, j) \in \mathcal{D}$ a la même probabilité $\mathbb{P}(R_{ij} = 1) = \rho := \boldsymbol{\psi}$ d'être observée, indépendamment les unes des autres.
- **Double standard sampling :** soit $\boldsymbol{\psi} := (\rho_1, \rho_0) \in [0, 1]^2$. L'échantillonnage *double standard* (ou *deux poids deux mesures*) est une stratégie pour laquelle la probabilité



d'observer une dyade dépend de sa valeur :

$$(R_{ij})_{i,j} \mid (Y_{ij}) \sim^{ind} \mathcal{B}(\rho_1 \mathbb{1}_{\{Y_{ij}=1\}} + \rho_0 \mathbb{1}_{\{Y_{ij}=0\}}).$$

Le cas $\rho_1 = \rho_0$ correspond au *random-dyad sampling*.

- **Block-dyad sampling :** soit $\boldsymbol{\psi} := (\rho_{q\ell})_{(q,\ell)\in\mathcal{Q}^2} \in \mathcal{M}_Q([0,1])$. L'échantillonnage *block-dyad* (ou *bloc-dyade*) est une stratégie où la probabilité d'observer une dyade (i,j) dépend des classes des nœuds i et j :

$$(R_{ij})_{i,j} \mid (\mathbf{Z}_i), (\mathbf{Z}_j) \sim^{ind} \mathcal{B}(\rho_{\mathbf{Z}_i \mathbf{Z}_j}).$$

Stratégies nœuds-centrées :

⋆ **Star and snowball samplings :** l'échantillonnage *star* (ou *en étoile*) consiste à sélectionner aléatoirement un ensemble de nœuds, puis à observer les lignes correspondantes de la matrice $\mathbf{Y}$. L'échantillonnage *snowball* (ou *boule de neige*) quant à lui se déroule en plusieurs « vagues » dont la première correspond à un échantillonnage *en étoile*. Les vagues successives consistent à observer les voisins des nœuds sélectionnés à la vague précédente. Ces échantillonnages correspondent au *multi-waves link tracing design* défini dans la sous-sous section 2.5 pour respectivement $k = 0$ et $k \geq 1$.

- **Star degree sampling :** l'échantillonnage *star degree* (ou *en étoile basé sur les degrés*) est une stratégie *en étoile* dans laquelle les probabilités d'échantillonner les nœuds $\{\rho_1, \ldots, \rho_n\}$ dépendent des degrés des nœuds de la façon suivante :

$$(V_i)_i \mid (D_i) \sim^{ind} \mathcal{B}(\text{logistic}(a + bD_i))$$

où $\boldsymbol{\psi} := (a,b) \in \mathbb{R}^2$, $D_i = \sum_j Y_{ij}$.

- **Block-node sampling :** *block-node sampling* (*en étoile basé sur les groupes*) est aussi un échantillonnage *en étoile* dans lequel la probabilité d'observer un nœud dépend du groupe auquel il appartient. On a donc

$$(V_i)_i \mid (\mathbf{Z}_i) \sim^{ind} \rho_{\mathbf{Z}_i}$$

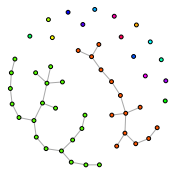
où $\boldsymbol{\psi} := (\rho_1, ..., \rho_Q) \in [0,1]^Q$.

Les échantillonnages marqués par une étoile ⋆ sont MAR (et même MCAR) si les probabilités de sélectionner les nœuds ou les dyades ne dépendent pas de la valeur des dyades ou des groupes auxquels appartiennent les nœuds. Ceux marqués par un rond plein • sont quant à eux NMAR car dépendant de la valeur des dyades : *double standard sampling*, *star degree sampling* ; ou des groupes des nœuds : *class-dyad sampling*, *class sampling*.

3.3 Inférence MAR et NMAR dans le SBM

Inférence dans le cas MAR.

L'inférence du SBM dans le cas ou les données manquantes sont MAR (*a fortiori* MCAR) est identique à celle du SBM sans données manquantes, à ceci près que l'on ne considère plus toutes les entrées de la matrice $\mathbf{Y}$ mais seulement l'ensemble des dyades observées, noté dans la thèse $\mathcal{D}^o$. De plus, le nombre de nœuds observés est composé de l'ensemble des nœuds pour lesquels au moins une dyade est observée : $\mathcal{N}^o := \{i : \exists\, j \in \mathcal{N} \text{ s.t. } R_{ij} = 1\}$.



Inférence dans les cas NMAR.

Nous ne traiterons que du cas des réseaux binaires. La principale difficulté est de trouver une approximation de loi type « champs moyens » de la loi conditionnelle $(\mathbf{Y}^o, \mathbf{R})|(\mathbf{Y}^m, \mathbf{Z})$. Dans Tabouy et al. (2019a), nous avons proposé l'approximation suivante :

$$\mathbb{Q}(\mathbf{Z}, \mathbf{Y}^o) = \prod_{i \in \mathcal{N}} \mathcal{M}(\mathbf{Z}_i, \boldsymbol{\tau}_i) \prod_{(i,j) \in \mathcal{D}^m} \mathcal{B}(Y_{ij}; \nu_{ij}),$$

où l'approximation de la loi de $\mathbf{Z}$ sachant $\mathbf{Y}^m$ est identique aux cas sans données manquantes (i.e. « champs moyen »). Par ailleurs, les Y_{ij} manquants suivent une loi de Bernoulli de paramètres $\nu_{ij} \in \mathbb{R}$, et sont supposés indépendants entre eux ainsi que de $\mathbf{Z}$ et $\mathbf{R}$. À partir de cette approximation on peut écrire un algorithme EM variationnel pour toutes les stratégies décrites dans la partie 3.2. Les détails sont développés dans le chapitre 2.

Sélection de modèle.

L'estimation du nombre de groupe est fait en utilisant le critère ICL (Biernacki et al., 2000) développé pour la sélection de modèle dans les modèles à variables latentes. Nous l'avons adapté au cas de données manquantes pour le SBM. Ce critère est une approximation de la vraisemblance classifiante des données. Comme alternative au critère ICL, on citera les travaux de Clauset et al. (2008) et Guimerà and Sales-Pardo (2009) sur les « liens manquants » et les « liens fallacieux » (*i.e.* observés avec du bruit) dans les réseaux. Les auteurs proposent de prendre en compte l'incertitude existant sur l'absence ou la présence de liens dans les données pour corriger les observations incertaines. C'est dans cette optique que les auteurs de Valles-Catala et al. (2018) et Ghasemian et al. (2018) ont étudiés la possibilité d'estimer le nombre de groupes dans des modèles comme le SBM et ses dérivés. Cependant, leurs observations soulignent la tendance de cette méthode à sur-estimer systématiquement le nombre de groupes.

3.4 Garanties théoriques

Identifiabilité.

Nous avons montré que les paramètres du SBM ainsi que ceux des échantillonnages MCAR *random dyad sampling* et *star sampling* ainsi que de l'échantillonnage NMAR *class sampling* sont identifiables.

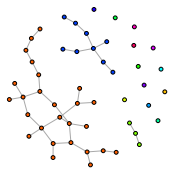
Dans le cas des échantillonnages MCAR, comme il y a indépendance entre $\mathbf{R}$ et $\mathbf{Y}$, l'identifiabilité se montre en deux temps. Dans un premier temps on montre que les paramètres $\boldsymbol{\psi}$ sont identifiables puis que $\boldsymbol{\theta}$ l'est aussi.

Proposition 2. *Le paramètre d'échantillonnage $\rho > 0$ du random-dyad (resp. star) sampling est identifiable par rapport à la distribution de l'échantillonnage.*

Théorème 2. *Soit $n \geq 2Q$. Supposons que pour tout $1 \leq q \leq Q$, $\rho > 0$ les coordonnées de $\boldsymbol{\pi\alpha}$ sont deux à deux différentes. Alors, sous le random-dyad (resp. star) sampling, les paramètres du SBM sont identifiables par rapport à la distribution de la partie observée du SBM, à la permutation des labels près.*

Le théorème suivant établit l'identifiabilité du SBM échantillonné suivant le *class sampling*. À la différence du cas MCAR, on ne peut pas prouver l'identifiabilité de $\boldsymbol{\psi}$ et de $\boldsymbol{\theta}$ séparément, cela doit être fait conjointement à cause de la dépendance entre $\mathbf{R}$ et $\mathbf{Y}$.

Théorème 3. *Soit $n \geq 2Q$ et supposons que pour tout $1 \leq q \leq Q$, $\rho_q > 0$, $\alpha_q > 0$ les coordonnées de $o = \boldsymbol{\pi\alpha}$ et de $t = (\sum_{k=1}^{Q} \pi_{1k}\rho_k\alpha_k, \ldots, \sum_{k=1}^{Q} \pi_{Qk}\rho_k\alpha_k)$ sont deux à deux différentes. Alors, sous le class sampling, les paramètres du SBM sont identifiables par rapport à la distribution du SBM, à la permutation des labels près.*



Les résultats d'identifiabilité associés aux échantillonnages *double standard sampling* et *star degree sampling* ne sont pas encore prouvés.

Consistance et normalité asymptotique.

Suivant les travaux de Celisse et al. (2012), Bickel et al. (2013) et de Brault et al. (2017), nous avons montré que dans le cadre du SBM avec une loi d'émission appartenant à la famille exponentielle à un paramètre et sous les conditions d'échantillonnage du *random dyad sampling*, les estimateurs du maximum de vraisemblance et les estimateurs issus de l'approximation variationnelle du SBM sont consistants et asymptotiquement normaux. De plus, les variances asymptotiques de ces estimateurs sont identiques et explicites. Finalement, nous avons aussi montré que le SBM est identifiable pour toute loi d'émission des dyades appartenant à la famille exponentielle à un paramètre en présence de données manquantes générées par un échantillonnage *random-dyad sampling*. Ces résultats sont développés dans le chapitre 3.

3.5 Prise en compte de covariables

Dans la section 2.1, nous avons vu qu'il était possible d'intégrer la connaissance de covariables sur les nœuds ou sur les dyades dans la modélisation du SBM. Cette inclusion des covariables dans la loi d'émission des dyades conditionnellement aux variables latentes est souvent source d'ambiguïté. Il s'agit non pas de prendre en compte l'effet des covariables sur la structure du graphe, mais bien de l'enlever.

Le but de notre travail est d'inclure dans le SBM la connaissance de covariables, ainsi que de définir des stratégies d'échantillonnage de graphes dépendants des covariables. Nous montrons qu'un SBM binaire avec données manquantes NMAR peut être équivalent à un SBM prenant en compte les covariables avec données manquantes MCAR. Ce travail est motivé par le fait que l'on n'a pas toujours accès aux covariables pendant l'échantillonnage ou dans les données.

Données manquantes, SBM et covariables.

Nous avons étudié trois modèles de SBM avec covariables. Dans le premier, les variables latentes dépendent directement des covariables, dans le second la loi des dyades dépend directement des covariables, finalement le troisième est la combinaison du modèle 1 et 2. L'échantillonnage quant à lui dépend directement des covariables. Les liens de dépendance conditionnelle entre les variables aléatoires $\mathbf{Y}$, $\mathbf{Z}$ et $\mathbf{R}$ associées à ces modèles sont décrits dans la figure 1.3.

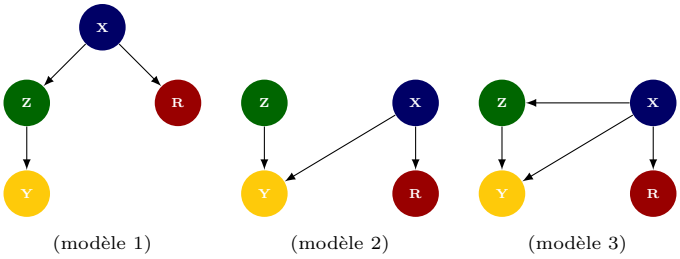


FIGURE 1.3 – Graphes acycliques dirigés de dépendances conditionnelles entre $\mathbf{Y}$, $\mathbf{Z}$ et $\mathbf{R}$ dans un cadre de données manquantes dans le SBM avec covariables.

Les modèles 1 et 2 sont définis ainsi :

Modèle 1. Soit $\boldsymbol{\alpha}_{\cdot} = (\boldsymbol{\alpha}_{\cdot 1}, ..., \boldsymbol{\alpha}_{\cdot Q})$ où $\boldsymbol{\alpha}_{\cdot q} \in [0,1]^Q$ pour tout $q \in \mathcal{Q}$, on définit

$$\alpha_{iq} = \frac{e^{\boldsymbol{\beta}_q^T X_i \mathbb{1}_{\{q \neq Q\}}}}{1 + \sum_{k=1}^{Q-1} \boldsymbol{\beta}_k^T X_i}, \quad \forall (i,q) \in \mathcal{N} \times \mathcal{Q},$$

avec $\beta_q \in \mathbb{R}^N$ pour tout $q \in [\![1, Q-1]\!]$ and $\beta_Q = 0$. De plus nous avons

$$\begin{aligned} (\mathbf{Z}_i)_i \mid (\mathbf{X}_i) &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}_i), \\ (Y_{ij})_{i,j} \mid (\mathbf{Z}_i), (\mathbf{Z}_j) &\sim^{\text{ind}} \mathcal{B}(\pi_{\mathbf{Z}_i \mathbf{Z}_j}), \end{aligned}$$

avec $\pi_{\mathbf{Z}_i \mathbf{Z}_j} \in [0,1]$ pour tout $(i,j) \in \mathcal{N}^2$.

Modèle 2.

$$\begin{aligned} (\mathbf{Z}_i)_i &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}), \\ (Y_{ij})_{i,j} \mid (\mathbf{Z}_i), (\mathbf{Z}_j), (\mathbf{X}_i), (\mathbf{X}_j) &\sim^{\text{ind}} \mathcal{B}(g(\gamma_{z_i z_j} + \boldsymbol{\beta}^T \phi(\mathbf{X}_i, \mathbf{X}_j))), \end{aligned}$$

où $\gamma_{q\ell} \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^m$, $\boldsymbol{\alpha} \in [0,1]^Q$, $g(x) = (1 + e^{-x})^{-1}$ et $\phi(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^m$ est une fonction symétrique (*i.e.* $\phi(x,y) = \phi(y,x)$) mesurant la proximité entre deux vecteurs.

Le modèle suivant est la combinaison des deux modèles précédents.

Modèle 3. Soit $\boldsymbol{\alpha}_{\cdot} = (\boldsymbol{\alpha}_{\cdot 1}, ..., \boldsymbol{\alpha}_{\cdot Q}) \in [0,1]^Q$ alors

$$\alpha_{iq} = \frac{e^{\boldsymbol{\beta}_q^t \mathbf{X}_i \mathbb{1}_{\{q \neq Q\}}}}{1 + \sum_{k=1}^{Q-1} \boldsymbol{\beta}_k^t \mathbf{X}_i}, \quad \forall (i,q) \in \mathcal{N} \times \mathcal{Q},$$

avec $\boldsymbol{\beta}_q \in \mathbb{R}^N$ pour tout $q \in [\![1, Q-1]\!]$ and $\boldsymbol{\beta}_Q = 0$. Alors

$$\begin{aligned} \mathbf{Z}_i \mid \mathbf{X}_i &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}_i), \quad \forall i \in \mathcal{N}, \\ Y_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, \mathbf{X}_{ij} &\sim^{\text{ind}} \mathcal{B}(g(\gamma_{z_i z_j} + \beta^t \mathbf{X}_{ij})), \quad \forall (i,j) \in \mathcal{N}^2, \end{aligned}$$

les paramètres et les notations sont identiques au modèle 2.

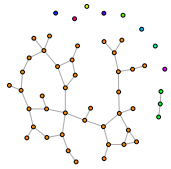
Stratégies d'échantillonnage.

Comme dans la section 3.2 pour le SBM, nous définissons deux stratégies d'échantillonnage (une dyade-centrée et une autre nœud-centrée), pour lesquelles la probabilité d'échantillonner un nœud ou une dyade dépend de la valeur des covariables.

- **Dyad covariates sampling :** soient $\alpha \in \mathbb{R}$, $\kappa \in \mathbb{R}^p$ et $\psi(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^p$. On a donc
$$(R_{ij})_{i,j} \mid (\mathbf{X}_i), (\mathbf{X}_j) \sim^{iid} \mathcal{B}(g(\alpha + \kappa^T \psi(\mathbf{X}_i, \mathbf{X}_j))).$$

- **Node covariates sampling :** soient $\nu \in \mathbb{R}$ et $\eta \in \mathbb{R}^N$. La probabilité d'observer toutes les dyades associées à un nœud est donnée par
$$(V_i)_i \mid (\mathbf{X}_i) \sim^{iid} \mathcal{B}(g(\nu + \eta^T \mathbf{X}_i)).$$

Considérant les modèles 1 et 2 décrit ci-dessus et conditionnellement à $\mathbf{X}$, les stratégies *dyad covariates sampling* et *node covariates sampling* sont MCAR. En particulier, d'après la proposition 1, l'inférence des paramètres des modèles 1 et 2 avec des données manquantes produites suivant ces stratégies d'échantillonnage se fait sur la partie observée des données.



Cas d'équivalence de modèles.

La proposition suivante illustre l'article Molenberghs et al. (2008). En effet, nous montrons qu'un modèle de SBM avec covariables et données manquantes MCAR peut dans certains cas être équivalent à un SBM sans covariables et données manquantes NMAR.

Proposition 3. *Soit $\boldsymbol{X} \in \mathcal{M}_{1\times n}(\mathcal{Q})$, on modélise des covariables et des variables latentes $\mathbf{Z}$ de la façon suivante :*

$$\boldsymbol{X}_i \sim^{iid} \mathcal{M}(1,\nu), \quad \forall i \in \mathcal{N},$$
$$\mathbb{P}(Z_{ia}=1|X_{ia}=1)=\delta \qquad et \qquad \mathbb{P}(Z_{ib}=1|X_{ia}=1)=\frac{1-\delta}{Q-1}, \quad b\neq a,\ \forall i\in\mathcal{N}.$$

De plus, on définit la probabilité d'observer une dyade comme suit :

$$p_{q\ell}=\mathbb{P}(R_{ij}=1|X_{iq}=1,X_{j\ell}=1).$$

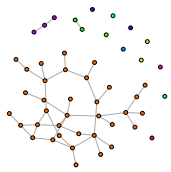
On utilise enfin le modèle 1 pour la loi des dyades conditionnellement à $\boldsymbol{X}$ et $\mathbf{Z}$.
Alors ce modèle est équivalent à un SBM binaire échantillonné avec une stratégie block-dyad sampling dont les paramètres d'échantillonnage seraient les suivants :

$$\begin{aligned}\rho_{q\ell} &= \mathbb{P}(R_{ij}=1|Z_{iq}=1,Z_{j\ell}=1),\\ &= \frac{\delta^2 p_{q\ell}\nu_q\nu_\ell+\left(\frac{1-\delta}{Q-1}\right)^2\sum_{\substack{a\neq q\\ b\neq\ell}}p_{ab}\nu_a\nu_b+2\delta\left(\frac{1-\delta}{Q-1}\right)\sum_{\substack{a\neq q\\ b=\ell}}p_{ab}\nu_a\nu_b}{(\delta\nu_q+\frac{1-\delta}{Q-1}\sum_{c\neq q}\nu_c)(\delta\nu_\ell+\frac{1-\delta}{Q-1}\sum_{c\neq\ell}\nu_c)}.\end{aligned}$$

Ce résultat théorique est illustré par des simulations dans le chapitre 4.

3.6 Package R : missSBM

Une partie des développements méthodologiques de la thèse ont été implémentés dans le package R **missSBM** (Tabouy et al., 2019c) disponible sur le CRAN (`https://CRAN.R-project.org/package=missSBM`). Il contient plusieurs fonctions permettant (i) de simuler un SBM binaire avec ou sans covariables (`simulate`), (ii) d'échantillonner un réseau selon les stratégies décrites dans la section 3.2 (`sampling`) et (iii) d'inférer les paramètres d'un SBM binaire avec ou sans covariables, de prédire les groupes des nœuds avec des données manquantes ou pas en prenant en compte la stratégie d'échantillonnage employée (`estimate`). L'inférence est conduite avec l'algorithme VEM décrit dans la section 2.3. Les détails du package, de sa structure et de son utilisation sont développés au chapitre 5.



Chapter 2

Variational inference of Stochastic Block Model from sampled data

This chapter has been published as an article in the Journal of the American Statistical Association (Tabouy et al., 2019a).

Contents

This paper deals with non-observed dyads during the sampling of a network and consecutive issues in the inference of the Stochastic Block Model (SBM). We review sampling designs and recover Missing At Random (MAR) and Not Missing At Random (NMAR) conditions for the SBM. We introduce variants of the variational EM algorithm for inferring the SBM under various sampling designs (MAR and NMAR) all available as an `R` package on the CRAN (see chapter 5 for more details). Model selection criteria based on Integrated Classification Likelihood are derived for selecting both the number of blocks and the sampling design. We investigate the accuracy and the range of applicability of these algorithms with simulations. We explore two real-world networks from ethnology (seed circulation network) and biology (protein-protein interaction network), where the interpretations considerably depends on the sampling designs considered.

1 Introduction

Networks arise in many fields of application for providing an intuitive way to represent interactions between entities. In this paper, a network is composed by a fixed set of nodes, and an interaction between a pair of nodes (dyad) is called an edge. We consider undirected binary networks with no loop, which can be represented by symmetric adjacency matrices filled with zeros and ones.

Various statistical models exist for depicting the probability distribution of the adjacency matrix (see, e.g. Goldenberg et al., 2010 and Snijders, 2011, for a survey). A highly desirable feature is their capability to describe the heterogeneity of real-world networks. In this perspective, the family of models endowed with a latent structure (reviewed in Matias and Robin, 2014) offers a natural way to introduce heterogeneity. Within this family the Stochastic Block Model (in short SBM, see Frank and Harary, 1982 and Holland et al., 1983) describes a broad variety of network topologies by positing a latent structure (or a clustering) on the nodes, then making the probability distribution of the adjacency matrix dependent on this latent structure. In order to estimate SBMs, Bayesian approaches were first developed (Snijders and Nowicki, 1997 and Nowicki and Snijders, 2001) prior to variational approaches (Daudin et al., 2008 and Latouche et al., 2012). On the theoretical side, Celisse et al. (2012) study the conditions for identifiability and the consistency of the variational estimators; Bickel et al. (2013) prove their asymptotic normality. Several generalizations are possible such as weighted or directed variants (Mariadassou et al., 2010), mixed-membership and overlapping SBM (Airoldi et al., 2008 and Latouche et al., 2011), degree-corrected SBM (Karrer and Newman, 2011), dynamic SBM (Matias and Miele, 2016), or multiplex SBM (Barbillon et al., 2015).

This paper deals with inference in the SBM when the network is not fully observed. We consider cases where all the nodes are observed but information regarding the presence/absence of an edge is missing for some dyads. In other words the adjacency matrix contains missing values, a situation often met with real-world networks. For instance in social sciences, network data consists in interactions between individuals: the set of individuals is fixed, possibly known from a census. Information about the presence/absence of an edge is only available when at least one of the two individuals is available for an interview, otherwise it is missing. See Thompson and Frank (2000), Thompson and Seber (1996), Kolaczyk (2009) and Handcock and Gile (2010) for a review of network sampling techniques. Even though some papers deal with SBM inference under missing data condition (Aicher et al., 2014 and Vinayak et al., 2014), the sampling mechanism responsible for the missing values is overlooked in the inference, contrary to the approach developed in our paper.

Our contributions. A typology of sampling designs is introduced in Section 2.2. We adapt the theory developed in Rubin (1976) and Little and Rubin (2014) to the SBM by splitting the sampling designs into the three usual classes of missing data:

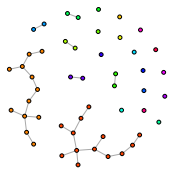
i) *Missing Completely At Random* (MCAR), where the sampling does not depend on the data, neither on the observed nor on the unobserved part of the network.

ii) *Missing At Random* (MAR), where the probability of being sampled is independent on the value of the missing data. For network data, the sampling does not depend on the presence/absence of an edge of an unobserved (or missing) dyad. MCAR is a particular case of MAR.

iii) *Not Missing At Random* (NMAR), where the sampling scheme is guided by unobserved dyads in some way.

Section 2.3 introduces several examples of sampling designs (MAR and NMAR) for which we derive conditions for identifiability of the SBM parameters.

Estimation of the SBM in the MAR cases can be handled with the Variational EM (VEM) of Daudin et al. (2008) by conducting the inference only on the observed part of the



network (Section 3.1). NMAR is more difficult to deal with as the sampling design must be taken into account in the inference. We introduce in Section 3.2 a general variational algorithm (Jordan et al., 1998) to deal with NMAR cases when the sampling design relies on a probability distribution which is explicitly known[1]. Our variational approach is based on a double mean-field approximation applied to the latent distribution of the clustering and to the distribution of the missing dyads. We implement VEM algorithms that produce unbiased estimators for three natural NMAR sampling designs: a dyad-centered strategy, a node-centered strategy, and a block-centered strategy. We also derive an Integrated Classification Likelihood criterion (ICL, Biernacki et al., 2000) for selecting the number of blocks. Although it is not possible to distinguish whether the sampling is MAR or NMAR (Molenberghs et al., 2008), the ICL can also be used to select which sampling design is the best fit for the data.

In Section 4.2 we show the good performance of our VEM algorithms on simulations for both MAR (Section 4.1) and NMAR conditions. Finally we investigate two very different real-world networks with missing values, namely a Kenyan seed exchange network (Section 5.1), and a protein-protein interaction (PPI) network (Section 5.2).

Related works. In the few papers dealing with missing data for networks, the sampling design is rarely discussed. Even if not explicitly stated they all assume MAR conditions. Aicher et al. (2014) propose a weighted SBM modeling simultaneously the presence/absence of an edge and its weight. Missing data are handled by dropping the corresponding terms in the likelihood and the inference is conducted by a variational algorithm. In Vincent and Thompson (2015) a Bayesian augmentation procedure is introduced to estimate simultaneously the size of the population and the clustering when the sampling design is a one-wave snowball. Apart from the SBM, the exponential random graph model has been studied in the MAR setting in Handcock and Gile (2010).

The matrix completion literature brings additional insights since SBM inference can be seen as a low-rank matrix estimation. Vinayak et al. (2014) introduce a convex program for the matrix completion problem where the underlying matrix has a simple affiliation structure defined via an SBM. The entries are sampled independently with the same probability, corresponding to a MAR case. In Davenport et al. (2014) the case of noisy 1-bit observations is studied and a likelihood-based strategy is developed with theoretical justifications ensuring good matrix completion. Chatterjee (2015) proves strong results for large matrices with noisy entries estimation, by means of a universal singular value thresholding.

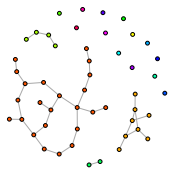
Another related question is when the status of some dyads (absence/presence) is not clear in errorfully observed graph. Such uncertainties can be taken into account (Priebe et al., 2015 and Balachandran et al., 2017). The latter reference studies the error propagation made by using estimators computed on observed sub-graphs, in order to estimate the number of existing edges in the real underlying graph.

2 Statistical framework

2.1 Stochastic Block Model

In an SBM, nodes from a set $\mathcal{N} \triangleq \{1, \dots, n\}$ are distributed among a set $\mathcal{Q} \triangleq \{1, \dots, Q\}$ of hidden blocks that model the latent structure of the graph. The blocks are described by the latent random vectors $\left(Z_{i\bullet} = (Z_{i1}, \dots, Z_{iQ})\right)_{i \in \mathcal{N}}$ with multinomial distribution $\mathcal{M}(1, \alpha = (\alpha_1, \dots, \alpha_Q))$. The probability of an edge between any dyad in $\mathcal{D} \triangleq \mathcal{N} \times \mathcal{N}$ only depends on the blocks the two nodes belong to. Hence, the presence of an edge between i and j, indicated by the binary variable Y_{ij}, is independent on the other edges conditionally on the

[1] More complex sampling schemes – for instance adversarial strategies – are thus not handled



latent blocks:

$$Y_{ij} \mid Z_{iq} = 1, Z_{j\ell} = 1 \sim^{\text{ind}} \mathcal{B}(\pi_{q\ell}), \qquad \forall (i,j) \in \mathcal{D}, \quad \forall (q,\ell) \in \mathcal{Q} \times \mathcal{Q},$$

where $\mathcal{B}$ stands for the Bernoulli distribution. In the following, $\pi = (\pi_{q\ell})_{(q,\ell)\in\mathcal{Q}\times\mathcal{Q}}$ is the $Q \times Q$ matrix of connectivity probabilities, $\mathbf{Y} = (Y_{ij})_{(i,j)\in\mathcal{D}}$ is the $n \times n$ adjacency matrix of the random graph, $Z = (Z_{iq})_{i\in\mathcal{N}, q\in\mathcal{Q}}$ is the $n \times Q$ matrix of the latent blocks and $\theta = (\alpha, \pi)$ are the unknown parameters. In the undirected binary case, $Y_{ij} = Y_{ji}$ for all $(i,j) \in \mathcal{D}$ and $Y_{ii} = 0$ for all $i \in \mathcal{N}$. Similarly, $\pi_{q\ell} = \pi_{\ell q}$ for all $(q,\ell) \in \mathcal{Q} \times \mathcal{Q}$.

2.2 Sampled data in the SBM framework

The sampled data is an $n \times n$ matrix with entries in $\{0, 1, \texttt{NA}\}$. It corresponds to the adjacency matrix $\mathbf{Y}$ where unobserved dyads have been replaced by `NA`'s. More formally, let R be the $n \times n$ sampling matrix recording the data sampled during this process, such that $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise; also define $\mathcal{D}^o = \{(i,j) : R_{ij} = 1\}$, $\mathcal{D}^m = \{(i,j) : R_{ij} = 0\}$, $\mathbf{Y}^o = \{Y_{ij} : (i,j) \in \mathcal{D}^o\}$ and $\mathbf{Y}^m = \{Y_{ij} : (i,j) \in \mathcal{D}^m\}$ to denote the sets of variables respectively associated with the *observed* and *missing* data. The number of nodes n is assumed to be known. The *sampling design* is the description of the stochastic process that generates $\mathbf{R}$. It is assumed that the network exists before the sampling design acts upon it. Moreover, the sampling design is fully characterized by the conditional distribution $p_\psi(\mathbf{R}|\mathbf{Y})$, the parameters of which are such that ψ and θ live in a product space $\Theta \times \Psi$. Hence the joint probability density function of the observed data satisfies

$$p_{\theta,\psi}(\mathbf{Y}, \mathbf{R}) = \int\int p_\theta(\mathbf{Y}^o, \mathbf{Y}^m, \mathbf{Z}) p_\psi(\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m, Z) \mathrm{d}\mathbf{Y}^m \mathrm{d}\mathbf{Z}. \tag{2.1}$$

Simplifications may occur in (2.1) depending on the sampling design, leading to the three usual types of missingness (MCAR, MAR and NMAR). This typology depends on the relations between the adjacency matrix $\mathbf{Y}$, the latent structure $\mathbf{Z}$ and the sampling $\mathbf{R}$, so that the missingness is characterized by four directed acyclic graphs displayed in Figure 2.1.

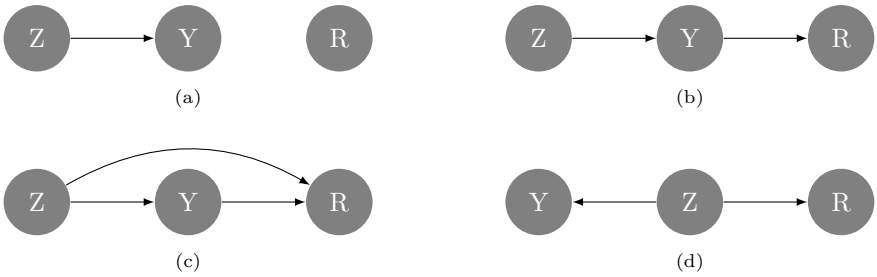
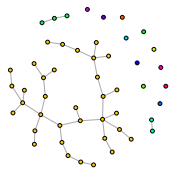


Figure 2.1 – DAGs of relationships between $\mathbf{Y}, \mathbf{Z}$ and $\mathbf{R}$ in the framework of missing data for SBM. DAG where $\mathbf{R}$ is a parent node are not reviewed since the network exists before the sampling design acts upon it. The systematic edge between $\mathbf{Z}$ and $\mathbf{Y}$ is due to the definition of the SBM. Note that the DAG (b) may correspond to MAR or NMAR samplings.

On the basis of these DAGs, the sampling design is MCAR if $\mathbf{R} \perp\!\!\!\perp (\mathbf{Y}^m, \mathbf{Z}, \mathbf{Y}^o)$, MAR if $\mathbf{R} \perp\!\!\!\perp (\mathbf{Y}^m, \mathbf{Z}) \mid \mathbf{Y}^o$, and NMAR otherwise. We derive Proposition 4 from these definitions.

Proposition 4. *If the sampling is MCAR or MAR then i)* $\arg\max_\theta p_{\theta,\psi}(\mathbf{Y}^o, \mathbf{R}) = \arg\max_\theta p_\theta(\mathbf{Y}^o)$ *for any* ψ *such that* $p_{\theta,\psi}(\mathbf{Y}^o, \mathbf{R}) \neq 0$ *and ii) the sampling design necessary satisfies DAG* (a) *or* (b).

Proof. To prove $i)$, if $\mathbf{R}$ satisfies MAR conditions, then $p_\psi(\mathbf{R}|\mathbf{Y}^o, \mathbf{Y}^m, Z) = p_\psi(\mathbf{R}|\mathbf{Y}^o)$. Moreover, θ and ψ lie in a product space so that (2.1) factorizes into $p_{\theta,\psi}(\mathbf{Y}^o, \mathbf{R}) = p_\theta(\mathbf{Y}^o) p_\psi(\mathbf{R}|\mathbf{Y}^o)$. This corresponds to the ignorability condition of Rubin (1976) and Handcock and Gile (2010). The proof of $ii)$ is postponed to the supplementary materials. □



2.3 Sampling design examples

MAR examples

Definition 1 (Random-dyad sampling)**.** *Each dyad $(i, j) \in \mathcal{D}$ has the same probability $\mathbb{P}(R_{ij} = 1) = \rho$ to be observed independently of the others.*

This design is trivially MCAR because each dyad is sampled with the same probability ρ which does not depend on $\mathbf{Y}$.

Definition 2 (Star and snowball sampling)**.** *The star sampling consists in selecting uniformly a set of nodes, then observing corresponding rows of matrix $\mathbf{Y}$. Snowball sampling is initialized by a star sampling which gives a first "wave" of nodes. The second wave is composed by the neighbors of the first. Successive waves can then be obtained. The final set of observed dyads corresponds to all dyads involving at least one of these nodes.*

These two designs are node-centered and MAR. Indeed, selecting nodes independently in star sampling or in the first wave of snowball sampling corresponds to MCAR sampling. Successive waves are then MAR since they are built on the basis of the previously observed part of $\mathbf{Y}$. Expressions of the corresponding distributions $p_\psi(\mathbf{R}|\mathbf{Y}^o)$ are given in Handcock and Gile (2010).

Identifiability of random-dyad and star sampling designs. Since random-dyad and star samplings are MCAR, the identifiability is assessed in two steps by proving the identifiability of, first, the sampling parameter $\psi = \rho$ and second, the SBM parameters $\theta = (\alpha, \pi)$ given ρ. Our proofs, postponed to the supplementary materials, follow Celisse et al. (2012) who established the identifiability of the SBM without missing data.

Proposition 5. *The sampling parameter $\rho > 0$ of random-dyad (resp. star) sampling is identifiable w.r.t. the sampling distribution.*

Theorem 1. *Let $n \geq 2Q$ and assume that for any $1 \leq q \leq Q$, $\rho > 0$, $\alpha_q > 0$ and that the coordinates of $\pi\alpha$ are pairwise distinct. Then, under random-dyad (resp. star) sampling, SBM parameters are identifiable w.r.t. the distribution of the observed part of the SBM up to label switching.*

NMAR examples

Definition 3 (Double standard sampling)**.** *Let $\rho_1, \rho_0 \in [0, 1]$. Double standard sampling consists in observing dyads with probabilities*

$$\mathbb{P}(R_{ij} = 1|Y_{ij} = 1) = \rho_1, \qquad \mathbb{P}(R_{ij} = 1|Y_{ij} = 0) = \rho_0. \tag{2.2}$$

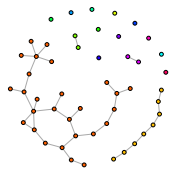
Denote $S^{\mathrm{o}} = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}} Y_{ij}$, $\bar{S}^{\mathrm{o}} = \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}}(1 - Y_{ij})$ and similarly for $S^m, \bar{S}^m$. In this dyad-centered sampling design satisfying DAG (b), the log-likelihood is

$$\log p_\psi(R|Y) = S^{\mathrm{o}} \log \rho_1 + \bar{S}^{\mathrm{o}} \log \rho_0 + S^{\mathrm{m}} \log(1 - \rho_1) + \bar{S}^{\mathrm{m}} \log(1 - \rho_0), \quad \text{with } \boldsymbol{\psi} = (\rho_0, \rho_1). \tag{2.3}$$

Definition 4 (Star sampling based on degrees – Star degree sampling)**.** *Star degree sampling consists in observing all dyads corresponding to nodes selected with probabilities $\{\rho_1, \ldots, \rho_n\}$ such that $\rho_i = \mathrm{logistic}(a+bD_i)$ for all $i \in \mathcal{N}$ where $(a, b) \in \mathbb{R}^2$, $D_i = \sum_j Y_{ij}$ and $\mathrm{logistic}(x) = (1 + e^{-x})^{-1}$.*

In this node-centered sampling design satisfying DAG (b), the log-likelihood is

$$\log p_\psi(R|Y) = \sum_{i\in\mathcal{N}^{\mathrm{o}}} \log \rho_i + \sum_{i\in\mathcal{N}^{\mathrm{m}}} \log(1 - \rho_i), \qquad \text{with } \boldsymbol{\psi} = (a, b). \tag{2.4}$$



Definition 5 (Class sampling)**.** *Class sampling consists in observing all dyads corresponding to nodes selected with probabilities* $\{\rho_1, \dots, \rho_Q\}$ *such that* $\rho_q = \mathbb{P}(i \in \mathcal{N}^{\mathrm{o}} \mid Z_{iq} = 1)$ *for all* $(i, q) \in \mathcal{N} \times \mathcal{Q}$.

In this node-centered sampling design satisfying DAG (d), the log-likelihood is

$$\log p_\psi(R|Z) = \sum_{i \in \mathcal{N}^{\mathrm{o}}} \sum_{q \in \mathcal{Q}} Z_{iq} \log \rho_q + \sum_{i \in \mathcal{N}^{\mathrm{m}}} \sum_{q \in \mathcal{Q}} Z_{iq} \log(1 - \rho_q), \qquad \text{with } \boldsymbol{\psi} = (\rho_1, \dots, \rho_Q). \tag{2.5}$$

Identifiability of class sampling. Theorem 2 establishes the identifiability of the SBM sampled under NMAR class sampling design (see the supplementary materials for the proof). Note that the identifiability of the sampling parameters $\psi = (\rho_1, \dots, \rho_Q)$ and of the SBM parameters must be proved jointly because of the dependence between the network and the sampling. It is worth mentioning that both α_q and ρ_q are identifiable and not only their product. Although somewhat counter-intuitive, this fact is supported by the inference algorithm for class sampling in Section 3.3, which weights the recovery of the latent clusters by taking the unbalanced sampling into account.

Theorem 2. *Let* $n \geq 2Q$ *and assume that for any* $1 \leq q \leq Q$, $\rho_q > 0$, $\alpha_q > 0$, *and that the coordinates of* $o = \pi\alpha$ *and* $t = (\sum_{k=1}^{Q} \pi_{1k}\rho_k\alpha_k, \dots, \sum_{k=1}^{Q} \pi_{Qk}\rho_k\alpha_k)$ *are pairwise distinct. Then, under class sampling, SBM and class sampling parameters are identifiable w.r.t. the distributions of the SBM and the sampling up to label switching.*

3 Variational Inference

Derivations of the practical variational algorithms considerably change depending on the missing data condition at play. We start by MAR to gently introduce the variational principle for SBM, then develop algorithms in a series of NMAR conditions

3.1 MAR inference

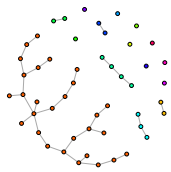
By Proposition 4 part (i), inference in the MAR case is conducted on $\mathbf{Y}^{\mathrm{o}}$. The EM algorithm is unfeasible since it requires the evaluation of the conditional mean of the complete log-likelihood $\mathbb{E}_{Z|\mathbf{Y}^{\mathrm{o}}}[\log p_\theta(\mathbf{Y}^{\mathrm{o}}, \mathbf{Z})]$ which is intractable when $\mathbf{Y}$ comes from an SBM. The variational approach circumvents this limitation by maximizing a lower bound of the log-likelihood based on an approximation $\tilde{p}_\tau$ of the true conditional distribution $p_\theta(\mathbf{Z}|\mathbf{Y}^{\mathrm{o}})$,

$$\begin{aligned} \log p_\theta(\mathbf{Y}^{\mathrm{o}}) \geq J_{\tau,\theta}(\mathbf{Y}^{\mathrm{o}}) &\triangleq \log(p_\theta(\mathbf{Y}^{\mathrm{o}})) - \mathrm{KL}[\tilde{p}_\tau(\mathbf{Z}) || p_\theta(\mathbf{Z}|\mathbf{Y}^{\mathrm{o}})], \\ &= \mathbb{E}_{\tilde{p}_\tau}[\log(p_\theta(\mathbf{Y}^{\mathrm{o}}, \mathbf{Z}))] - \mathbb{E}_{\tilde{p}_\tau}[\log \tilde{p}_\tau(\mathbf{Z})], \end{aligned} \tag{2.6}$$

where τ are some variational parameters and KL is the Kullback-Leibler divergence. The approximated distribution is chosen so that the integration over the latent variables simplifies by factorization. Recall from Section 2.1 that the latent vectors $\big(Z_{i\bullet} = (Z_{i1}, \dots, Z_{iQ})\big)_{i \in \mathcal{N}}$ are independent with a multinomial prior distribution. Thus, in order to factorize the likelihood in a convenient way, a natural variational counterpart to $p_\theta(\mathbf{Z}|\mathbf{Y}^{\mathrm{o}})$ is $\tilde{p}_\tau(\mathbf{Z}) = \prod_{i \in \mathcal{N}} m(Z_{i\bullet}; \boldsymbol{\tau}_i)$, where $\boldsymbol{\tau}_i = (\tau_{i1}, \dots, \tau_{iQ})$, and $m(\cdot; \boldsymbol{\tau}_i)$ is the multinomial probability density function with parameters $\boldsymbol{\tau}_i$. The VEM sketched in Algorithm 1 consists in alternatively maximizing J w.r.t. $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_n\}$ (the variational E-step) and w.r.t. $\boldsymbol{\theta}$ (the M-step). The two maximization problems are solved straightforwardly following Daudin et al. (2008):

1. The parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ maximizing $J_{\boldsymbol{\theta}(\mathbf{Y}^{\mathrm{o}})}$ when $\boldsymbol{\tau}$ is held fixed are

$$\hat{\alpha}_q = \frac{\sum_{i \in \mathcal{N}^{\mathrm{o}}} \hat{\tau}_{iq}}{\mathrm{card}\,(\mathcal{N}^{\mathrm{o}})}, \qquad \hat{\pi}_{q\ell} = \frac{\sum_{(i,j) \in \mathcal{D}^{\mathrm{o}}} \hat{\tau}_{iq}\hat{\tau}_{j\ell} Y_{ij}}{\sum_{(i,j) \in \mathcal{D}^{\mathrm{o}}} \hat{\tau}_{iq}\hat{\tau}_{j\ell}}.$$



2. The variational parameters $\boldsymbol{\tau}$ maximizing $J_{\boldsymbol{\tau}(\mathbf{Y}^{\mathrm{o}})}$ when $\boldsymbol{\theta}$ is held fixed are obtained with the following fixed point relation:

$$\hat{\tau}_{iq} \propto \alpha_q \left(\prod_{(i,j)\in\mathcal{D}^{\mathrm{o}}} \prod_{\ell\in\mathcal{Q}} b(Y_{ij};\pi_{q\ell})^{\hat{\tau}_{j\ell}} \right),$$

where $b(x,\pi) = \pi^x(1-\pi)^{1-x}$ the Bernoulli probability density function.

Algorithm 1: Variational EM for MAR inference in SBM

Initialization: Set up $\boldsymbol{\tau}^{(0)}$ with some clustering algorithm

repeat

$$\boldsymbol{\theta}^{(h+1)} = \arg\max_{\boldsymbol{\theta}} J\left(\mathbf{Y}^{\mathrm{o}};\boldsymbol{\tau}^{(h)},\boldsymbol{\theta}\right) \qquad \textbf{M-step}$$

$$\boldsymbol{\tau}^{(h+1)} = \arg\max_{\boldsymbol{\tau}} J\left(\mathbf{Y}^{\mathrm{o}};\boldsymbol{\tau},\boldsymbol{\theta}^{(h+1)}\right) \qquad \textbf{variational E-step}$$

until $\left\|\boldsymbol{\theta}^{(h+1)} - \boldsymbol{\theta}^{(h)}\right\| < \varepsilon$

Algorithm 1 generates a sequence $\{\boldsymbol{\tau}^{(h)},\boldsymbol{\theta}^{(h)}; h \geqslant 0\}$ with increasing $J(\mathbf{Y}^{\mathrm{o}};\boldsymbol{\tau}^{(h)},\boldsymbol{\theta}^{(h)})$. Since there is no guarantee for convergence to the global maximum, we run the algorithm from several different initializations to finally retain the best solution.

Model selection of the number of blocks. The Integrated Classification Likelihood (ICL) criterion of Biernacki et al. (2000) is relevant for latent variable models where the likelihood – and thus BIC – is intractable. Daudin et al. (2008) derive a variational ICL for the SBM which we adapt to missing data conditions: if $\hat{\boldsymbol{\theta}} = \arg\max \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{\mathrm{o}},\mathbf{Z})$ then

$$\mathrm{ICL}(Q) = -2\mathbb{E}_{\tilde{p}_{\boldsymbol{\tau}}}\left[\log p_{\hat{\boldsymbol{\theta}}}(\mathbf{Y}^{\mathrm{o}},\mathbf{Z};Q)\right] + \frac{Q(Q+1)}{2}\log\mathrm{card}\left(\mathcal{D}^{\mathrm{o}}\right) + (Q-1)\log\mathrm{card}\left(\mathcal{N}^{\mathrm{o}}\right).$$

Note that each dyad is only counted once since we work with symmetric networks.

3.2 NMAR inference: the general case

In contrast to the MAR case, conducting inference on the observed dyads only may bias the estimates in the NMAR case. In fact, all observed data (including the sampling matrix $\mathbf{R}$ in addition to $\mathbf{Y}^{\mathrm{o}}$) must be taken into account. The likelihood of the observed data is thus $\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R})$ and the corresponding completed likelihood has the following decomposition:

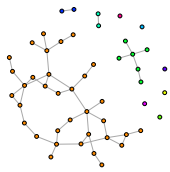
$$\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R},\mathbf{Y}^{\mathrm{m}},\mathbf{Z}) = \log p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{Z}) + \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{Z}), \tag{2.7}$$

where an explicit form of $p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{Z})$ requires further specification of the sampling. The joint distribution $p_{\boldsymbol{\theta}}(\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{Z})$ has a form similar to the MAR case. Now, the approximation is required both on latent blocks $\mathbf{Z}$ and missing dyads $\mathbf{Y}^{\mathrm{m}}$ to approximate $p_{\boldsymbol{\theta}}(\mathbf{Z},\mathbf{Y}^{\mathrm{m}}|\mathbf{Y}^{\mathrm{o}})$. We suggest a variational distribution where complete independence is forced on $\mathbf{Z}$ and $\mathbf{Y}^{\mathrm{m}}$, using a multinomial (resp. Bernoulli) distribution for $\mathbf{Z}$ (resp. for $\mathbf{Y}^{\mathrm{m}}$):

$$\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}(\mathbf{Z},\mathbf{Y}^{\mathrm{m}}) = \tilde{p}_{\boldsymbol{\tau}}(\mathbf{Z})\,\tilde{p}_{\boldsymbol{\nu}}(\mathbf{Y}^{\mathrm{m}}) = \prod_{i\in\mathcal{N}} m(Z_{i\cdot};\boldsymbol{\tau}_i) \prod_{(i,j)\in\mathcal{D}^{\mathrm{m}}} b(Y_{ij};\nu_{ij}), \tag{2.8}$$

where $\boldsymbol{\tau}$ and $\boldsymbol{\nu} = \{\nu_{ij},(i,j)\in\mathcal{D}^{\mathrm{m}}\}$ are two sets of variational parameters respectively associated with $\mathbf{Z}$ and $\mathbf{Y}^{\mathrm{m}}$. This leads to the following lower bound for $\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R})$:

$$J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R}) = \mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}}\left[\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R},\mathbf{Y}^{\mathrm{m}},\mathbf{Z})\right] - \mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}}\left[\log \tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}(\mathbf{Z},\mathbf{Y}^{\mathrm{m}})\right].$$



By means of Decomposition (2.7) of the completed log-likelihood, variational approximation (2.8) and entropies of multinomial and Bernoulli distributions, one has

$$\begin{aligned} J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R}) &= \mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}}\left[\log p_{\psi}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{Z})\right] \\ &+ \sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}}\sum_{(q,\ell)\in\mathcal{Q}^2}\tau_{iq}\tau_{j\ell}\log b(Y_{ij},\pi_{q\ell}) + \sum_{(i,j)\in\mathcal{D}^{\mathrm{m}}}\sum_{(q,\ell)\in\mathcal{Q}^2}\tau_{iq}\tau_{j\ell}\log b(\nu_{ij},\pi_{q\ell}) \\ &+ \sum_{i\in\mathcal{N}}\sum_{q\in\mathcal{Q}}\tau_{iq}\log(\alpha_q/\tau_{iq}) - \sum_{(i,j)\in\mathcal{D}^{\mathrm{m}}}\nu_{ij}\log(\nu_{ij}) + (1-\nu_{ij})\log(1-\nu_{ij}). \end{aligned} \tag{2.9}$$

In (2.9), $\mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}}\left[\log p_{\psi}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{Z})\right]$ can be integrated over the variational distribution $\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}(\mathbf{Z},\mathbf{Y}^{\mathrm{m}})$, as expected. The practical computations depend on the sampling design.

The general VEM algorithm used to maximize (2.9) is sketched in its main lines in Algorithm 2. Both the E-step and the M-step split into two parts: the maximization must be performed on the SBM parameters $\boldsymbol{\theta}$ and the sampling design parameters $\boldsymbol{\psi}$ respectively. The variational E-step is performed on the parameters $\boldsymbol{\tau}$ of the latent block $\mathbf{Z}$ and on the parameters $\boldsymbol{\nu}$ of the missing data $\mathbf{Y}^{\mathrm{m}}$.

Algorithm 2: Variational EM for NMAR inference in SBM

Initialisation: set up $\boldsymbol{\tau}^{(0)}$, $\boldsymbol{\nu}^{(0)}$ and $\boldsymbol{\psi}^{(0)}$

repeat

$$\begin{aligned} \boldsymbol{\theta}^{(h+1)} &= \arg\max_{\boldsymbol{\theta}} J\left(\mathbf{Y}^{\mathrm{o}},\mathbf{R};\ \boldsymbol{\tau}^{(h)},\boldsymbol{\nu}^{(h)},\boldsymbol{\psi}^{(h)},\boldsymbol{\theta}\right) && \textbf{M-step a)} \\ \boldsymbol{\psi}^{(h+1)} &= \arg\max_{\boldsymbol{\psi}} J\left(\mathbf{Y}^{\mathrm{o}},\mathbf{R};\ \boldsymbol{\tau}^{(h)},\boldsymbol{\nu}^{(h)},\boldsymbol{\psi},\boldsymbol{\theta}^{(h+1)}\right) && \textbf{M-step b)} \\ \boldsymbol{\tau}^{(h+1)} &= \arg\max_{\boldsymbol{\tau}} J\left(\mathbf{Y}^{\mathrm{o}},\mathbf{R};\ \boldsymbol{\tau},\boldsymbol{\nu}^{(h)},\boldsymbol{\psi}^{(h+1)},\boldsymbol{\theta}^{(h+1)}\right) && \textbf{VE-step a)} \\ \boldsymbol{\nu}^{(h+1)} &= \arg\max_{\boldsymbol{\nu}} J\left(\mathbf{Y}^{\mathrm{o}},\mathbf{R};\boldsymbol{\tau}^{(h+1)},\boldsymbol{\nu},\boldsymbol{\psi}^{(h+1)},\boldsymbol{\theta}^{(h+1)}\right) && \textbf{VE-step b)} \end{aligned}$$

until $\left\|\boldsymbol{\theta}^{(h+1)}-\boldsymbol{\theta}^{(h)}\right\| < \varepsilon$

Interestingly, resolution of the two steps concerned with the optimization of the parameters related with the SBM – that is to say, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ – can be stated almost independently of any further specification of the sampling design.

Proposition 6. *Consider the lower bound $J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R})$ given by (2.9).*

1. *The parameters $\boldsymbol{\theta}=(\boldsymbol{\alpha},\boldsymbol{\pi})$ maximizing (2.9) when all others are held fixed are*

$$\hat{\alpha}_q = \frac{1}{n}\sum_{i\in\mathcal{N}}\hat{\tau}_{iq}, \qquad \hat{\pi}_{q\ell} = \frac{\sum_{(i,j)\in\mathcal{D}^{\mathrm{o}}}\hat{\tau}_{iq}\hat{\tau}_{j\ell}Y_{ij} + \sum_{(i,j)\in\mathcal{D}^{\mathrm{m}}}\hat{\tau}_{iq}\hat{\tau}_{j\ell}\hat{\nu}_{ij}}{\sum_{(i,j)\in\mathcal{D}}\hat{\tau}_{iq}\hat{\tau}_{j\ell}}.$$

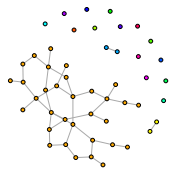
2. *The optimal $\boldsymbol{\tau}$ in (2.9) when all other parameters are held fixed verifies*

$$\hat{\tau}_{iq} \propto \lambda_{iq}\alpha_q\left(\prod_{(i,j)\in\mathcal{D}^{\mathrm{o}}}\prod_{\ell\in\mathcal{Q}} b(Y_{ij};\pi_{q\ell})^{\hat{\tau}_{j\ell}}\right)\left(\prod_{(i,j)\in\mathcal{D}^{\mathrm{m}}}\prod_{\ell\in\mathcal{Q}} b(\nu_{ij};\pi_{q\ell})^{\hat{\tau}_{j\ell}}\right)$$

with λ_{iq} a simple constant depending on the sampling design.

Proof. These results are simply obtained by differentiation of (2.9). □

The two steps concerned with $\boldsymbol{\psi}$ and $\boldsymbol{\nu}$ are specific to the sampling designs used to describe $\mathbf{R}$. Further details are provided below for the designs presented in Section 2.3.



3.3 NMAR: specificities related to the choice of the sampling

In light of Figure 2.1, NMAR conditions specified by DAGs $(b), (c)$ or (d) induce different simplifications for the conditional distribution of the sampling design R:

DAG (b) $p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{Z}) = p_{\psi}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}})$,

DAG (c) $p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{Z}) = p_{\psi}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{Z})$,

DAG (d) $p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{Z}) = p_{\psi}(\mathbf{R}|\mathbf{Z})$.

This induces different evaluations of $\mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}}\left[\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{R}, \mathbf{Y}^{\mathrm{m}}, \mathbf{Z})\right]$ in the lower bound (2.9) for double standard sampling, star degree sampling and class sampling. We obtain below explicit formulas of $\boldsymbol{\psi}$ and $\boldsymbol{\nu}$ by differentiation of the corresponding variational lower bounds. The computations are tedious but straightforward and thus eluded in the following.

Double-standard sampling. Let $s^{\mathrm{m}} = \sum_{(i,j)\in\mathcal{D}^{\mathrm{m}}} \nu_{ij}$, $\bar{s}^{\mathrm{m}} = \sum_{(i,j)\in\mathcal{D}^{\mathrm{m}}} (1-\nu_{ij})$ be the variational counterparts of S^{m} and $\bar{S}^{\mathrm{m}}$. From (2.3) we have

$$\mathbb{E}_{\tilde{p}} \log p_{\psi}(\mathbf{R}|\mathbf{Y}) = S^{\mathrm{o}} \log \rho_1 + \bar{S}^{\mathrm{o}} \log \rho_0 + s^{\mathrm{m}} \log(1-\rho_1) + \bar{s}^{\mathrm{m}} \log(1-\rho_0).$$

Proposition 7 (double standard sampling)**.**

1. *The parameters $\boldsymbol{\psi} = (\rho_0, \rho_1)$ maximizing (2.9) when all others are held fixed are*

$$\hat{\rho}_0 = \frac{\bar{S}^{\mathrm{o}}}{\bar{S}^{\mathrm{o}} + \bar{s}^{\mathrm{m}}}, \qquad \hat{\rho}_1 = \frac{S^{\mathrm{o}}}{S^{\mathrm{o}} + s^{\mathrm{m}}}. \tag{2.10}$$

2. *The optimal $\boldsymbol{\nu}$ in (2.9) when all other parameters are held fixed are*

$$\hat{\nu}_{ij} = \mathrm{logistic}\left(\log\left(\frac{1-\rho_1}{1-\rho_0}\right) + \sum_{(q,\ell)\in\mathcal{Q}^2} \tau_{iq}\tau_{j\ell} \log\left(\frac{\pi_{q\ell}}{1-\pi_{q\ell}}\right)\right).$$

Moreover, $\lambda_{iq} = 1\ \forall (i,q) \in \mathcal{N} \times \mathcal{Q}$ for optimization of $\boldsymbol{\tau}$ in Proposition 6.b).

Class sampling. According to (2.5) we have

$$\mathbb{E}_{\tilde{p}} \log p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}) = \sum_{i\in\mathcal{N}^{\mathrm{o}}} \sum_{q\in\mathcal{Q}} \tau_{iq} \log(\rho_q) + \sum_{i\in\mathcal{N}^{\mathrm{m}}} \sum_{q\in\mathcal{Q}} \tau_{iq} \log(1-\rho_q).$$

Proposition 8 (class sampling)**.**

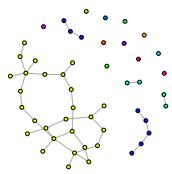
1. *The parameters $\boldsymbol{\psi} = (\rho_1 ... \rho_Q)$ maximizing (2.9) when all others are held fixed are*

$$\hat{\rho}_q = \frac{\sum_{i\in\mathcal{N}^{\mathrm{o}}} \tau_{iq}}{\sum_{i\in\mathcal{N}} \tau_{iq}}. \tag{2.11}$$

2. *The optimal $\boldsymbol{\nu}$ in (2.9) when all other parameters are held fixed verify*

$$\hat{\nu}_{ij} = \mathrm{logistic}\left(\sum_{(q,\ell)\in\mathcal{Q}^2} \tau_{iq}\tau_{j\ell} \log\left(\frac{\pi_{q\ell}}{1-\pi_{q\ell}}\right)\right).$$

Moreover $\lambda_{iq} = \rho_q^{\mathbb{1}\{i\in\mathcal{N}^{\mathrm{o}}\}}(1-\rho_q)^{\mathbb{1}\{i\in\mathcal{N}^{\mathrm{m}}\}}$ for optimization of $\boldsymbol{\tau}$ in Proposition 6.b).



Star degree sampling. From Expression (2.4) of the likelihood, one has

$$\mathbb{E}_{\tilde{p}} \log p_{\psi}(\mathbf{R}|\mathbf{Y}) = -\sum_{i\in\mathcal{N}^{\mathrm{m}}} \left(a + b\tilde{D}_i\right) + \sum_{i\in\mathcal{N}} \mathbb{E}_{\tilde{p}}\left[-\log(1+e^{-(a+bD_i)})\right],$$

where $\tilde{D}_i = \mathbb{E}_{\tilde{p}}\left[D_i\right] = \sum_{i\in\mathcal{N}^{\mathrm{m}}} \nu_{ij} + \sum_{i\in\mathcal{N}^{\mathrm{o}}} Y_{ij}$ is the approximation of the degrees. Because $\mathbb{E}_{\tilde{p}}\left[-\log(1+e^{-(a+bD_i)})\right]$ has no explicit form, an additional variational approximation is needed (Jordan et al., 1998). This technique was recently used in random graph framework (Latouche et al., 2018). It relies on the following approximation of the logistic function:

$$g(x) \geq g(\zeta) + \frac{x-\zeta}{2} + h(\zeta)(x^2-\zeta^2), \quad h(\zeta) = \frac{-1}{2\zeta}\left[\mathrm{logistic}(\zeta) - \frac{1}{2}\right] \tag{2.12}$$

for all $(x,\zeta) \in \mathbb{R}\times\mathbb{R}^+$. This leads to a lower bound of the initial lower bound:

$$\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R}) \geq J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R}) \geq J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{R}), \tag{2.13}$$

with $\boldsymbol{\zeta} = (\zeta_i, i\in\mathcal{N})$ such that $\zeta_i > 0$ is an additional set of variational parameters used to approximate $-\log(1+e^{-x})$. The second lower bound $J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}}$ is derived from Equation (2.12) and given in the supplementary materials for completeness . At the end of the day, we have an additional set of variational parameters to optimize, and a corresponding additional step in Algorithm 2. Expression of all the parameters specific to star degree sampling by differentiating $J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}}$.

Proposition 9 (star degree sampling)**.** *Let* $\widetilde{D_i^2} = \mathbb{E}_{\tilde{p}}\left[{D_i}^2\right]$ *and* $\tilde{D}_k^{-\ell} = \tilde{D}_k - \nu_{k\ell}$.

1. *The parameters* $\boldsymbol{\psi} = (a,b)$ *maximizing* $J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},R)$ *when others are held fixed are*

$$\hat{b} = \frac{2\left(\frac{n}{2} - \mathrm{card}\left(\mathcal{N}^{\mathrm{m}}\right)\right)\sum_{i\in\mathcal{N}}(h(\zeta_i)\tilde{D}_i) - \left(\frac{1}{2}\sum_{i\in\mathcal{N}}\tilde{D}_i - \sum_{i\in\mathcal{N}^{\mathrm{m}}}\tilde{D}_i\right)\times\sum_{i\in\mathcal{N}}h(\zeta_i)}{2\sum_{i\in\mathcal{N}}(h(\zeta_i)\widetilde{D_i^2})\times\sum_{i\in\mathcal{N}}h(\zeta_i) - \left(2\sum_{i\in\mathcal{N}}h(\zeta_i)\tilde{D}_i\right)^2},$$

$$\hat{a} = -\frac{\hat{b}\sum_{i\in\mathcal{N}}\left(h(\zeta_i)\tilde{D}_i\right) + \frac{n}{2} - \mathrm{card}\left(\mathcal{N}^{\mathrm{m}}\right)}{\sum_{i\in\mathcal{N}}h(\zeta_i)}.$$

2. *The parameters* $\boldsymbol{\zeta}$ *maximizing* $J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},R)$ *when others are held fixed are*

$$\hat{\zeta}_i = \sqrt{a^2 + b^2\widetilde{D_i^2} + 2ab\tilde{D}_i}, \ \forall i\in\mathcal{N}.$$

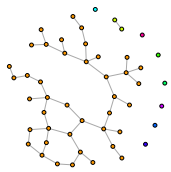
3. *The optimal* $\boldsymbol{\nu}$ *in* $J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},R)$ *when all other parameters are held fixed verify*

$$\hat{\nu}_{ij} = \mathrm{logistic}\Bigg(\sum_{(q,\ell)\in\mathcal{Q}^2}\tau_{iq}\tau_{j\ell}\log\left(\frac{\pi_{q\ell}}{1-\pi_{q\ell}}\right) - b + 2h(\zeta_i)\left(ab + b^2(1+\tilde{D}_i^{-j})\right) + 2h(\zeta_j)\left(ab + b^2(1+\tilde{D}_j^{-i})\right)\Bigg). \tag{2.14}$$

Moreover, $\lambda_{iq} = 1\ \forall(i,q)\in\mathcal{N}\times\mathcal{Q}$ *for optimization of* $\boldsymbol{\tau}$ *in Proposition 6.b).*

Model selection. In NMAR cases, ICL can be useful not only to select the appropriate number of blocks but also for selecting the most appropriate sampling design when it is unknown. Contrary to the MAR case, ICL is no longer a straightforward generalization of Daudin et al. (2008). Indeed, the complete likelihood and thus the penalization needs to take into account the sampling design. Let us consider a model with Q blocks and a sampling design with K parameters (*i.e.* the dimension of $\boldsymbol{\psi}$). The ICL criterion is a Laplace approximation of the complete likelihood $p(\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{R},\mathbf{Z}|Q,K)$ with $p(\boldsymbol{\theta},\boldsymbol{\psi}|Q,K)$ the prior distributions on the parameters such that

$$p(\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{R},\mathbf{Z}|Q,K) = \int_{\Theta\times\Psi} p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}},\mathbf{Y}^{\mathrm{m}},\mathbf{R},\mathbf{Z}|Q,K)p(\boldsymbol{\theta},\boldsymbol{\psi}|Q,K)d\boldsymbol{\theta}d\boldsymbol{\psi}.$$



Proposition 10. *For a model with Q blocks, a sampling design with a vector of parameters $\boldsymbol{\psi} \in \mathbb{R}^K$ and $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\psi}}) = \arg\max_{(\boldsymbol{\theta},\boldsymbol{\psi})} \log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{R}, \mathbf{Z})$, then*

$$\mathrm{ICL}(Q) = -2\mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}};\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\psi}}}\left[\log p_{\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\psi}}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, \mathbf{R}, \mathbf{Z}|Q, K)\right] + \mathrm{pen}_{ICL}(Q),$$

$$\mathrm{pen}_{ICL} = \begin{cases} \left(K + \frac{Q(Q+1)}{2}\right)\log\left(\frac{n(n-1)}{2}\right) + (Q-1)\log(n) & \textit{for dyad-centered sampling} \\ \frac{Q(Q+1)}{2}\log\left(\frac{n(n-1)}{2}\right) + (K+Q-1)\log(n) & \textit{for node-centered sampling} \end{cases}$$

Note that an ICL criterion for MAR sampling designs can be constructed in the same fashion for the purpose of comparison with NMAR sampling designs.

4 Simulation study

In this section, we illustrate the relevance of our approaches on network data simulated under the SBM and sampled under MAR and NMAR conditions. The quality of the inference is assessed by computing the distance between the estimated and the true connectivity matrices $\boldsymbol{\pi}$ in terms of Frobenius norm. The quality of the clustering recovery is measured with the adjusted Rand index (ARI, Rand, 1971) between the true classification and the clustering obtained by maximum posterior probabilities for each $\boldsymbol{\tau}_i$.

4.1 MAR condition

Algorithm 1 for MAR samplings is tested on affiliation networks with 3 blocks. The number of blocks is assumed to be known. For this topology the probability of connection within a block is η and is ten times stronger than the probability of connections between nodes from different blocks. We generate networks with $n = 200$ nodes and marginal probabilities of belonging to blocks $\boldsymbol{\alpha} = (1/3, 1/3, 1/3)$. The sampling design is chosen as a random-dyad sampling with a varying ρ. The difficulty is controlled by two parameters: the sampling effort ρ and the overall connectivity in matrix $\boldsymbol{\pi}$, defined by $c = \sum_{q\ell} \alpha_q \alpha_\ell \pi_{q\ell}$, which is directly related to the choice of η: the lower the η, the sparser the network and the harder the inference. The simulation is repeated 500 times for each configuration (c, ρ). Figure 2.2 displays the results in terms of estimation of $\boldsymbol{\pi}$ and of classification recovery, for varying connectivity c and sampling effort ρ. Our method achieves good performances even with a low sampling effort provided that the connectivity is not too low.

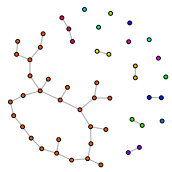
4.2 NMAR condition

Under NMAR conditions we conduct an extensive simulation study by considering various network topologies (namely affiliation, star and bipartite), the connectivity matrix of which are given in Figure 2.3. We use a common tuning parameter ϵ to control the connectivity of the networks in each topology: the lower the ϵ, the more contrasted the topology.

Among the three schemes developed in Section 2.3, we investigate thoroughly the double standard sampling, for which we exhibit a large panel of situations where the gap is large between the performances of the algorithms designed for MAR or NMAR cases. Other sampling designs are explored in the supplementary materials.

Simulated networks have $n = 100$ nodes, with ϵ varying in $\{0.05, 0.15, 0.25\}$. Prior probabilities α are chosen specifically for affiliation, star and bipartite topologies, respectively $(1/3, 1/3, 1/3)$, $(1/6, 1/3, 1/6, 1/3)$ and $(1/4, 1/4, 1/4, 1/4)$. The exploration of the sampling parameters $\psi = (\rho_0, \rho_1)$ is done on a grid $[0.1, 0.9] \times [0.1, 0.9]$ discretized by steps of 0.1. Algorithm 2 is initialized with several random initializations and spectral clustering.

In Figure 2.4, the estimation error is represented as a function of the difference between the sampling design parameters (ρ_0, ρ_1): the closer this difference to zero, the closer to the MAR case. As expected, Algorithm 1 designed for MAR only performs well when



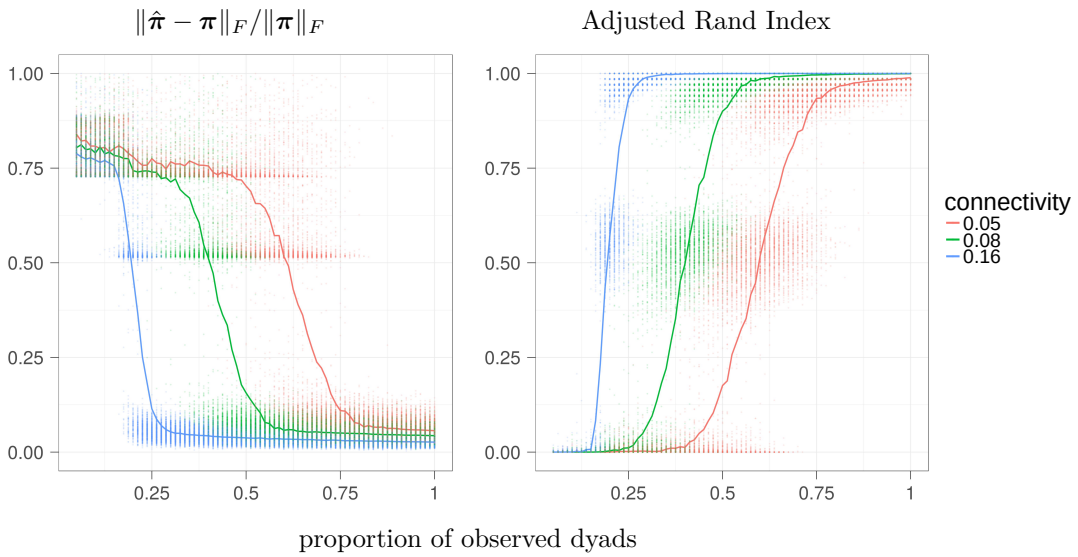
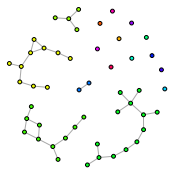


Figure 2.2 – Estimation error of π and Adjusted Rand Index averaged over 500 simulations in the MAR setting. The adjacency matrix $\mathbf{Y}$ is generated under random-dyad sampling strategy for various connectivity $c = \sum_{q\ell} \alpha_q \alpha_\ell \pi_{q\ell}$.

$$\begin{pmatrix} 1-\epsilon & \epsilon & \epsilon \\ \epsilon & 1-\epsilon & \epsilon \\ \epsilon & \epsilon & 1-\epsilon \end{pmatrix} \qquad \begin{pmatrix} 1-\epsilon & 1-\epsilon & 0 & 0 \\ 1-\epsilon & 0 & \epsilon & 0 \\ 0 & \epsilon & 1-\epsilon & 1-\epsilon \\ 0 & 0 & 1-\epsilon & 0 \end{pmatrix} \qquad \begin{pmatrix} \epsilon & 1-\epsilon & \epsilon & \epsilon \\ 1-\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-\epsilon \\ \epsilon & \epsilon & 1-\epsilon & \epsilon \end{pmatrix}$$

(a) affiliation (b) star (c) bipartite

Figure 2.3 – Matrix π in different topologies with inter/intra block probabilities.



$\rho_1 - \rho_0 \approx 0$. Algorithm 2 designed for NMAR double-standard sampling shows relatively flat curves which means that its performances are roughly constant no matter the sampling condition. Figure 2.5 reports estimation accuracy for the sampling parameters ρ_0 and

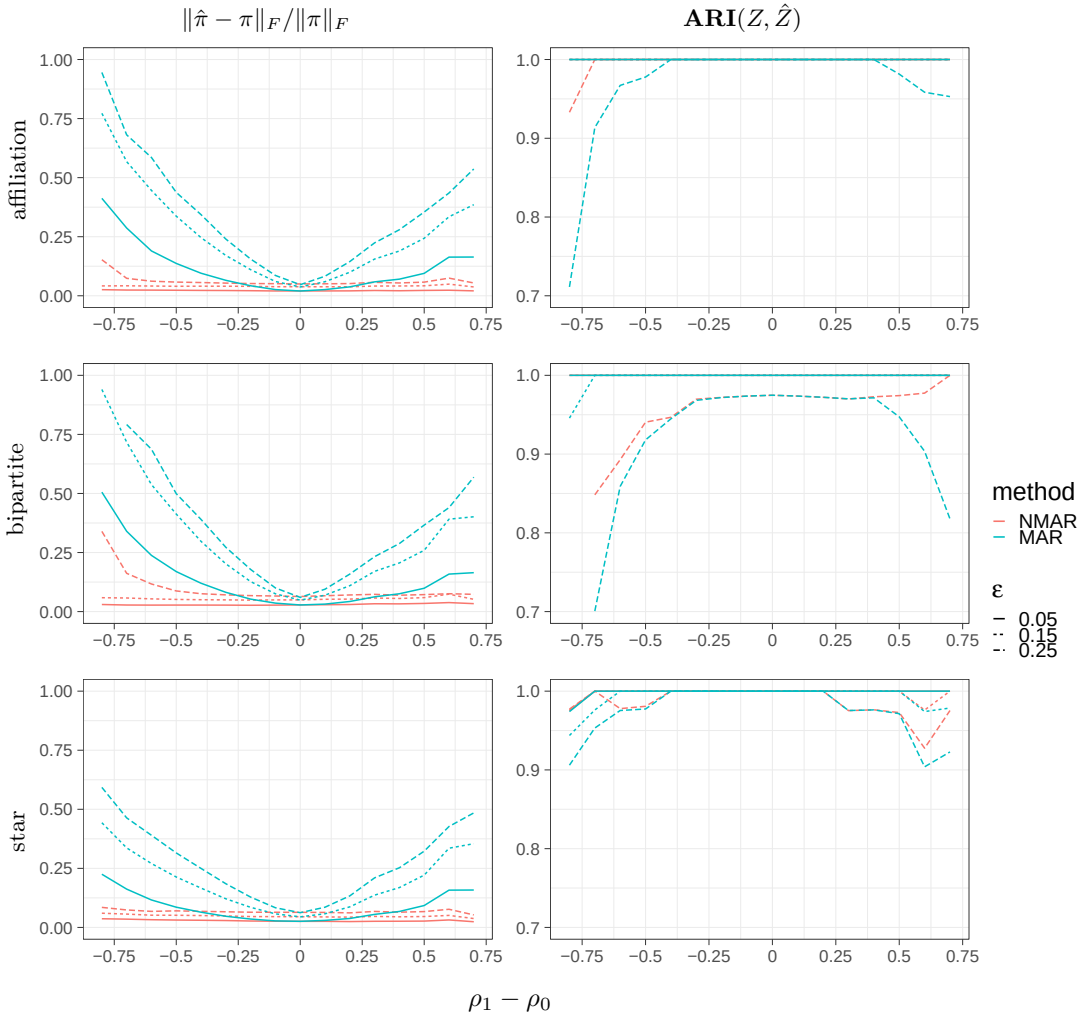


Figure 2.4 – Double standard setting: estimation error of π and adjusted Rand index averaged over 500 simulations for affiliation, bipartite and star topologies.

ρ_1. Results show a good ability of the VEM to estimate these parameters. As expected, performances deteriorate for uncontrasted topologies with low sampling rate.

Model selection. Simulations are also conducted to study the performances of ICL. We compare results for the different topologies described in Figure 2.3 for $\epsilon = 0.05$. Rates of correct answers for selecting the number of blocks Q under a double standard sampling with different sampling rates are displayed in Table 2.1. The ARI is also provided. The ICL shows a satisfactory ability to select the true Q even if the selection task obviously needs a larger sampling effort than the estimation task. It is worth mentioning that a whole block may not be sampled, which leads the ICL to select a lower number of blocks. In such a case the ARI is a meaningful additional information to demonstrate that the clustering remains coherent with the true one.

In Table 2.2, results concern the rates of correct selections of the sampling design when the two designs in competition are the random-dyad and the double standard samplings.

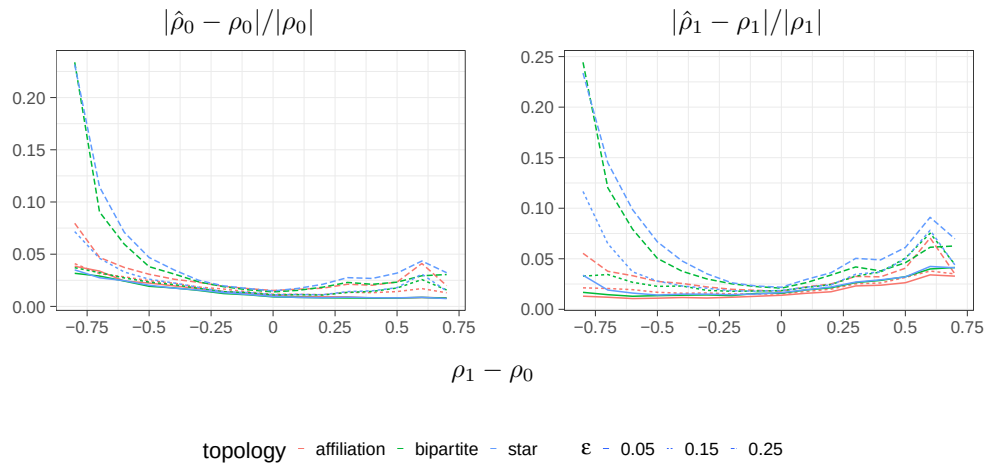
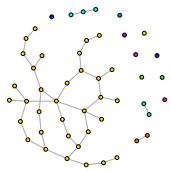


Figure 2.5 – Double standard setting: estimation error of ρ_0 and ρ_1 averaged over 500 simulations for affiliation, bipartite and star topologies.

sampling rate	affiliation	bipartite	star
(0.154, 0.405]	0.58/0.96	0.46/0.84	0.45/0.84
(0.405, 0.656]	0.95/0.99	0.87/0.98	0.90/0.98
(0.656, 0.908]	1/1	0.99/1	0.99/1

Table 2.1 – Performance of the ICL criterion: rates of correct answers when choosing the number of blocks and Adjusted Rand Indexes. Tested configurations are different sampling rates and three topologies (affiliation, bipartite and star) under a double standard sampling. Each configuration is simulated 500 times.



As expected, the rate of correct answers increases with the sampling rate.

sampling rate	sampling	affiliation	bipartite	star
(0.096, 0.367]	MAR	0.73	0.67	0.63
	NMAR	0.72	0.75	0.75
(0.367, 0.638]	MAR	1	1	1
	NMAR	0.91	0.78	0.82
(0.638, 0.909]	MAR	1	1	1
	NMAR	0.91	0.8	0.95

Table 2.2 – Rates of correct answers of the ICL criterion when choosing between Random-dyad sampling (MAR) and double standard sampling (NMAR) in each of the 18 configurations. A configuration is the combination of a topology (affiliation, bipartite and star), a sampling rate and a sampling design. Each configuration is simulated 500 times.

5 Importance of accouting for missing values in real networks

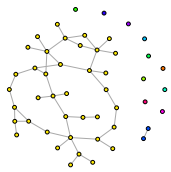
5.1 Seed exchange network in the region of Mount Kenya

In a context of subsistence farming, studies which investigate the relationships between crop genetic diversity and human cultural diversity patterns have shown that seed exchanges are embedded in farmers' social organization. Data on seed exchanges of sorghum in the region of Mount Kenya were collected and analyzed in Labeyrie et al. (2016, 2014). The sampling is node-centered since the exchanges are documented by interviewing farmers who are asked to declare to whom they gave seeds and from whom they receive seeds. Since an interview is time consuming, the sampling is not exhaustive. A limited space area was defined where all the farmers were interviewed. The network is thus collected with missing dyads since information on the potential links between two farmers who were cited but not interviewed is missing. With the courtesy of Vanesse Labeyrie, we analyzed the Mount Kenya seed exchange network involving 568 farmers among which 155 were interviewed. Although other farmers in this region might be connected to non-interviewed farmers, we focus on this closed network of 568 nodes.

Since we only know that the sampling is node-centered, we fit SBM under the three node-centered sampling designs presented in Section 2.2 (star (MAR), class and star degree sampling). The ICL criterion is minimal for 10 blocks under the star degree sampling and for 11 blocks under the class degree sampling. The clusterings between the SBMs obtained with either class or star degree sampling remain close from each other (ARI: 0.6) and both unravel a strong community structure. The model selected by ICL for MAR sampling is composed by 11 blocks. The ARIs between MAR clustering and the two other clusterings are lower (around 0.4). Finally, note that interviewed and non-interviewed farmers are mixed up in the blocks of the three selected models. The ICL criteria computed for the three sampling designs are a slightly in favor of the MAR sampling.

On top of network data, categorical variables are available for discriminating the farmers such as the ntora[2] they belong to (10 main ntoras plus 1 grouping all the others) and the dialect they speak (4 dialects). In Figure 2.6, we compute ARIs between the ntoras (left panel), the dialects (right panel) and the clusterings obtained with the SBM under the three node-centered sampling designs for a varying number of blocks. Even though the ARIs remain low, the clusterings from class or star degree sampling seem to catch a non negligible fraction of the social organization, larger than the one caught by the clustering

[2]The ntora is a small village or a group of neighborhoods



from the MAR sampling. These two categorical variables, reflecting some aspects of the social organization, could partially explain the structure of the exchange network.

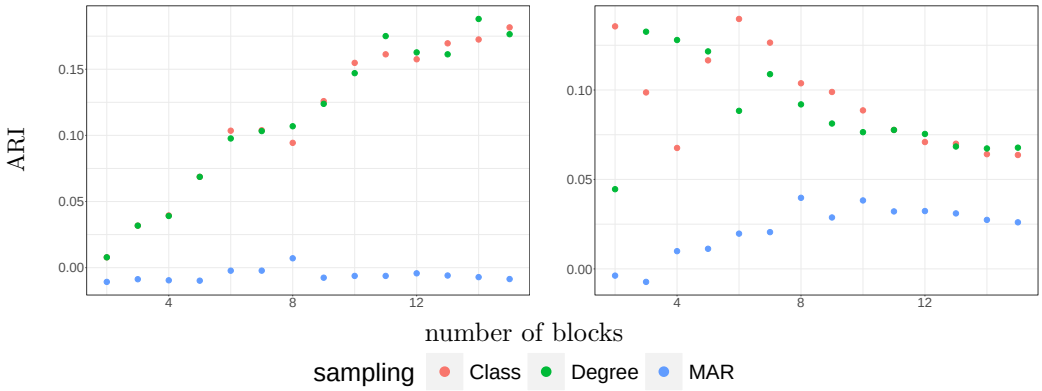


Figure 2.6 – ARIs computed between the clusterings given by an SBM under class, star degree and MAR samplings with a varying number of blocks Q and ntora of farmers (left-hand-side) or dialect spoken by farmers (right-hand-side)

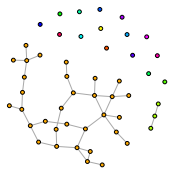
5.2 ER (ESR1) Protein-Protein Interaction network in breast cancer

Estrogen receptor 1 (ESR1) is a gene that encodes an estrogen receptor protein (ER), a central actor in breast cancer. Uncovering its relations with other proteins is essential for a better understanding of the disease. To this end, various bioinformatics tools are available to centralize knowledge about possible relations between proteins into networks known as *Protein-Protein Interaction* (PPI) networks. The platform `string` (Szklarczyk et al., 2015) accessible via `http://www.string-db.org` is one of the most popular tools for this task. Given a set of one (or several) initial protein(s) provided by the user, it is possible to recover a valued network between all proteins connected to the initial set. The value of an edge in this network corresponds to a score obtained by aggregating different types of knowledge (wet-lab experiments, textmining, co-expression data, etc. . .), reflecting a level of confidence. Thus, it is possible for a given protein – we choose ER here – to recover the PPI network between all proteins involved. Our ambition is to rely on a SBM with missing data to finely analyze such networks: we rather describe a dyad as missing (thus not choosing between 0 or 1) if its level of confidence is too low.

The PPI network in the neighborhood of ER is composed by 741 proteins connected by edges with values in $(0, 1]$. We remove ER from this set of proteins, as well as the zinc finger protein 44. Indeed, they were both connected to most of the other proteins and would thus only blur the underlying clustering structure. We denote ω_{ij} the weight associated with dyad (i, j). By means of a tuning parameter γ reflecting the level of confidence, the adjacency matrix is defined as follows:

$$\mathbf{A}^\gamma = (A^\gamma)_{ij} = \begin{cases} 1 & \text{if } \omega_{ij} > 1-\gamma, \\ \texttt{NA} & \text{if } \gamma \leq \omega_{ij} \leq 1-\gamma, \\ 0 & \text{if } \omega_{ij} < \gamma. \end{cases} \tag{2.15}$$

In order to analyze the ER-centered network, Algorithm 1 (random-dyad MAR sampling) and Algorithm 2 (double-standard NMAR sampling) were applied on $\mathbf{A}^\gamma$ for γ varying in $\{.15, .25, .35\}$, hence taking the uncertainties on the missing dyads into account with various thresholds. The ICL criterion in Figure 2.7 systematically chooses the NMAR modeling against the MAR modeling, whatever the value of γ.



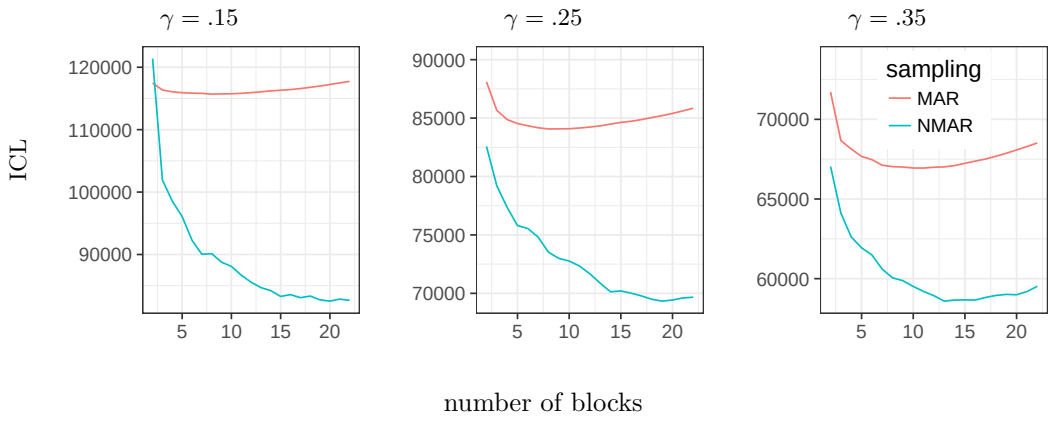


Figure 2.7 – ICL criteria for SBMs with random-dyad MAR sampling and double-standard NMAR sampling in the thresholded ER network.

We study the best MAR and NMAR models associated with $\gamma = 0.35$, which value exhibits a clearer choice of the ICL than for $\gamma = \{.15, .25\}$ for both MAR and NMAR modelings. The two corresponding SBMs have 11 clusters for MAR sampling and 13 clusters for NMAR sampling. The ARI between the two clusterings is around 0.39: this is mainly due to a large block in the random-dyad MAR clustering which contains much more nodes than any of the blocks in the NMAR clustering. The latter dispatches many of these nodes in four blocks (see the supplementary materials for a more detailed exposition of results). To prove that this finest clustering of the nodes is more relevant from the biological point of view, we propose a validation based on external biological knowledge. To this end, we rely on the Gene Ontology (GO) annotation (Ashburner et al., 2000) which provides a DAG of ontologies to which genes are annotated if the proteins encoded by these genes are involved in a known biological process. Here, we use GO to perform enrichment analysis (that is to say identifying classes of genes that are over-represented in a large set of genes, via a simple hypergeometric test) on genes corresponding to the proteins present in the large block for MAR, and the corresponding four blocks for NMAR. Interestingly, at a significance level of 1%, we find a single significant biological process for MAR modeling while 13 are found significant in the NMAR case. We check that it is not due to a simple threshold effect by looking at the ranks of the p-values of the 13 NMAR significant processes in the 100 first most significant terms found in the MAR model: only 5 of the NMAR processes are found, with high ranks (24, 33, 39, 56 and 77) far from the smallest MAR p-values.

6 Conclusion

This paper shows how to deal with missing data on dyads in the SBM. We study MAR and NMAR sampling designs motivated by network data and propose variational approaches to perform inference under this series of designs, accompanied with model selection criteria. Relevance of the method is illustrated on numerical experiments both on simulated and real-world networks. An R-package called missSBM is available on the CRAN (see chapter 5 for more details).

This work focuses on undirected binary networks. However, it can be adapted to other SBMs, in particular those developed in Mariadassou et al. (2010) for (un)directed valued networks with a distribution of weights belonging to the exponential family. It could also be adapted to the degree-corrected SBM (Karrer and Newman, 2011), where the sampling design would depend on the degree correction parameters. This should lead to a design close to the star degree sampling. In future works, we plan to investigate the consistency of

the variational estimators of SBM under missing data conditions, looking for similar results as the ones obtained in Bickel et al. (2013) for fully observed networks. Another path of research is to consider missing data where we cannot distinguish between a missing dyad and the absence of an edge like in Priebe et al. (2015) and Balachandran et al. (2017).

7 Supplementary

7.1 Proof of Proposition 4.*ii*)

Recall the following result from graphical model theory (see Giraud, 2014, Formula 7.1).

Lemma. *Let X, Y, W and Z be random variables, then for all h measurable,*

$$X \perp\!\!\!\perp (Y, W) \mid Z \Rightarrow \; X \perp\!\!\!\perp Y \mid (h(W), Z).$$

Applying this lemma with h the identity, function, we get $\mathbf{R} \perp\!\!\!\perp (\mathbf{Y}^{\mathrm{m}}, \mathbf{Z}) \mid \mathbf{Y}^{\mathrm{o}} \Rightarrow \mathbf{R} \perp\!\!\!\perp \mathbf{Z} \mid (\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}})$, and then $\mathbf{R} \perp\!\!\!\perp \mathbf{Z} \mid (\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}})$ implies $\mathbf{R} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{Y}$.

7.2 Identifiability

When the sampling is node-centered, we denote $V_i = 1$ if node i is observed and $V_i = 0$ otherwise.

Proof of Proposition 5. Let $\rho, \rho' > 0$ be such that $p_\rho(\mathbf{R}) = p_{\rho'}(\mathbf{R})$ (resp. $p_\rho(\mathbf{V}) = p_{\rho'}(\mathbf{V})$). Since $\mathbf{R}$ (resp.$\mathbf{V}$) does not depend on $\mathbf{Y}$, then $\mathbb{P}_\rho(R_{ij} = 1) = \rho = \rho' = \mathbb{P}_{\rho'}(R_{ij} = 1)$ (resp. $\mathbb{P}_\rho(V_i = 1) = \rho = \rho' = \mathbb{P}_{\rho'}(V_i = 1)$). □

Proof of Theorem 1. Let $P^o_{[n]}$ denote the probability distribution function of $\mathbf{Y}^o$. We show that there exists a unique $(\boldsymbol{\alpha}, \boldsymbol{\pi})$ corresponding to $P^o_{[n]}$.

Define $s_q = \mathbb{P}(Y_{ij}R_{ij} = 1 | Z_{iq} = 1) = \rho(\boldsymbol{\pi}\ \boldsymbol{\alpha})_q$ (resp. $s_q = \mathbb{P}(Y_{ij} = 1 | Z_{iq} = 1) = (\boldsymbol{\pi}\ \boldsymbol{\alpha})_q$ for star sampling). Up to reordering, $s_1 < s_2 < \ldots < s_Q$ are the coordinates of the vector s in the increasing order. Let S denote the Vandermonde matrix defined by $S_{i,q} = s_q^i$, for $0 \le i < Q$ and $1 \le q \le Q$. S is invertible since the coordinates of s are all different. For $i \ge 1$, $S_{i,q} = \mathbb{P}(Y_{12}R_{12} = 1, \ldots, Y_{1i+1}R_{1i+1} = 1 | Z_{1q} = 1)$ for random-dyad sampling (resp. $S_{i,q} = \mathbb{P}(Y_{12} = 1, \ldots, Y_{1i+1} = 1 | Z_{1q} = 1)$ for star degree sampling). Let us also define

$$u_i = \sum_{1 \le k \le Q} \alpha_k s_k^i \quad (\text{resp. } u_i = \sum_{1 \le k \le Q} \rho \alpha_k s_k^i), \quad i = 0, \ldots, 2Q - 1 \ .$$

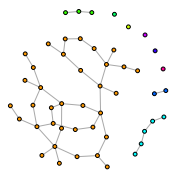
For $i \ge 1$, $u_i = \mathbb{P}(Y_{12}R_{12} = 1, \ldots, Y_{1i+1}R_{1i+1} = 1)$ (resp. $u_i = \mathbb{P}(Y_{12} = 1, \ldots, Y_{1i+1} = 1, V_1 = 1)$). Note that $n \ge 2Q$ is a necessary requirement on n since $Y_{i,i} = 0$ by assumption. Hence, given $P^o_{[n]}$ and ρ, $u_0 = 1$ and $u_1, \ldots, u_{2Q-1}$ are known.

Furthermore, set M the $(Q+1) \times Q$ matrix given by $M_{i,j} = u_{i+j}$ for every $0 \le i \le Q$ and $0 \le j < Q$, and let M_i denote the square matrix obtained by removing the row i from M. The coefficients of M_Q, for $0 \le i, j < Q$, are

$$M_{i,j} = \sum_{1 \le k \le Q} s_k^i \alpha_k s_k^j \quad (\text{resp. } M_{i,j} = \sum_{1 \le k \le Q} \rho s_k^i \alpha_k s_k^j) \ , \text{ with } 0 \le i, j < Q \ .$$

Defining the diagonal matrix $A = \mathrm{Diag}(\boldsymbol{\alpha})$, it comes that $M_Q = SAS^t$ (resp. $M_Q = \rho SAS^t$), where S and A are invertible, but unknown at this stage, and $\rho > 0$. With $D_k = \det(M_k)$ and the polynomial $B(x) = \sum_{k=0}^{Q} (-1)^{k+Q} D_k\, x^k$, it yields $D_Q = \det(M_Q) \neq 0$ and the degree of B is equal to Q.

Set $C_i = (1, s_i, \ldots, s_i^Q)^t$ and let us notice that $B(s_i)$ is the determinant of the square matrix produced when appending C_i as the last column to M. The $Q+1$ columns of this matrix are linearly dependent, since they are all linear combinations of the Q vectors C_1, C_2, …, C_Q. Hence $B(s_i) = 0$ and s_i is a root of B for every $1 \le i \le Q$. This proves that $B = D_Q \prod_{i=1}^{Q} (x - s_i)$. Then, one knows that $s = (s_1, \ldots, s_Q)$ (as the roots of B defined from M) and S. It results that $A = S^{-1} M_Q (S^t)^{-1}$, which yields a unique $(\alpha_1, \ldots, \alpha_Q)$ (resp. $A = \rho^{-1} S^{-1} M_Q (S^t)^{-1}$).



It only remains to determine π. For $0 \leq i, j < Q$, let us introduce $U_{i,j}$ the probability that the first row of Y^o begins with $i+1$ occurrences of 1, and the second row of Y^o ends up with j occurrences of 1 ($i+1+j \leq n-1$ implies $n \geq 2Q$). Then, $U_{i,j} = \sum_{k,l} S_{i,k}\alpha_k\pi_{k,l}\alpha_l S_{j,l}$ (resp. $U_{i,j} = \sum_{k,l} \rho^2 S_{i,k}\alpha_k\pi_{k,l}\alpha_l S_{j,l}$), for $0 \leq i, j < Q$, and the $Q \times Q$ matrix $U = SA\boldsymbol{\pi}AS^t$. The conclusion results from $\boldsymbol{\pi} = A^{-1}S^{-1}U(S^t)^{-1}A^{-1}$ (resp. $\boldsymbol{\pi} = \rho^{-2}A^{-1}S^{-1}U(S^t)^{-1}A^{-1}$). $\square$

Proof of Theorem 2. Let $P_{[n]}$ denote the probability distribution function of $(\mathbf{Y}^o, \mathbf{R})$. We show that there exists a unique $(\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho})$ corresponding to $P_{[n]}$.

Identifiability of $\boldsymbol{\alpha}$. Up to reordering, let $t_1 < t_2 < \ldots < t_Q$ denote the coordinates of the vector t in the increasing order, we have : $t_q = \mathbb{P}(Y_{ij} = 1, V_j = 1 | Z_{iq} = 1)$.

Let T denote the Vandermonde matrix defined by $T_{i,q} = t_q^i$, for $0 \leq i < Q$ and $1 \leq q \leq Q$. T is invertible since the coordinates of t are all different. For $i \geq 1$, $T_{i,q} = \mathbb{P}(Y_{12} = 1, \ldots, Y_{1i+1} = 1, V_2 = 1, \ldots, V_{i+1} = 1 | Z_{1q} = 1)$. Let us also define

$$v_i = \sum_{1 \leq k \leq Q} \alpha_k t_k^i, \quad i = 0, \ldots, 2Q-1 \ .$$

For $i \geq 1$, $v_i = \mathbb{P}(Y_{12} = 1, \ldots, Y_{1i+1} = 1, V_2 = 1, \ldots, V_{i+1} = 1)$. Hence given $P_{[n]}$, $v_0 = 1$ and $v_1, \ldots, v_{2Q-1}$ are known. Furthermore, set N the $(Q+1) \times Q$ matrix given by $N_{i,j} = v_{i+j}$ for $0 \leq i \leq j \leq Q$, and let N_i denote the square matrix obtained by removing the row i from N. The coefficients of N_Q are

$$N_{i,j} = v_{i+j} = \sum_{1 \leq k \leq Q} t_k^i \alpha_k t_k^j \ , \quad \text{with} \quad 0 \leq i, j < Q \ .$$

Defining the diagonal matrix $A = \mathrm{Diag}(\boldsymbol{\alpha})$, it comes that $N_Q = TAT^t$, where T and A are invertible. With $D_k = \det(N_k)$ and the polynomial $B(x) = \sum_{k=0}^{Q}(-1)^{k+Q} D_k\, x^k$, it yields $D_Q = \det(N_Q) \neq 0$ and the degree of B is equal to Q.

Set $C_i = (1, t_i, \ldots, t_i^Q)^t$ and let us notice that $B(t_i)$ is the determinant of the square matrix produced when appending C_i as last column to N. The $Q+1$ columns of this matrix are linearly dependent, since they are all linear combinations of the Q vectors C_1, C_2, …, C_Q. Hence $B(t_i) = 0$ and t_i is a root of B for every $1 \leq i \leq Q$. This proves that $B = D_Q \prod_{i=1}^{Q}(x - t_i)$. Then, one knows that $t = (t_1, \ldots, t_Q)$ (as the roots of B defined from N) and T. It results that $A = T^{-1}N_Q(T^t)^{-1}$, which yields a unique $(\alpha_1, \ldots, \alpha_Q)$.

Identifiability of $\boldsymbol{\rho}$. Up to reordering, let $o_1 < o_2 < \ldots < o_Q$ denote the coordinates of the vector o in the increasing order, then $s_q = \mathbb{P}(Y_{ij} = 1, V_i = 1 | Z_{iq} = 1) = \rho_q o_q$. Let O denote the Vandermonde matrix defined by $O_{i,q} = o_q^i$, for $0 \leq i < Q$ and $1 \leq q \leq Q$. O is invertible since the coordinates of o are all different. For $i \geq 1$, $O_{i,q} = \mathbb{P}(Y_{12} = 1, \ldots, Y_{1i+1} = 1 | Z_{1q} = 1)$. Let us also define

$$u_i = \sum_{1 \leq k \leq Q} \rho_k \alpha_k o_k^i, \quad i = 0, \ldots, 2Q-1 \ .$$

For $i \geq 1$, $u_i = \mathbb{P}(Y_{12} = 1, \ldots, Y_{1i+1} = 1, V_1 = 1)$. Hence given $P_{[n]}$, $u_0 = 1$ and $u_1, \ldots, u_{2Q-1}$ are known. Furthermore, set M the $(Q+1) \times Q$ matrix given by $M_{i,j} = u_{i+j}$ for every $0 \leq i \leq Q$ and $0 \leq j < Q$, and let M_i denote the square matrix obtained by removing the row i from M. The coefficients of M_Q are

$$M_{i,j} = u_{i+j} = \sum_{1 \leq k \leq Q} o_k^i \alpha_k \rho_k o_k^j \ , \quad \text{with} \quad 0 \leq i, j < Q \ .$$

Defining the diagonal matrix $B = \mathrm{Diag}(\boldsymbol{\rho})$, it comes that $M_Q = OABO^t$, where O, B and A are invertible. Using the same algebraic argument than for the identifiability of $\boldsymbol{\alpha}$, it results that $B = A^{-1}O^{-1}M_Q(O^t)^{-1}$, which yields, because of the identifiability of $\boldsymbol{\alpha}$, a unique $(\rho_1, \ldots, \rho_Q)$.

Identifiability of $\boldsymbol{\pi}$. For $0 \leq i, j < Q$, let us introduce $U_{i,j}$ the probability that the first row of $\mathbf{Y}^o$ begins with $i+1$ occurrences of 1, and the second row of $\mathbf{Y}^o$ ends up with j occurrences of 1.

$$U_{i,j} = \mathbb{P}\Big(\{Y_{12} = 1, \ldots, Y_{1i+2} = 1, V_2 = 1, \ldots, V_{i+2} = 1\} \bigcap \\ \{Y_{2n-j} = 1, \ldots, Y_{2n} = 1, V_{n-j} = 1, \ldots, V_n = 1\} \Big),$$

Then, $U_{i,j} = \sum_{k,l} T_{i,k}\alpha_k\pi_{k,l}\rho_k\alpha_l T_{j,l}$, for $0 \leq i, j < Q$, and the $Q \times Q$ matrix $U = TA\boldsymbol{\pi}ABT^t$. The conclusion results from $\boldsymbol{\pi} = A^{-1}T^{-1}U(T^t)^{-1}B^{-1}A^{-1}$.

□

7.3 Derivation of second lower bound in star degree sampling

$$J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\zeta},\boldsymbol{\theta},\boldsymbol{\psi}} = C_{miss} + \sum_{i=1}^{n} \mathbb{E}_{\tilde{p}}\left[\text{logistic}(\zeta_i) + \frac{(a+bD_i)-\zeta_i}{2} + h(\zeta_i)((a+bD_i)^2 - \zeta_i^2)\right],$$
$$= C_{miss} + \sum_{i=1}^{n}\left[\text{logistic}(\zeta_i) + \frac{(a+b\tilde{D}_i)-\zeta_i}{2} + h(\zeta_i)(a^2 + 2ab\tilde{D}_i + b^2\hat{D}_i - \zeta_i^2)\right],$$

where $C_{miss} = \sum_{i\in\mathcal{N}^{\mathrm{m}}}\left(a + b\tilde{D}_i\right)$ and $\hat{D}_i = \mathbb{E}_{\tilde{p}}\left[{D_i}^2\right] = \mathbb{V}_{\tilde{p}}\left(D_i\right) + \mathbb{E}_{\tilde{p}}\left[D_i\right]^2 = \sum_{j\in\mathcal{N}^{\mathrm{m}}}\nu_{ij}(1-\nu_{ij}) + \left(\sum_{j\in\mathcal{N}^{\mathrm{m}}}\nu_{ij} + \sum_{j\in\mathcal{N}^{\mathrm{o}}}Y_{ij}\right)^2$.

7.4 Proof of Proposition 10

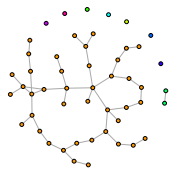
If the sampling design is node-centered (respectively dyad-centered), the first term in (2.7) consists in n Bernoulli random variables (respectively $n(n-1)/2$ Bernoulli random variables). The term $\log p_\psi(R|\mathbf{Y}^{\mathrm{o}}, \mathbf{Y}^{\mathrm{m}}, Z, Q)$ can be derived using a BIC approximation:

$$\log p_{\boldsymbol{\psi}}(\mathbf{R}|\mathbf{Y}, \mathbf{Z}, Q) \simeq \arg\max_{\boldsymbol{\psi}} \log p(\mathbf{R}|\mathbf{Y}, \mathbf{Z}, \psi, Q) - \frac{1}{2}\text{pen}_{\text{BIC}},$$

$$\text{where} \quad \text{pen}_{\text{BIC}} = \begin{cases} K \times \log(n) & \text{if the sampling design is node-centered,} \\ K \times \log\left(\frac{n(n-1)}{2}\right) & \text{otherwise.} \end{cases}$$

7.5 Simulations for star degree and class samplings

With node-centered samplings such as star-degree and class samplings, it is more difficult to find configurations different from MAR sampling designs: with dyad-centred double-standard design, the sampling parameters were directly related to the probability of an edge conditionally on the observation of the corresponding dyad, which is no longer the case for the node-centered designs. The sampling designs being hardly different from a MAR design, the NMAR inference does not show much improvement compared to the MAR inference. Still, we exhibit some interesting situations where star degree and class samplings deserve an appropriate treatment, presented herein. We simulate networks with $n = 100$ nodes under an affiliation topology with intra-community probability (resp. inter community probability) equal to 0.5 (resp. 0.05) and $\boldsymbol{\alpha} = (0.25, 0.5, 0.25)$. Sampling parameters are chosen such that $\boldsymbol{\psi} = (a, b) = (-3.6, 0.1)$ for star degree sampling, which makes nodes with highest degrees preferably selected. In class sampling, these parameters are set to $\boldsymbol{\psi} = (\rho_1, \rho_2, \rho_3) = (0.75, 0.5, 0.05)$, which makes nodes from the largest block and from a small block preferably selected while the other small block is under-sampled. In Figure 2.8, the estimation errors and the ARI are pictured for both cases. The sampling rates (i.e. rates of observed dyads over total number of dyads) lie in the intervals $[0.558, 0.844]$ for



class sampling and $[0.162, 0.622]$ for star degree sampling. These two intervals arise from the values of the parameter $\boldsymbol{\psi}$ explored for these two designs. We compare the performances of Algorithm 2 to an oracle (when inference is conducted via a classical VEM algorithm on a fully observed network) and with Algorithm 1. When facing NMAR condition, Algorithm 2 shows a slight improvement over Algorithm 1 even if it remains far from the oracle.

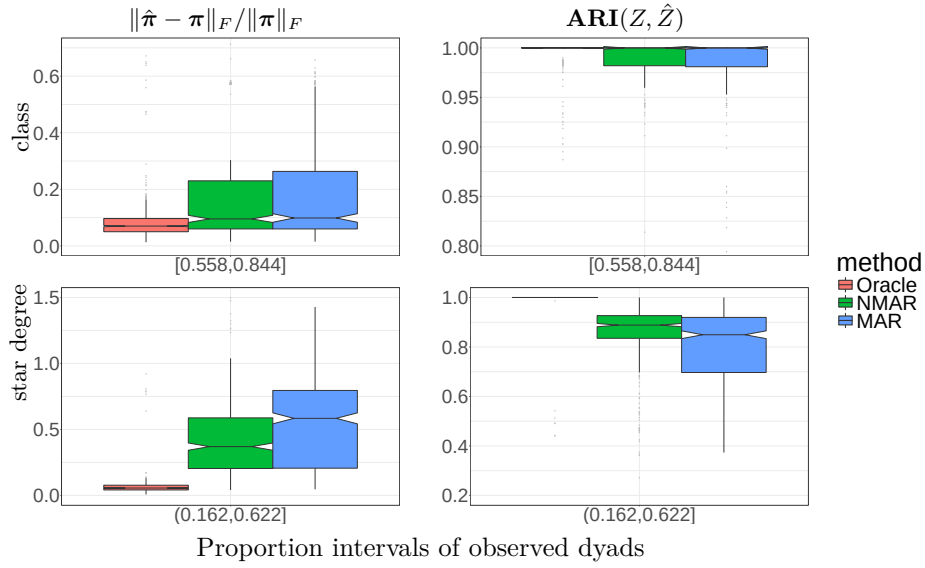
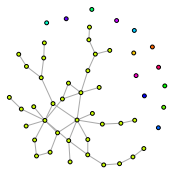
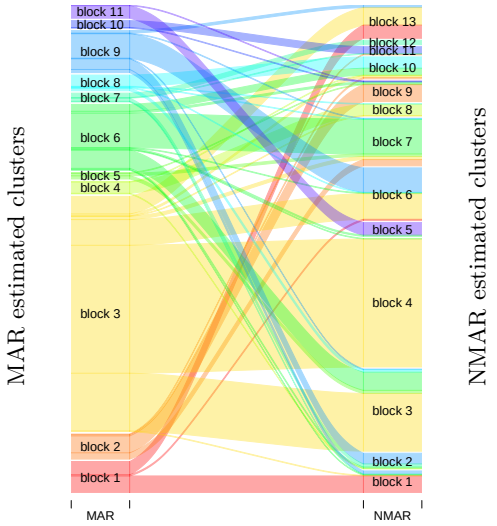


Figure 2.8 – Estimation error of $\boldsymbol{\pi}$ and ARI averaged over 500 simulations in star degree and class settings. The topology is affiliation with $\epsilon = 0.05$.

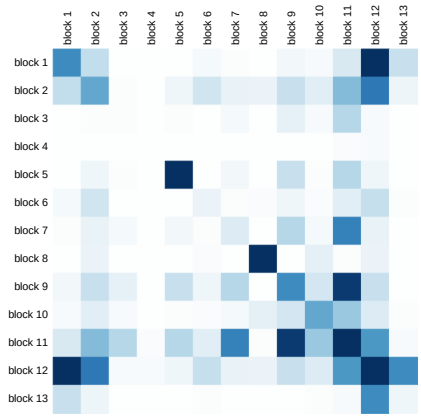
7.6 Additional results for the ER Protein-Protein Interaction network

In this appendix, we provide more details on the NMAR clustering of the ER network discussed in Section 5.2. We represent the estimated connectivity matrix $\hat{\pi}$ in Figure 2.9b for NMAR, which exhibits a network-structure with 13 blocks (or sets of proteins) the sizes of which can be sketched from Figure 2.9a. In Figure 2.9c, we represent on the same matrix the network with the original missing data and the imputed missing dyads with the variational parameters ν. Interestingly, many of the imputed values are close to 1, which might help validating some relationships which were still uncertain in the biological literature. Figure 2.9a shows that the MAR clustering leads to a large cluster (block 3) which is mainly split into 4 blocks in the NMAR clustering (blocks 3, 4, 6 and 13).

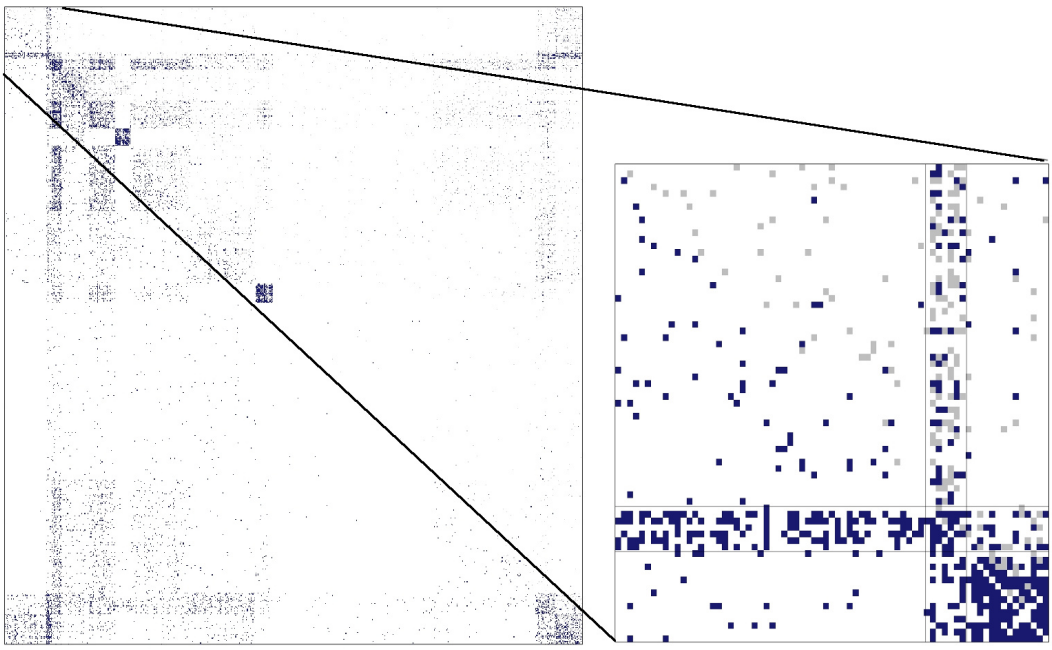




(a) Clustering comparison between MAR and NMAR (11 vs 13 blocks).



(b) Matrix of connectivity $\hat{\boldsymbol{\pi}}$ for NMAR inference (double standard) ; intensity of the color is proportional to the probability of connection between blocks.



(c) ER PPI network reordered by blocks inferred with SBM with NMAR modeling. Left panel: original data with `NA` entries colored in gray (upper triangle) and data imputed with ν_{ij} (lower triangle); right panel: zoom of blocks (1,2,3).

Figure 2.9 – ER PPI network analysis with SBM under missing data conditions

Chapter 3

Consistency and Asymptotic Normality of Stochastic Block Models Estimators from Sampled Data

This chapter is a joint work with Mahendra Mariadassou and is available as a preprint on arXiv (Mariadassou and Tabouy, 2019).

Contents

Statistical analysis of network is an active research area and the literature counts a lot of papers concerned with network models and statistical analysis of networks. However, very few papers deal with missing data in network analysis and we reckon that, in practice, networks are often observed with missing values. In this paper we focus on the Stochastic Block Model with valued edges and consider a MCAR setting by assuming that every dyad (pair of nodes) is sampled identically and independently of the others with probability $\rho > 0$. We prove that maximum likelihood estimators and its variational approximations are consistent and asymptotically normal in the presence of missing data as soon as the sampling probability ρ satisfies $\rho \gg \log(n)/n$.

1 Introduction

For the last decade, statistical network analyses has a been a very active research topic and the statistical modeling of networks has found many applications in social sciences and biology for example Aicher et al. (2014), Barbillon et al. (2015), Mariadassou et al. (2010), Wasserman and Faust (1994) and Zachary (1977).

Many random graphs models have been widely studied, either from a theoretical or an empirical point of view. The first model studied was Erdős-Rényi model (Erdős and Renyi, 1959) which assumes that each pair of nodes (dyad) is connected independently to the others with the same probability. This model assumes homogeneity of all nodes across the network. In order to alleviate this constraint, many families of models have been introduced. Most are endowed with a latent structure (reviewed in Matias and Robin, 2014) to capture heterogeneity across nodes. Among those, the Stochastic Block Model (in short SBM, see Frank and Harary, 1982 and Holland et al., 1983) is one of the oldest and most studied as it is highly flexible and can capture a large variety of structures (affiliation, hub, bipartite and many other). In order to estimate this model, Bayesian approaches were first proposed (Snijders and Nowicki, 1997 and Nowicki and Snijders, 2001) but have been superseded by variational methods (Daudin et al., 2008 and Latouche et al., 2012). The former class of approaches are exact but lack the computational efficiency and scalability that the latter offers.

Theoretical guarantees concerning maximum likelihood estimators (in short MLE) and variational methods for the binary SBM estimation is not an easy task and have been widely studied. In Celisse et al. (2012), consistency of MLE and variational estimates is proven but asymptotic normality requires that the estimators converges at rate at least n^{-1}, which is not proven in the paper, although some results were available for some particular cases (affiliation for example). Ambroise and Matias (2012) tackles the specific case of affiliation model with equal group proportion and proves the consistency and asymptotic normality of parameter estimates. Bickel et al. (2013) extends those results to arbitrary binary SBM graphs and improves Celisse et al. (2012) by removing the condition on the convergence rate. Following along the path of Bickel et al. (2013), Brault et al. (2017) proved consistency and asymptotic normality of estimators (MLE and variational) to weighted Latent Block Models where the weights distribution belongs to a regular one-dimensional exponential family. In particular, considering non-bounded edge values invalidates several parts of the proofs for binary graphs and requires substantial adaptations and additional results, notably concentration inequalities for sums of unbounded, non-gaussian random variables.

Some results are also available for the related semi-parametric problem of assignment reconstruction. Mariadassou and Matias (2015) show that the conditional distribution of the (latent) assignments converge to a degenerate distribution and Rohe et al. (2010) prove that, when the data are generated according to a SBM model, spectral methods are consistent. Choi et al. (2012a) extend those results to settings where the density of the graph goes to 0 as $\Omega(\log^{\alpha}(n)/n)$ (for α large enough) and/or the number of groups goes to $+\infty$ as $\sqrt{n}$. Finally, Wang and Bickel (2017) and Hu et al. (2017) also show that model selection for the number of groups is consistent for dense graphs, they suggest using a penalized likelihood criteria with penalty of the form $\frac{k(k+1)}{2}\log(n) + \lambda n \log(k)$ where λ is a tuning parameter.

In this paper we consider a simple setting with fixed number of groups and fixed density but weighted edges and missing values. In most network studies, there is a strong asymmetry between the presence of an edge and its absence: the lack of proof that an edge exists is taken as proof that the edge does not exist and edges with uncertain status are considered as non existent in the graph. This is the strategy adopted in most sparse asymptotic settings where the density of edges goes to 0 asymptotically Bickel et al. (2013). We adopt a different point of view where edges with uncertain status are considered as missing, rather than absent and explicitly accounted for their missing nature. We use the framework of Rubin (1976) and its application to network data, see Kolaczyk (2009) and Handcock and Gile (2010), for parameter inference in presence of missing values and more specifically its applications to SBM Tabouy et al. (2019a). We prove that, in the MCAR setting where each dyad is missing independently and with the same probability, the MLE and variational estimates are still consistent and asymptotically normal.

The article is organized as follows. We first present the model and missing data theory applied to our context with some examples of sampling designs. We then posit some defi-

nitions and discuss the assumptions required for our results in Section 2. In Section 3 we establish asymptotic normality for the complete-observed model (*i.e.* observed SBM where latent variables are known). Section 4 is the main result of this paper and states that the observed-likelihood behaves like the complete-observed likelihood (*i.e.* joint likelihood of the observed data and latent variables) close to its maximum. The proof is sketched in Section 5. Consequences for the MLE and variational estimator, as well as comparison to existing results, are in discussed in Section 6. Technical lemmas and details of the proofs are available in the appendices.

2 Statistical framework

2.1 Stochastic Block Model

In SBM, nodes from a set $\mathcal{N} \triangleq \{1, \dots, n\}$ are distributed among a set $\mathcal{Q} \triangleq \{1, \dots, Q\}$ of hidden blocks that model the latent structure of the graph. The block-memberships are encoded by $(z_i, i \in \mathcal{N})$ where the z_i are independent random variables with prior probabilities $\alpha = (\alpha_1, \dots, \alpha_Q)$, such that $\mathbb{P}(z_i = q) = \alpha_q$, for all $q \in \mathcal{Q}$. The value y_{ij} of any dyad (i, j) in $\mathcal{D} = \mathcal{N} \times \mathcal{N}$, with $i \neq j$, only depends on the blocks i and j belong to. The variables (y_{ij})s are thus independent conditionally on the (z_i)s:

$$y_{ij} \mid z_i = q, z_j = \ell \sim^{\text{ind}} \varphi(., \pi_{q\ell}), \quad \forall (i, j) \in \mathcal{D}, \quad i \neq j, \quad \forall (q, \ell) \in \mathcal{Q} \times \mathcal{Q}.$$

In the following, $\mathbf{y} = (y_{ij})_{i,j \in \mathcal{D}}$ is the $n \times n$ adjacency matrix of the random graph, $\mathbf{z} = (z_1, \dots, z_n)$ the n-vector of the latent blocks. With a slight abuse of notation, we associate to z_i a binary vector $(z_{i1}, \dots, z_{iQ})$ such that $z_i = q \Leftrightarrow z_{iq} = 1, z_{i\ell} = 0$, for all $\ell \neq q$. In this case $\mathbf{z}$ is a $n \times Q$ matrix.

We note the complete parameter set as $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi}) \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta}$ stands for the parameter space. When performing inference from data, we note $\boldsymbol{\theta}^\star = (\boldsymbol{\alpha}^\star, \boldsymbol{\pi}^\star)$ the true parameter set, *i.e.* the parameter values used to generate the data, and $\mathbf{z}^\star$ the true (and usually unobserved) memberships of nodes. For any $\mathbf{z}$, we also note:

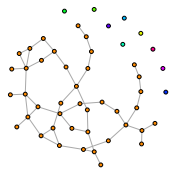
- $z_{+q} = \sum_i z_{iq}$ the size of block q for membership $\mathbf{z}$
- $z^\star_{+q}$ its counterpart for $\mathbf{z}^\star$.

2.2 Missing data for SBM

Regarding SBM inference, a missing value corresponds to a missing entry in the adjacency matrix $\mathbf{y}$, typically denoted by `NA`'s. We rely on the $n \times n$ sampling matrix $\mathbf{r}$ to record the missing state of each entry:

$$(r_{ij}) = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

As a shortcut, we use $\mathbf{y^o} = \{y_{ij} : r_{ij} = 1\}$ and $\mathbf{y^m} = \{y_{ij} : r_{ij} = 0\}$ to respectively denote the *observed* and *missing* dyads. The *sampling design* is the description of the stochastic process that generates $\mathbf{r}$. It is assumed that the network exists before the sampling design acts upon it, which is fully characterized by the conditional distribution $p_\psi(\mathbf{r}|\mathbf{y})$, the parameters of which are such that ψ and θ live in a product space $\Theta \times \Psi$. In this paper we are going to focus on a specific type of missingness, called missing completely at random (MCAR) for which $p_\psi(\mathbf{r}|\mathbf{y}) = p_\psi(\mathbf{r})$ and leave aside more complex forms of dependencies such as Missing at random (MAR) and Not missing at random (NMAR).



We then follow the framework of (Rubin, 1976) and Tabouy et al. (2019a) for missing data and define the joint probability density function as

$$p_{\theta,\psi}(\mathbf{y^o}, \mathbf{z}, \mathbf{r}) = \int p_\theta(\mathbf{y^o}, \mathbf{y^m}, \mathbf{z}) p_\psi(\mathbf{r}|\mathbf{y^o}, \mathbf{y^m}, \mathbf{z}) \mathrm{d}\mathbf{y^m}. \tag{2.2}$$

Property 1. *According to Equation (2.2), if the sampling design is MCAR, then maximizing $p_{\theta,\psi}(\mathbf{y^o}, \mathbf{z}, \mathbf{r})$ or $p_{\theta,\psi}(\mathbf{y^o}, \mathbf{r})$ in θ is equivalent to maximizing $p_\theta(\mathbf{y^o})$ in θ, this corresponds to the ignorability notion defined in Rubin (1976).*

2.3 Sampling design examples

We present here some examples of sampling designs to illustrate differences between notions of MCAR, MAR and NMAR.

Definition 3 (Random dyad sampling)**.** *Each dyad $(i, j) \in \mathcal{D}$ has the same probability $\mathbb{P}(r_{ij} = 1) = \rho$ of being observed, independently of the others. This design is MCAR.*

Definition 4 (Random node sampling)**.** *The random node sampling consists in selecting independently with probability ρ a set of nodes and then observing the corresponding rows and columns of matrix $\mathbf{y}$.*

The major point in both examples is that the probability (ρ in *random dyad sampling* and $1 - (1 - \rho)^2$ in the *random node sampling*) of observing a dyad does not depend on its value. In contrast, the following dyad-centered sampling design adapted to binary networks is NMAR since the probability to observe a dyad depends on its value:

Definition 5 (Double standard sampling)**.** *Each dyad $(i, j) \in \mathcal{D}$ is observed, independently of other dyads, with a probability depending on its value: $\mathbb{P}(r_{ij} = 1|y_{ij} = 0) = \rho_0$ and $\mathbb{P}(r_{ij} = 1|y_{ij} = 1) = \rho_1$.*

For non-binary networks, specifying the sampling design is more involved and requires defining the sampling density for every possible value of y_{ij}, *e.g.* $(\mathbb{P}(r_{ij} = 1|y_{ij} = k))_{k\in\mathbb{N}}$ for Poisson-valued edges.

2.4 Observed-likelihoods

When the labels are known, the *complete-observed log-likelihood* is given by:

$$\mathcal{L}_{co}(\mathbf{z}; \boldsymbol{\theta}) = \log p(\mathbf{y^o}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i,q} z_{iq} \log \alpha_q + \sum_{\substack{i,j,q,\ell \\ i\neq j}} z_{iq} z_{j\ell} r_{ij} \log \varphi(y_{ij}; \pi_{q\ell}) \tag{2.3}$$

But the labels are usually unobserved, and the *observed log-likelihood* is obtained by integration over all memberships:

$$\mathcal{L}_o(\boldsymbol{\theta}) = \log p(\mathbf{y^o}; \boldsymbol{\theta}) = \log \left(\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{y^o}, \mathbf{z}; \boldsymbol{\theta}) \right). \tag{2.4}$$

2.5 Models and Assumptions

We focus here on parametric models where φ belongs to a regular one-dimension exponential family in canonical form:

$$\varphi(y, \pi) = b(y) \exp(\pi y - \psi(\pi)), \tag{2.5}$$

where π belongs to the space $\mathcal{A}$, so that $\varphi(\cdot, \pi)$ is well defined for all $\pi \in \mathcal{A}$. Classical properties of exponential families ensure that ψ is convex, infinitely differentiable on $\mathring{\mathcal{A}}$, that $(\psi')^{-1}$ is well defined on $\psi'(\mathring{\mathcal{A}})$. Furthermore, when $y_\pi \sim \varphi(., \pi)$, $\mathbb{E}[y_\pi] = \psi'(\pi)$ and

$\mathbb{V}[y_\pi] = \psi''(\pi)$.

In the following, we assume that missing data are produced according to a *random dyad sampling* with parameter $\rho > 0$.

Moreover, we make the following assumptions on the parameter space :

A_1 : There exists a positive constant c, and a compact interval C_π such that

$$\rho \in [c, 1-c], \quad \boldsymbol{\Theta} \subset [c, 1-c]^{\mathcal{Q}} \times C_\pi^{\mathcal{Q}\times\mathcal{Q}} \quad \text{with} \quad C_\pi \subset \mathring{\mathcal{A}}.$$

A_2 : The true parameter $\boldsymbol{\theta}^\star = (\boldsymbol{\alpha}^\star, \boldsymbol{\pi}^\star)$ lies in the interior of $\boldsymbol{\Theta}$.

A_3 : The map $\pi \mapsto \varphi(\cdot, \pi)$ is injective.

A_4 : The coordinates of $\boldsymbol{\pi}^\star \psi'(\boldsymbol{\alpha}^\star)$, where ψ' is applied component-wise, are pairwise distinct.

The previous assumptions are standard. Assumption A_1 ensure that the group proportions and the sampling parameter are bounded away from 0 and 1 so that no group disappears when n goes to infinity. It also ensures that π is bounded away from the boundaries of the $\mathcal{A}$. This is essential for the sub-exponential properties of Propositions 13 and 14. A_2 and A_3 are necessary for identifiability purposes: the model is trivially not identifiable if the map $\pi \mapsto \varphi(., \pi)$ is not injective. A_4 states the identifiability of SBM parameters under *random dyad sampling*. Note that, combined with A_3, it implies that all columns and all rows of $\boldsymbol{\pi}^\star$ are distinct and therefore there are no two groups with identical connectivity profiles. In the following, we consider that $\mathcal{Q}$, the number of classes (or groups) is known.

2.6 Identifiability

Since $\mathbf{r}$ is independent on $\mathbf{y}$, the identifiability of SBM with emission law in the one-dimension exponential family under *random dyad sampling* can be stated in two steps. First the sampling parameter ρ and secondly the SBM parameters $\boldsymbol{\theta}^\star = (\boldsymbol{\alpha}^\star, \boldsymbol{\pi}^\star)$ given ρ.

Proposition 11. *The sampling parameter $\rho > 0$ of random dyad sampling is identifiable w.r.t. the sampling distribution.*

Proof. See Tabouy et al. (2019a). The proof does not depend on $\mathbf{y}$ being binary but also holds for $\mathbf{y}$ distributed as in Eq. (2.5). □

Proposition 12. *Let $n \geq 2Q$ and assume that for any $1 \leq q \leq Q$, $\rho > 0$, $\pi_q^\star > 0$ and that the coordinates of $\boldsymbol{\alpha}^\star \psi'(\boldsymbol{\pi}^\star)$, where ψ' is applied component-wise, are pairwise distinct. Then, under random dyad sampling, SBM parameters are identifiable w.r.t. the distribution of the observed part of the SBM up to label switching.*

Proof. The proof is nearly identical to the one written in Tabouy et al. (2019a) and inspired by Celisse et al. (2012) for the binary SBM under *random dyad sampling*. However, substituting $\mathbb{E}[y_{ij}|z_i = q]$ to s_q in the proof ensures that $\boldsymbol{\alpha}^\star$ is identifiable. Finally, the fact that $(\psi')^{-1}$ is a one-to-one map ensures that $\boldsymbol{\pi}^\star$ is identifiable. □

Note that asymptotically, the assumption $n \geq 2Q$ is always satisfied since Q is fixed and n grows to infinity.

2.7 Sub-exponential variables

Remarque 2. *Since we restricted π in a bounded subset of $\mathring{A}$, the variance of y_π is bounded away from 0 and $+\infty$. We note*

$$\bar{\sigma}^2 = \sup_{\pi \in C_\pi} \mathbb{V}(y_\pi) < +\infty \quad \text{and} \quad \underline{\sigma}^2 = \inf_{\pi \in C_\pi} \mathbb{V}(y_\pi) > 0. \tag{2.6}$$

Similarly, since π belongs to a bounded subset of a open interval, there exists a constant $\kappa > 0$, such that $[\pi - \kappa, \pi + \kappa] \subset \mathring{A}$ uniformly over all $\pi \in C_\pi$

Proposition 13. *With the previous notations, if $\pi \in C_\pi$ and $y_\pi \sim \varphi(., \pi)$, then y_π is sub-exponential with parameters $(\bar{\sigma}^2, \kappa^{-1})$.*

Proposition 14. *Considering $x = y_\pi r_{ij} + \lambda r_{ij}$ (we recall that $r_{ij} \sim \mathcal{B}(\rho)$), with r_{ij} independent of y_π and $\lambda \in \mathbb{R}$ bounded. There are non-negative numbers ν and b such that x is sub-exponential with parameters (ν^2, b^{-1}).*

Proof. These results derive directly from theorem 20 (statement 2.). □

2.8 Symmetry

We now introduce the concepts of assignments and parameter symmetries, that must be accounted for when studying the asymptotic properties of the MLE. Complications stemming from symmetries are related to but no equivalent to the problem of label-switching in mixture models.

Definition 6 (permutation)**.** *Let s be a permutation on $\{1, \dots, \mathcal{Q}\}$. If $\boldsymbol{A}$ is a matrix with $\mathcal{Q}$ columns and n rows, we define $\boldsymbol{A}^s$ as the matrix obtained by permuting the columns of $\boldsymbol{A}$ according to s,* i.e. *for any row i and column q of $\boldsymbol{A}$, $A^s_{iq} = A_{is(q)}$. If $\boldsymbol{C}$ is a matrix with $\mathcal{Q}$ rows and $\mathcal{Q}$ columns, $\boldsymbol{C}^s$ is defined similarly:*

$$\boldsymbol{A}^s = \left(A_{is(q)}\right)_{i,q} \quad \boldsymbol{C}^s = \left(C_{s(q)s(\ell)}\right)_{q,\ell}$$

Definition 7 (equivalence)**.** *We define the following equivalence relationships:*

- *Two assignments $\mathbf{z}$ and $\mathbf{z}'$ are* equivalent*, noted $\sim$, if they are equal up to label permutation,* i.e. *there exists a permutation s such that $\mathbf{z}' = \mathbf{z}^s$.*
- *Two parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are* equivalent*, noted $\sim$, if they are equal up to label permutation,* i.e. *there exists a permutation s such that $(\boldsymbol{\alpha}^s, \boldsymbol{\pi}^s) = (\boldsymbol{\alpha}', \boldsymbol{\pi}')$.*
- *$(\boldsymbol{\theta}, \mathbf{z})$ and $(\boldsymbol{\theta}', \mathbf{z}')$ are* equivalent*, noted $\sim$, if they are equal up to label permutation on $\boldsymbol{\pi}$ and $\mathbf{z}$,* i.e. *there exists a permutation s such that $(\boldsymbol{\pi}^s, \mathbf{z}^s) = (\boldsymbol{\pi}', \mathbf{z}')$. This is* label-switching.

Definition 8 (symmetry)**.** *We say that the parameter $\boldsymbol{\theta}$* exhibits symmetry *for the permutation s if*

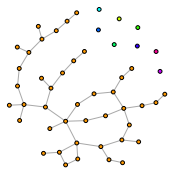
$$(\boldsymbol{\alpha}^s, \boldsymbol{\pi}^s) = (\boldsymbol{\alpha}, \boldsymbol{\pi}).$$

$\boldsymbol{\theta}$ exhibits symmetry *if it exhibits symmetry for any non trivial permutations s. Finally the set of permutations for which $\boldsymbol{\theta}$ exhibits symmetry is noted* $\mathrm{Sym}(\boldsymbol{\theta})$.

Remarque 3. *The set of parameters that exhibit symmetry is a manifold of null Lebesgue measure in $\boldsymbol{\Theta}$. The notion of symmetry allows us to deal with a notion of non-identifiability of the class labels that is subtler than and different from label switching. More precisely*

$$\begin{aligned} \textit{Label switching is when :} \quad p(\mathbf{y^o}, \mathbf{z}, \boldsymbol{\theta}) &= p(\mathbf{y^o}, \mathbf{z}^s, \boldsymbol{\theta}^s), \ \theta \neq \theta^s \ \forall s \\ \textit{Symmetry is when :} \quad p(\mathbf{y^o}, \mathbf{z}, \boldsymbol{\theta}) &= p(\mathbf{y^o}, \mathbf{z}^s, \boldsymbol{\theta}), \ \forall s \in Sym(\theta) \end{aligned}$$

In particular, in label-switching, $\mathbf{z}$ and $\mathbf{z}^s$ have the same likelihood but under equivalent yet different parameters $\boldsymbol{\theta}$s. In contrast, in the presence of symmetry, multiple assignments can have exactly the same likelihood under $\boldsymbol{\theta}$.



The issue of symmetry forces us to use a notion of distance between assignment that is invariant to label permutation.

Definition 9 (distance)**.** *We define the following distance, up to equivalence, between configurations* $\mathbf{z}$ *and* $\mathbf{z}^\star$*:*

$$\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} = \inf_{\mathbf{z}' \sim \mathbf{z}} \|\mathbf{z}' - \mathbf{z}^\star\|_0$$

where, for all matrix $\mathbf{z}$*, we use the Hamming norm* $\|\cdot\|_0$ *defined by*

$$\|\mathbf{z}\|_0 = \frac{1}{2} \sum_{i,q} \mathbb{1}\{z_{iq} \neq 0\}.$$

Definition 10 (Set of local assignments)**.** *We note* $S(\mathbf{z}^\star, r)$ *the set of configurations that have a representative (for* $\sim$*) within relative radius* r *of* $\mathbf{z}^\star$*:*

$$S(\mathbf{z}^\star, r) = \{\mathbf{z} : \|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \leq rn\}$$

2.9 Other definitions

We finally introduce a few useful notions that will be instrumental in the proofs. The first is "regular" assignments, for which each group has "enough" nodes:

Definition 11 (c-regular assignments)**.** *Let* $\mathbf{z} \in \mathcal{Z}$*. For any* $c > 0$*, we say that* $\mathbf{z}$ *is* c-regular *if*

$$\min_q z_{+q} \geq cn. \tag{2.7}$$

Class distinctness $\delta(\boldsymbol{\pi})$ captures the differences between groups: lower values of $\delta(\boldsymbol{\pi})$ means that at least two classes are very similar. $\delta(\boldsymbol{\pi})$ is intrinsically linked to the convergence rate of several estimates.

Definition 12 (class distinctness)**.** *For* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi}) \in \boldsymbol{\Theta}$*. We define:*

$$\delta(\boldsymbol{\pi}) = \min_{q,q'} \max_{\ell} \mathcal{KL}(\pi_{q\ell}, \pi_{q'\ell})$$

with $\mathcal{KL}(\pi, \pi') = \mathbb{E}_\pi[\log(\varphi(Y, \pi)/\varphi(Y, \pi'))] = \psi'(\pi)(\pi - \pi') + \psi(\pi') - \psi(\pi)$ *the Kullback-Leibler divergence between* $\varphi(., \pi)$ *and* $\varphi(., \pi')$*, when* φ *comes from an exponential family.*

Remarque 4. *Since all* $\boldsymbol{\pi}$ *have distinct rows and columns,* $\delta(\boldsymbol{\pi}) > 0$*.*

Finally, the confusion matrix allows to compare groups between assignments:

Definition 13 (confusion matrix)**.** *For given assignments* $\mathbf{z}$ *and* $\mathbf{z}^\star$*, we define the* confusion matrix *between* $\mathbf{z}$ *and* $\mathbf{z}^\star$*, noted* $\mathbb{R}(\mathbf{z})$*, as follows:*

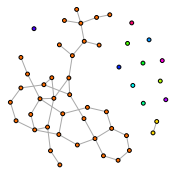
$$\mathbb{R}(\mathbf{z})_{qq'} = \frac{1}{n} \sum_i z^\star_{iq} z_{iq'} \tag{2.8}$$

Definition 14. *For more conciseness, we define*

$$\boldsymbol{S}^\star = (S^\star_{q\ell})_{q\ell} = \left(\psi'(\pi^\star_{q\ell})\right)_{q\ell} \tag{2.9}$$

3 Complete-observed Model

In the following we study the asymptotic properties of the complete-observed data model, *i.e.* when the true assignment $\mathbf{z}^\star$ is known.



Proposition 15. *Under random dyad sampling, defining* $N_i = \sum_{i,j} R_{ij}$ *and* $\Omega_{0,n} = \{\forall i \in \{1, ..., n\},\ N_i \geqslant 1\}$ *the set of nodes with at least one dyaddy observed. Then*

$$\mathbb{P}\left(\lim_{n\to+\infty} \Omega_{0,n}\right) = 1.$$

Proof. This proposition is a direct consequence of Borel-Cantelli's theorem. Details are available in appendix 10. □

Remarque 5. *This result shows that, with high probability, the network has no unobserved node. In the remainder, we work conditionally on* $\Omega_{0,n}$.

Let $\widehat{\boldsymbol{\theta}}_c = (\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\pi}})$ be the MLE of $\boldsymbol{\theta}$ in the complete-observed data model. Simple manipulations of Equation (2.3) yield:

$$\begin{aligned} \hat{\alpha}_q = \widehat{\alpha}_q(\mathbf{z}) &= \frac{z_{+q}}{n} \\ \widehat{y}_{q\ell}(\mathbf{z}) = \frac{\sum_{i\neq j} y_{ij} r_{ij} z_{iq} z_{j\ell}}{\sum_{i\neq j} r_{ij} z_{iq} z_{j\ell}} &\quad \widehat{\pi}_{q\ell} = \widehat{\pi}_{q\ell}(\mathbf{z}) = (\psi')^{-1}\left(\widehat{y}_{q\ell}(\mathbf{z})\right) \end{aligned} \tag{3.1}$$

Since there are missing values in the adjacency matrix, we need the following technical lemma to prove asymptotic normality of $\pi_{q\ell}$'s in the complete data model.

Lemme 1.

$$U_n = \frac{1}{n(n-1)} \sum_{i\neq j} r_{ij} z_{iq} z_{j\ell} \xrightarrow[n\to+\infty]{\mathbb{P}} \rho \alpha_q \alpha_l$$

Proof. The proof of this lemma is based on Hoeffding's decomposition for U-statistics and on the proof of Hoeffding's concentration inequality. Details are postponed to appendix 10. □

Proposition 16. *Let* $\Sigma_{\boldsymbol{\alpha}^\star} = \mathrm{Diag}(\boldsymbol{\alpha}^\star) - \boldsymbol{\alpha}^\star (\boldsymbol{\alpha}^\star)^T$. $\Sigma_{\boldsymbol{\alpha}^\star}$ *is semi-definite positive, of rank* $\mathcal{Q} - 1$, *and* $\widehat{\boldsymbol{\alpha}}$ *is asymptotically normal:*

$$\sqrt{n}\left(\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star) - \boldsymbol{\alpha}^\star\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\alpha}^\star}) \tag{3.2}$$

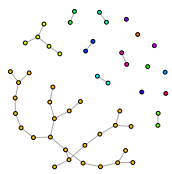
Similarly, let $V(\boldsymbol{\pi}^\star)$ *be the matrix defined by* $[V(\boldsymbol{\pi}^\star)]_{q\ell} = 1/\psi''(\pi^\star_{q\ell})$ *and* $\Sigma_{\boldsymbol{\pi}^\star} = \rho^{-1}\,\mathrm{Diag}^{-1}(\boldsymbol{\alpha}^\star) V(\boldsymbol{\pi}^\star)\,\mathrm{Diag}^{-1}(\boldsymbol{\alpha}^\star)$. *Then the estimates* $\hat{\pi}_{q\ell}(\mathbf{z}^\star)$ *are independent and asymptotically Gaussian with limit distribution:*

$$\sqrt{n(n-1)}\left(\widehat{\pi}_{q\ell}(\mathbf{z}^\star) - \pi^\star_{q\ell}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma_{\boldsymbol{\pi}^\star, q\ell}) \quad \textit{for all } q, \ell \tag{3.3}$$

Proof. The proof is postponed to appendix 10. The first part is a direct application of central limit theorem for i.i.d. variables and the second part relies on a variant of the central limit theorem for random sums of random variables. □

Proposition 17 (Local asymptotic normality)**.** *Let* $\mathcal{L}^\star_{co}$ *be the complete likelihood function defined on* $\boldsymbol{\Theta}$ *by* $\mathcal{L}^\star_{co}(\boldsymbol{\alpha}, \boldsymbol{\pi}) = \log p(\mathbf{y}^o, \mathbf{z}^\star; \boldsymbol{\theta})$. *For any* s, t *and* u *in a compact set, we have:*

$$\begin{aligned} \mathcal{L}^\star_{co}\left(\boldsymbol{\alpha}^\star + \frac{s}{\sqrt{n}}, \boldsymbol{\pi}^\star + \frac{u}{\sqrt{n(n-1)}}\right) &= \mathcal{L}^\star_{co}(\boldsymbol{\theta}^\star) + s^T \mathbf{Y}_{\boldsymbol{\alpha}^\star} + Tr(u^T \mathbf{Y}_{\boldsymbol{\pi}^\star}) \\ &- \left(\frac{1}{2} s^T \Sigma_{\boldsymbol{\alpha}^\star} s + \frac{1}{2} Tr\left((u \odot u)^T \Sigma_{\boldsymbol{\pi}^\star}\right)\right) \\ &+ o_P(1) \end{aligned}$$



where $\odot$ denote the Hadamard product of two matrices (element-wise product) and $\Sigma_{\boldsymbol{\alpha}^\star}$ and $\Sigma_{\boldsymbol{\pi}^\star}$ are defined in Proposition 16. $\mathbf{Y}_{\boldsymbol{\alpha}^\star}$ is asymptotically Gaussian with zero mean and variance matrix $\Sigma_{\boldsymbol{\alpha}^\star}$. $\mathbf{Y}_{\boldsymbol{\pi}^\star}$ is a random matrix with independent entries that are asymptotically Gaussian zero mean and variance $\Sigma_{\boldsymbol{\pi}^\star}$.

Proof. This result is based on a Taylor expansion of $\mathcal{L}^\star_{co}$ in a neighborhood of $(\boldsymbol{\alpha}^\star, \boldsymbol{\pi}^\star)$. Details are available in appendix 10. □

4 Main Result

Our main result compares the observed likelihood ratio $p(\mathbf{y}^o; \boldsymbol{\theta})/p(\mathbf{y}^o; \boldsymbol{\theta}^\star)$ with the complete likelihood $p(\mathbf{y}^o, \mathbf{z}^\star; \boldsymbol{\theta}')/p(\mathbf{y}^o, \mathbf{z}^\star; \boldsymbol{\theta}^\star)$ to show that they have the same argmax. To ease the comparison, we work only on the high probability set Ω_1 of $c/2$-regular configurations, *i.e.* that have $\Omega(n)$ nodes in each group as defined in Section 2,

Proposition 18. *Define $\mathcal{Z}_1$ as the subset of $\mathcal{Z}$ made of $c/2$-regular assignments, with c defined in assumption H_1. Note Ω_1 the event $\{\mathbf{z}^\star \in \mathcal{Z}_1\}$, then:*

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(\bar{\Omega}_1\right) \leq \mathcal{Q} \exp\left(-\frac{nc^2}{2}\right).$$

Proof. This proposition is a consequence of Hoeffding's inequality. See appendix 10 for more details. □

We can now state our main result:

Theorem 15 (complete-observed)**.** *Assume that A_1 to A_4 with random-dyad sampling hold for the Stochastic Block Model of known order with $n \times n$ observations coming from an univariate exponential family and define $\#\,\mathrm{Sym}(\boldsymbol{\theta})$ as the set of permutation s for which $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ exhibits symmetry. Then, for n tending to infinity, the observed likelihood ratio behaves like the complete likelihood ratio, up to a bounded multiplicative factor:*

$$\frac{p(\mathbf{y^o}; \boldsymbol{\theta})}{p(\mathbf{y^o}; \boldsymbol{\theta}^\star)} = \frac{\#\,\mathrm{Sym}(\boldsymbol{\theta})}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)} \max_{\boldsymbol{\theta}' \sim \boldsymbol{\theta}} \frac{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}')}{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)} (1 + o_P(1)) + o_P(1)$$

where the o_P is uniform over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

The maximum over all $\boldsymbol{\theta}'$ that are equivalent to $\boldsymbol{\theta}$ stems from the fact that because of label-switching, $\boldsymbol{\theta}$ is only identifiable up to its $\sim$-equivalence class from the observed likelihood, whereas it is completely identifiable from the complete likelihood. The multiplicative factor arises from the fact that equivalent assignments have exactly the same complete likelihood and contribute equally to the observed likelihood.

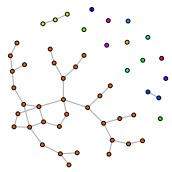
Corollary 16. *If $\boldsymbol{\Theta}$ contains only parameters with no symmetry:*

$$\frac{p(\mathbf{y^o}; \boldsymbol{\theta})}{p\left(\mathbf{y^o}; \boldsymbol{\theta}^\star\right)} = \max_{\boldsymbol{\theta}' \sim \boldsymbol{\theta}} \frac{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}')}{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)} (1 + o_P(1)) + o_P(1)$$

where the o_P is uniform over all $\boldsymbol{\Theta}$.

5 Proof Sketch

The proof of theorem relies on controlling deviations of the log-likelihood ratios from their expectations. We introduce a few notations for those quantities.



5.1 log-likelihood ratios

Definition 17. *We define the conditional log-likelihood ratio LR and its expectation ELR as:*

$$LR(\boldsymbol{\theta}, \mathbf{z}) = \log \frac{p(\mathbf{y^o}|\mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{y^o}|\mathbf{z}^\star; \boldsymbol{\theta}^\star)} \quad \text{and} \quad ELR(\boldsymbol{\theta}, \mathbf{z}) = \mathbb{E}_{\boldsymbol{\theta}^\star}\left[LR(\boldsymbol{\theta}, \mathbf{z})|\, \mathbf{z}^\star\right] \tag{5.1}$$

We also define the profile ratio Λ and its counterpart $\tilde{\Lambda}$ as:

$$\Lambda(\mathbf{z}) = \max_{\boldsymbol{\theta}} LR(\boldsymbol{\theta}, \mathbf{z}) \quad \text{and} \quad \tilde{\Lambda}(\mathbf{z}) = \max_{\boldsymbol{\theta}} ELR(\boldsymbol{\theta}, \mathbf{z}). \tag{5.2}$$

Proposition 19. *Conditionally on $\mathbf{z}^\star$, we have*

$$\bar{y}_{q\ell}(\mathbf{z}) := \mathbb{E}_{\boldsymbol{\theta}^\star}[\widehat{y}_{q\ell}(\mathbf{z})|\mathbf{z}^\star] = \frac{\left[\mathbb{R}(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}(\mathbf{z})\right]_{q\ell}}{\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})} \tag{5.3}$$

with $\bar{y}_{q\ell}(\mathbf{z}) = 0$ for $\mathbf{z}$ such that $\widehat{\alpha}_q(\mathbf{z}) = 0$ or $\widehat{\alpha}_\ell(\mathbf{z}) = 0$.

Remarque 6. *Note the absence of the random variable $\mathbf{r}$ in $\bar{y}_{q\ell}(\mathbf{z})$.*

The following decomposition of $p(\mathbf{y^o}; \boldsymbol{\theta})$ highlights the importance of $F_n(\boldsymbol{\theta}, \mathbf{z})$:

$$p(\mathbf{y^o}; \boldsymbol{\theta}) = \sum_{(\mathbf{z})} p(\mathbf{y^o}, \mathbf{z}; \boldsymbol{\theta}) = p(\mathbf{y^o}|\mathbf{z}^\star; \boldsymbol{\theta}^\star) \sum_{(\mathbf{z})} p(\mathbf{z}; \boldsymbol{\theta}) \exp(LR(\boldsymbol{\theta}, \mathbf{z})).$$

Since $LR(\boldsymbol{\theta}, \mathbf{z}) \leq \Lambda(\mathbf{z})$, the profile ratio is useful to remove the dependency on $\boldsymbol{\theta}$ and reduce the study to a series of problems depending only on $\mathbf{z}$. The following propositions show when those quantities reach their maximum values and what the corresponding values are.

Proposition 20 (maximum of ELR and $\tilde{\Lambda}$ in $\boldsymbol{\theta}$)**.** *The functions $LR(\boldsymbol{\theta}, \mathbf{z})$ and $ELR(\boldsymbol{\theta}, \mathbf{z})$ are maximum respectively in $\boldsymbol{\pi}$ for $\widehat{\boldsymbol{\pi}}(\mathbf{z})$ and $\bar{\boldsymbol{\pi}}(\mathbf{z})$ defined by:*

$$\widehat{\pi}(\mathbf{z})_{q\ell} = (\psi')^{-1}(\widehat{y}_{q\ell}(\mathbf{z})) \quad \text{and} \quad \bar{\pi}(\mathbf{z})_{q\ell} = (\psi')^{-1}(\bar{y}_{q\ell}(\mathbf{z}))$$

so that

$$\Lambda(\mathbf{z}) = LR(\widehat{\boldsymbol{\pi}}(\mathbf{z}), \mathbf{z}) \quad \text{and} \quad \tilde{\Lambda}(\mathbf{z}) = ELR(\bar{\boldsymbol{\pi}}(\mathbf{z}), \mathbf{z}).$$

Proposition 21 (Local upper bound for $\tilde{\Lambda}$)**.** *Conditionally upon Ω_1, there exists a positive constant C such that for all $\mathbf{z} \in S(\mathbf{z}^\star, C)$:*

$$\tilde{\Lambda}(\mathbf{z}) \leq -c\rho n \frac{3\delta(\boldsymbol{\pi}^\star)}{4} \|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \tag{5.4}$$

Proposition 22 (maximum of ELR and $\tilde{\Lambda}$ in $(\boldsymbol{\theta}, \mathbf{z})$)**.** *$ELR$ can be written:*

$$ELR(\boldsymbol{\theta}, \mathbf{z}) = -\rho n^2 \sum_{q,q'} \sum_{\ell,\ell'} \mathbb{R}(\mathbf{z})_{q,q'} \mathbb{R}(\mathbf{z})_{\ell,\ell'} \mathcal{KL}(\pi^{\star q\ell}, \pi_{q'\ell'}) \leq 0. \tag{5.5}$$

Conditionally on the set Ω_1 of regular assignments and for $n > 2/c$,

(i) ELR is maximized at $(\boldsymbol{\pi}^\star, \mathbf{z}^\star)$ and its equivalence class and $ELR(\boldsymbol{\pi}^\star, \mathbf{z}^\star) = 0$.

(ii) $\tilde{\Lambda}$ is maximized at $\mathbf{z}^\star$ and its equivalence class and $\tilde{\Lambda}(\mathbf{z}^\star) = 0$.

(iii) The maximum of $\tilde{\Lambda}$ (and hence the maximum of ELR) is well separated.

Proofs of Propositions 19, 20, 22 and 21 are postponed to Appendix 11.

5.2 High level view of the proof

The proof proceeds with an examination of the asymptotic behavior of LR on three types of configurations that partition $\mathcal{Z}$:

1. *global control*: for $\mathbf{z}$ such that $\tilde{\Lambda}(\mathbf{z}) = \Omega(-n^2)$, Proposition 23 proves a large deviation behavior and shows that $LR = -\Omega_P(n^2)$. In turn, those assignments contribute a o_P of $p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star))$ to the sum (Proposition 24).

2. *local control*: a small deviation result (Proposition 25) is needed to show that the combined contribution of assignments close to but not equivalent to $\mathbf{z}^\star$ is also a o_P of $p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)$ (Proposition 26).

3. *equivalent assignments*: Proposition 27 examines which of the remaining assignments, all equivalent to $\mathbf{z}^\star$, contribute to the sum.

These results are presented in next section 5.3 and their proofs postponed to Appendix 11. They are then put together in section 5.4 to prove our main result. The remainder of the section is devoted to the asymptotic of the ML and variational estimators as a consequence of the main result.

5.3 Different asymptotic behaviors

Global Control

Proposition 23 (large deviations of LR)**.** *Let* $\mathrm{Diam}(\boldsymbol{\Theta}) = \sup_{\boldsymbol{\theta},\boldsymbol{\theta}'} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty$. *For all* $\varepsilon_n < \nu b$ *and* n *large enough that* $2\sqrt{2n^2}\epsilon_n \geq Q^2$

$$\sup_{\boldsymbol{\theta},\mathbf{z}} \left\{ LR(\boldsymbol{\theta}, \mathbf{z}) - \tilde{\Lambda}(\mathbf{z}) \right\} = \mathcal{O}_p(n^2\epsilon_n) \tag{5.6}$$

Proposition 24 (contribution of global assignments)**.** *Choose* t_n *decreasing to* 0 *such that* $\frac{\rho n t_n}{\sqrt{\log(n)}} \to +\infty$. *Then conditionally on* Ω_1 *and for* n *large enough that* $2\sqrt{2n^2}\epsilon_n \geq \mathcal{Q}^2$, *we have:*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\mathbf{z}\notin S(\mathbf{z}^\star, t_n)} p(\mathbf{z}, \mathbf{y^o}; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{y^o}; \boldsymbol{\theta}^\star))$$

Local Control

Proposition 25 (small deviations LR)**.** *Conditionally upon* Ω_1,

$$\sup_{\mathbf{z}\nsim\mathbf{z}^\star} \frac{\Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}^\star)}{n\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}} = o_P(1) \tag{5.7}$$

The next proposition uses Propositions 25 and 22 to show that the combined contribution to the observed likelihood of assignments close to $\mathbf{z}^\star$ is also a o_P of $p(\mathbf{z}^\star, \mathbf{y}^o; \boldsymbol{\theta}^\star)$:

Proposition 26 (contribution of local assignments)**.** *With the previous notations and* C *the positive constant defined in Proposition 21:*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\substack{\mathbf{z}\in S(\mathbf{z}^\star, C)\\ \mathbf{z}\nsim\mathbf{z}^\star}} p(\mathbf{z}, \mathbf{y}^o; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{y}^o; \boldsymbol{\theta}^\star))$$

Equivalent assignments

It remains to study the contribution of equivalent assignments.

Proposition 27 (contribution of equivalent assignments)**.** *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we have*

$$\sum_{\mathbf{z}\sim\mathbf{z}^\star} \frac{p(\mathbf{y^o}, \mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)} = \#\,\mathrm{Sym}(\boldsymbol{\theta}) \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}')}{p(\mathbf{y^o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)}(1 + o_P(1))$$

where the o_P is uniform in $\boldsymbol{\theta}$.

5.4 Proof of the main result

Proof. We work conditionally on Ω_1. Choose $\mathbf{z}^\star \in \mathcal{Z}_1$ and a sequence t_n decreasing to 0 but satisfying $\rho n t_n/\sqrt{\log(n)} \to +\infty$. According to Proposition 24,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\mathbf{z}\notin S(\mathbf{z}^\star, t_n)} p(\mathbf{z}, \mathbf{y}^o; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{y}^o; \boldsymbol{\theta}^\star))$$

Since t_n decreases to 0, it gets smaller than C (used in proposition 26) for n large enough. As this point, Proposition 26 ensures that:

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{\substack{\mathbf{z}\in S(\mathbf{z}^\star, t_n)\\ \mathbf{z}\nsim\mathbf{z}^\star}} p(\mathbf{z}, \mathbf{y}^o; \boldsymbol{\theta}) = o_P(p(\mathbf{z}^\star, \mathbf{y}^o; \boldsymbol{\theta}^\star))$$

And therefore the observed likelihood ratio reduces as:

$$\begin{aligned}\frac{p(\mathbf{y}^o; \boldsymbol{\theta})}{p(\mathbf{y}^o; \boldsymbol{\theta}^\star)} &= \frac{\sum_{\mathbf{z}\sim\mathbf{z}^\star} p(\mathbf{y}^o, \mathbf{z}; \boldsymbol{\theta}) + \sum_{\mathbf{z}\nsim\mathbf{z}^\star} p(\mathbf{y}^o, \mathbf{z}; \boldsymbol{\theta})}{\sum_{\mathbf{z}\sim\mathbf{z}^\star} p(\mathbf{y}^o, \mathbf{z}; \boldsymbol{\theta}^\star) + \sum_{\mathbf{z}\nsim\mathbf{z}^\star} p(\mathbf{y}^o, \mathbf{z}; \boldsymbol{\theta}^\star)} \\ &= \frac{\sum_{\mathbf{z}\sim\mathbf{z}^\star} p(\mathbf{y}^o, \mathbf{z}; \boldsymbol{\theta}) + p(\mathbf{y}^o; \mathbf{z}^\star, \boldsymbol{\theta}^\star) o_P(1)}{\sum_{\mathbf{z}\sim\mathbf{z}^\star} p(\mathbf{y}^o, \mathbf{z}; \boldsymbol{\theta}^\star) + p(\mathbf{y}^o; \mathbf{z}^\star, \boldsymbol{\theta}^\star) o_P(1)}\end{aligned}$$

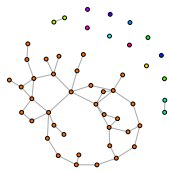
And Proposition 27 allows us to conclude

$$\frac{p(\mathbf{y}^o; \boldsymbol{\theta})}{p(\mathbf{y}^o; \boldsymbol{\theta}^\star)} = \frac{\#\,\mathrm{Sym}(\boldsymbol{\theta})}{\#\,\mathrm{Sym}(\boldsymbol{\theta}^\star)} \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{y}^o, \mathbf{z}^\star; \boldsymbol{\theta}')}{p(\mathbf{y}^o, \mathbf{z}^\star; \boldsymbol{\theta}^\star)}(1 + o_P(1)) + o_P(1).$$

□

6 Variational and Maximum Likelihood Estimates

This section is devoted to the asymptotic of the ML and variational estimators in the incomplete data model as a consequence of the main result 15. Note that, with high probability, ML and variational estimators have no symmetry since the set $\{\boldsymbol{\theta} : \#\,\mathrm{Sym}(\boldsymbol{\theta}) > 1\}$ is a manifold of null Lebesque's measure in $\boldsymbol{\Theta}$.



6.1 ML estimator

The asymptotic behavior of the maximum likelihood estimator in the incomplete data model is a direct consequence of Theorem 15 and Proposition 17.

Corollary 18 (Asymptotic behavior of $\widehat{\boldsymbol{\theta}}_{MLE}$)**.** *Denote $\widehat{\boldsymbol{\theta}}_{MLE}$ the maximum likelihood estimator and use the notations of Proposition 16. There exist permutations s of $\{1,\ldots,\mathcal{Q}\}$ such that*

$$\begin{aligned}\hat{\boldsymbol{\alpha}}\left(\mathbf{z}^{\star}\right)-\widehat{\boldsymbol{\alpha}}_{MLE}^{s} &= o_P\left(n^{-1/2}\right),\\ \hat{\boldsymbol{\pi}}\left(\mathbf{z}^{\star}\right)-\widehat{\boldsymbol{\pi}}_{MLE}^{s} &= o_P\left(n^{-1}\right).\end{aligned}$$

Hence, the maximum likelihood estimator for the SBM under random-dyad sampling condition is consistent and asymptotically normal, with the same behavior as the maximum likelihood estimator in the complete data model. The proof is postponed to appendix 11.10.

6.2 Variational estimator

Due to the complex dependency structure of the observations, the maximum likelihood estimator of the SBM is not numerically tractable, even with the *Expectation Maximization* algorithm. In practice, a variational approximation is often used (see Daudin et al., 2008): for any joint distribution $\mathbb{Q}\in\mathcal{Q}$ on $\mathcal{Z}$ a lower bound of $\mathcal{L}(\boldsymbol{\theta})$ is given by

$$\begin{aligned}J\left(\mathbb{Q},\boldsymbol{\theta}\right) &= \mathcal{L}(\boldsymbol{\theta})-\mathcal{KL}\left(\mathbb{Q},p\left(.;\boldsymbol{\theta},\mathbf{y}^{o}\right)\right)\\ &= \mathbb{E}_{\mathbb{Q}}\left[\mathcal{L}_{co}\left(\mathbf{z};\boldsymbol{\theta}\right)\right]+\mathcal{H}\left(\mathbb{Q}\right).\end{aligned}$$

where $\mathcal{H}\left(\mathbb{Q}\right)=-\mathbb{E}_{\mathbb{Q}}[\log(\mathbb{Q})]$. Choosing $\mathcal{Q}$ to be the set of product distributions, such that for all $\mathbf{z}$

$$\mathbb{Q}\left(\mathbf{z}\right)=\prod_{i,q}\mathbb{Q}\left(z_{iq}=1\right)^{z_{iq}}$$

allows us to obtain tractable expressions of $J\left(\mathbb{Q},\boldsymbol{\theta}\right)$. The variational estimate $\widehat{\boldsymbol{\theta}}_{var}$ of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}}_{var}\in\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\max_{\mathbb{Q}\in\mathcal{Q}}J\left(\mathbb{Q},\boldsymbol{\theta}\right).$$

The following corollary states that $\widehat{\boldsymbol{\theta}}_{var}$ has the same asymptotic properties as $\widehat{\boldsymbol{\theta}}_{MLE}$ and $\widehat{\boldsymbol{\theta}}_{MC}$, in particular is consistent and asymptotically normal.

Theorem 19 (Variational estimate)**.** *Under the assumptions of Theorem 15, there exist permutations s of $\{1,\ldots,\mathcal{Q}\}$ such that*

$$\begin{aligned}\hat{\boldsymbol{\alpha}}\left(\mathbf{z}^{\star}\right)-\widehat{\boldsymbol{\alpha}}_{var}^{s} &= o_P\left(n^{-1/2}\right),\\ \hat{\boldsymbol{\pi}}\left(\mathbf{z}^{\star}\right)-\widehat{\boldsymbol{\pi}}_{var}^{s} &= o_P\left(n^{-1}\right).\end{aligned}$$

The proof is very similar to the proof of Theorem 18 and postponed to appendix 11.10.

7 Discussion

Close examination of the different proofs, especially of Prop. 26, reveals that the quantities driving convergence of the estimates are $\rho n\delta(\boldsymbol{\pi}^{\star})$, which must go to $+\infty$ with n to ensure validity of Prop. 24, and $\rho n t_n\delta(\boldsymbol{\pi}^{\star})$, which must be larger than $\sqrt{\log(n)}$ while $t_n\to 0$, to ensure validity of Prop. 26. Both conditions are met as soon as $\rho\gg\log(n)/n$, allowing for a

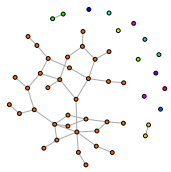
large fraction of missing edges. Note that this limiting rate for missingness is the same as the one found for graph density in sparse settings to achieve consistency and local asymptotic normality of $\boldsymbol{\theta}$ (Bickel et al., 2013).

In this paper, we focused on data sampled according to *random dyad sampling*. However, as described in section 2.3, there are many other ways to sample a network. In the case of node-centered sampling design, like *random node sampling*, the main difficulty to prove consistency and asymptotic normality is the dependency between the r_{ij} variables. Indeed, in random node sampling, the variable $r_{i_0 j_0}$ depends on all r_{ij_0} and $r_{i_0 j}$ (for all $i, j \in \mathcal{N}$). As a consequence, many results proved in this paper are not valid under *random node sampling*. NMAR sampling designs raises problem of their own: each design requires its own estimation procedure (Tabouy et al., 2019a) and therefore its own analysis. For example, even parameter estimation under the *double standard sampling* for binary networks mentioned in section 2.3 is still an unsolved problem: numerical experiments suggest that $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ and $\boldsymbol{\psi} = (\rho_0, \rho_1)$ are jointly identifiable but there is no formal proof.

8 Acknowledgment

The authors thank Pierre Barbillon (INRA-MIA, AgroParisTech), Julien chiquet (INRA-MIA, AgroParisTech), Stéphane Robin (INRA-MIA, AgroParisTech) and James Ridgway (CFM) for their helpful remarks and suggestions.

This work is supported by two public grants overseen by the French National research Agency (ANR): first as part of the « Investissement d'Avenir » program, through the « IDI 2017 » project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02, and second by the « EcoNet » project.



9 Supplementary

10 Technical results

10.1 Proof of proposition 15

Proof. Noticing that $N_i \sim \mathrm{Bin}(n-1,\rho)$, then $\mathbb{P}(N_i \geqslant 1) = 1-(1-\rho)^{n-1}$. As a consequence $\mathbb{P}(\overline{\Omega_{0,n}}) \leqslant \sum_i \mathbb{P}(N_i = 0) = n(1-\rho)^{n-1} \underset{n\to+\infty}{\longrightarrow} 0$, and $\mathbb{P}(\Omega_{0,n}) \underset{n\to+\infty}{\longrightarrow} 1$. Then $\mathbb{P}(\limsup(\overline{\Omega_{0,n}})) = 0$ by Borel-Cantelli's theorem (because $\sum_n \mathbb{P}(\overline{\Omega_{0,n}})$ converge), and as $\overline{\limsup \overline{\Omega_{0,n}}} = \overline{\bigcap_{n\geqslant 0}\bigcup_{q\geqslant n}\overline{\Omega_{0,n}}} = \bigcup_{n\geqslant 0}\bigcap_{q\geqslant n}\Omega_{0,n} = \liminf \Omega_{0,n}$, the result follow. □

10.2 Proof of lemma 1

Proof. Noticing that $\mathbb{E}[r_{ij}z_{iq}z_{j\ell}] = \rho\alpha_q\alpha_l$ and defining $q_{i,j}^{q,\ell} = r_{ij}z_{iq}z_{j\ell} - \rho\alpha_q\alpha_l$. By Hoeffding decomposition for U-statistics (see Hoeffding (1948))

$$U_n' = \frac{1}{n(n-1)}\sum_{i\neq j}(r_{ij}z_{iq}z_{j\ell} - \rho\alpha_q\alpha_l) = \frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n}\frac{1}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q_{\sigma(i),\sigma(i+\lfloor\frac{n}{2}\rfloor)}^{q,\ell}, \tag{10.1}$$

where for each permutation $\sigma\in\mathfrak{S}$, $\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q_{\sigma(i),\sigma(i+\lfloor\frac{n}{2}\rfloor)}^{q,\ell}$ is a sum of independent r.v. Then, for $\gamma > 0$ by Jensen's inequality and Hoeffding's lemma about bounded r.v.

$$\begin{aligned}\mathbb{E}\left[\exp(\gamma U_n')\right] &\leq \frac{1}{n!}\sum_{\sigma\in\mathfrak{S}_n}\mathbb{E}\exp\left(\frac{\gamma}{\lfloor\frac{n}{2}\rfloor}\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} q_{\sigma(i),\sigma(i+\lfloor\frac{n}{2}\rfloor)}^{q,\ell}\right)\\ &\leq \exp\left(\frac{\gamma^2}{8\lfloor\frac{n}{2}\rfloor}\right).\end{aligned}$$

Finally, using the same proof than Hoeffding's inequality allows us to conclude. □

10.3 Proof of proposition 16

Proof. Since $\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star) = (\hat{\alpha}_1(\mathbf{z}^\star),\ldots,\hat{\alpha}_g(\mathbf{z}^\star))$ is the sample mean of n i.i.d. multinomial random variables with parameters 1 and $\boldsymbol{\alpha}^\star$, a simple application of the central limit theorem (CLT) gives:

$$\Sigma_{\boldsymbol{\alpha}^\star,qq'} = \begin{cases}\alpha_q^\star(1-\alpha_q^\star) & \text{if} \quad q = q'\\ -\alpha_q^\star\alpha_{q'}^\star & \text{if} \quad q \neq q'\end{cases}$$

which proves Equation (3.2) where $\Sigma_{\boldsymbol{\alpha}^\star}$ is semi-definite positive of rank $\mathcal{Q}-1$.

Similarly, $\psi'(\widehat{\pi}_{q\ell}(\mathbf{z}^\star))$ is the average of $\sum_{i\neq j} r_{ij}z_{iq}^\star z_{j\ell}^\star$ i.i.d. random variables with mean $\psi'\left(\pi_{q\ell}^\star\right)$ and variance $\psi''\left(\pi_{q\ell}^\star\right)$. $\sum_{i\neq j} r_{ij}z_{iq}^\star z_{j\ell}^\star$ is itself random but thanks to lemma 1 : $\frac{1}{n(n-1)}\sum_{i\neq j} r_{ij}z_{iq}^\star z_{j\ell}^\star \xrightarrow[n\to+\infty]{\mathbb{P}} \rho\alpha_q^\star\alpha_l^\star$. Therefore, by Slutsky's lemma and the CLT for random sums of random variables Shanthikumar and Sumita (1984), we have:

$$\begin{aligned}&\sqrt{n(n-1)\rho\alpha^{\star q}\alpha^{\star\ell}}\left(\psi'\left(\widehat{\pi}_{q\ell}(\mathbf{z}^\star)\right) - \psi'(\pi^{\star q\ell})\right)\\ &= \sqrt{n(n-1)\rho\alpha_q^\star\alpha^{\star\ell}}\left(\frac{\sum_{i\neq j} y_{ij}r_{ij}z_{iq}^\star z_{j\ell}^\star}{\sum_{i\neq j} r_{ij}z_{iq}^\star z_{j\ell}^\star} - \psi'(\pi^{\star q\ell})\right)\\ &\xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\left(0,\psi''(\pi^{\star q\ell})\right)\end{aligned}$$

The differentiability of $(\psi')^{-1}$ and the delta method then gives:

$$\sqrt{n(n-1)}\left(\widehat{\pi}_{q\ell}\left(\mathbf{z}^\star\right)-\pi^{\star_{q\ell}}\right)\xrightarrow[n\to+\infty]{\mathcal{D}}\mathcal{N}\left(0,\frac{1}{\rho\alpha^{\star_q}\alpha^{\star_\ell}\psi''\left(\pi^{\star_{q\ell}}\right)}\right)$$

and the independence results from the independence of $\widehat{\pi}_{q\ell}\left(\mathbf{z}^\star\right)$ and $\widehat{\pi}_{q'\ell'}\left(\mathbf{z}^\star\right)$ as soon as $q\neq q'$ or $\ell\neq\ell'$, as they involve different sets of i.i.d. variables. □

10.4 Proof of proposition 17

Proof. By Taylor expansion,

$$\begin{aligned}
&\mathcal{L}^\star_{co}\left(\boldsymbol{\alpha}^\star+\frac{s}{\sqrt{n}},\boldsymbol{\pi}^\star+\frac{u}{\sqrt{n(n-1)}}\right)\\
=\;&\mathcal{L}^\star_{co}\left(\boldsymbol{\theta}^\star\right)+\frac{1}{\sqrt{n}}s^T\nabla\mathcal{L}^\star_{co\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^\star\right)+\frac{1}{\sqrt{n(n-1)}}\mathrm{Tr}\left(u^T\nabla\mathcal{L}^\star_{co\boldsymbol{\pi}}\left(\boldsymbol{\theta}^\star\right)\right)\\
&+\frac{1}{n}s^T\mathbf{H}_{\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^\star\right)s+\frac{1}{n(n-1)}\mathrm{Tr}\left((u\odot u)^T\mathbf{H}_{\boldsymbol{\pi}}\left(\boldsymbol{\theta}^\star\right)\right)+o_P(1)
\end{aligned}$$

where $\nabla\mathcal{L}^\star_{co\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^\star\right)$ and $\nabla\mathcal{L}^\star_{co\boldsymbol{\pi}}\left(\boldsymbol{\theta}^\star\right)$ denote the respective components of the gradient of $\mathcal{L}^\star_{co}$ evaluated at $\boldsymbol{\theta}^\star$ and $\mathbf{H}_{\boldsymbol{\alpha}}$ and $\mathbf{H}_{\boldsymbol{\pi}}$ denote the conditional hessian of $\mathcal{L}^\star_{co}$ evaluated at $\boldsymbol{\theta}^\star$. By inspection, $\mathbf{H}_{\boldsymbol{\alpha}}/n$ and $\mathbf{H}_{\boldsymbol{\pi}}/(n(n-1))$ converge in probability to constant matrices Σ_α,Σ_π and the random vectors $\nabla\mathcal{L}^\star_{co\boldsymbol{\alpha}}\left(\boldsymbol{\theta}^\star\right)/\sqrt{n}$ and $\nabla\mathcal{L}^\star_{co\boldsymbol{\pi}}\left(\boldsymbol{\theta}^\star\right)/\sqrt{n(n-1)}$ converge in distribution by central limit theorem. □

10.5 Proof of proposition 18

Proof. In regular configurations, each group has $\Omega(n)$ members, where $u_n=\Omega(n)$ if there exists two constant $a,b>0$ such that for n enough large $an\leq u_n\leq bn$. $c/2$-regular assignments, with c defined in Assumption H_1, have high $\mathbb{P}_{\boldsymbol{\theta}^\star}$-probability in the space of all assignments, uniformly over all $\boldsymbol{\theta}^\star\in\boldsymbol{\Theta}$.

Each z_{+q} is a sum of n i.i.d Bernoulli r.v. with parameter $\alpha_q\geq\alpha_{\min}\geq c$. A simple Hoeffding bound shows that

$$\mathbb{P}_{\boldsymbol{\theta}^\star}\left(z_{+q}\leq n\frac{c}{2}\right)\leq\mathbb{P}_{\boldsymbol{\theta}^\star}\left(z_{+q}\leq n\frac{\alpha_q}{2}\right)\leq\exp\left(-2n\left(\frac{\alpha_q}{2}\right)^2\right)\leq\exp\left(-\frac{nc^2}{2}\right)$$

taking a union bound over $\mathcal{Q}$ values of q leads to Proposition 18. □

11 Main Results

11.1 Proof of proposition 19)

Proof. First of all we will prove equation 5.3,

$$\begin{aligned}
\bar{y}_{q\ell}(\mathbf{z}) &= \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\frac{\sum_{i\neq j}z_{iq}z_{j\ell}r_{ij}y_{ij}}{\sum_{i\neq j}z_{iq}z_{j\ell}r_{ij}}\middle|\mathbf{z}^\star\right]\\
&= \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\mathbb{E}_{\boldsymbol{\theta}^\star}\left[\frac{\sum_{i\neq j}z_{iq}z_{j\ell}r_{ij}y_{ij}}{\sum_{i\neq j}z_{iq}z_{j\ell}r_{ij}}\middle|R,\mathbf{z}^\star\right]\middle|\mathbf{z}^\star\right]\\
&= \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\frac{\sum_{i\neq j}z_{iq}z_{j\ell}r_{ij}S^\star_{Z_iZ_j}}{\sum_{i\neq j}z_{iq}z_{j\ell}r_{ij}}\middle|\mathbf{z}^\star\right],
\end{aligned}$$

where $Z_i = q \Leftrightarrow z_{iq} = 1$. Noticing that the (i,j) for which $z_{iq}z_{j\ell} = 0$ does not contributes in any of the two terms of the ratio. The calculus of this expectation is then equivalent to calculate an expectation of the general form $\mathbb{E}_{\boldsymbol{\theta}^\star}\left[\frac{\sum_{i=1}^n a_i R_i}{\sum_{i=1}^n R_i}\right]$, $(a_i)_{i\in\{1,..,n\}} \in \mathbb{R}^n$ and $T_i \overset{iid}{\sim} \mathcal{B}(\rho)$.

Lemme 2.

$$\mathbb{E}_{\boldsymbol{\theta}^\star}\left[\frac{\sum_{i=1}^n a_i T_i}{\sum_{i=1}^n T_i}\right] = \frac{\sum_{i=1}^n a_i}{n}.$$

Proof. Define $N = \sum_{i=1}^n T_i$ and noticing that $\mathbb{E}[T_i|N=k] = \frac{k}{n}$. Conditionally to $N \geq 1$

$$\begin{aligned}\mathbb{E}\left[\frac{\sum_{i=1}^n a_i T_i}{\sum_{i=1}^n T_i}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n a_i T_i}{N}\middle| N\right]\right] \\ &= \frac{\sum_{i=1}^n a_i}{n}.\end{aligned}$$

□

Now, applying lemma 2 with $N^o_{q\ell}(z) = \sum_{i\neq j} z_{iq}z_{j\ell}r_{ij}$ leads to

$$\mathbb{E}_{\boldsymbol{\theta}^\star}\left[\frac{\sum_{i\neq j} z_{iq}z_{j\ell}r_{ij}S^\star_{Z_iZ_j}}{\sum_{i\neq j} z_{iq}z_{j\ell}r_{ij}}\middle| \mathbf{z}^\star, N^o_{q\ell}(z) \geq 1\right] = \frac{\left[\mathbb{R}(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}(\mathbf{z})\right]_{q\ell}}{\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})}\mathbb{1}_{N^o_{q\ell}(z)\geq 1}.$$

Finally, $\mathbb{E}_{\boldsymbol{\theta}^\star}[\widehat{y}_{q\ell}(\mathbf{z})|\mathbf{z}^\star, N^o_{q\ell}(z) = 0]$ can be arbitrarily defined at the same value than $\mathbb{E}_{\boldsymbol{\theta}^\star}[\widehat{y}_{q\ell}(\mathbf{z})|\mathbf{z}^\star, N^o_{q\ell}(z) \geq 1]$ which conclude the proof. □

11.2 Proof of proposition 20

Proof. Defining $\nu(y,\pi) = y\pi - \psi(\pi)$. For y fixed, $\nu(y,\pi)$ is maximized at $\pi = (\psi')^{-1}(y)$. Manipulations yield

$$\begin{aligned}&LR(\boldsymbol{\theta},\mathbf{z}) = \log p(\mathbf{y^o};\mathbf{z},\boldsymbol{\theta}) - \log p(\mathbf{y^o};\mathbf{z}^\star,\boldsymbol{\theta}^\star) \\ &= \left[\sum_q\sum_\ell N^o_{q\ell}(z)\nu(\widehat{y}_{q\ell}(\mathbf{z}),\pi_{q\ell}) - \sum_q\sum_\ell N^o_{q\ell}(z^\star)\nu(\widehat{y}_{q\ell}(\mathbf{z}^\star),\pi^\star_{q\ell})\right]\end{aligned}$$

which is maximized at $\pi_{q\ell} = (\psi')^{-1}(\widehat{y}_{q\ell}(\mathbf{z}))$. Similarly with $N_{q\ell}(z) = \sum_{i\neq j} z_{iq}z_{j\ell}$,

$$\begin{aligned}&ELR(\boldsymbol{\theta},\mathbf{z}) = \mathbb{E}_{\boldsymbol{\theta}^\star}[\log p(\mathbf{y^o};\mathbf{z},\boldsymbol{\theta}) - \log p(\mathbf{y^o};\mathbf{z}^\star,\boldsymbol{\theta}^\star)|\mathbf{z}^\star] \\ &= \rho\left[\sum_q\sum_\ell N_{q\ell}(z)\nu(\bar{y}_{q\ell}(\mathbf{z}),\pi_{q\ell}) - \sum_q\sum_\ell N_{q\ell}(z^\star)\nu(\psi'(\pi^\star_{q\ell}),\pi^\star_{q\ell})\right]\end{aligned}$$

is maximized at $\pi_{q\ell} = (\psi')^{-1}(\bar{y}_{q\ell}(\mathbf{z}))$. □

11.3 Proof of Proposition 22 (maximum of ELR and $\tilde{\Lambda}$)

Proof. We condition on $\mathbf{z}^\star$ and prove Equation (5.5):

$$\begin{aligned}
ELR(\boldsymbol{\theta}, \mathbf{z}) &= \mathbb{E}_{\boldsymbol{\theta}^\star}\left[\log \frac{p(\mathbf{y}^\mathbf{o}; \mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{y}^\mathbf{o}; \mathbf{z}^\star, \boldsymbol{\theta}^\star)}\middle| \mathbf{z}^\star\right] \\
&= \sum_i \sum_j \sum_{q,q'} \sum_{\ell,\ell'} \mathbb{E}_{\boldsymbol{\theta}^\star}\left[y_{ij}(\pi_{q'\ell'} - \pi^\star_{q\ell}) - (\psi(\pi_{q'\ell'}) - \psi(\pi^\star_{q\ell}))\right] \rho z^\star_{iq} z_{iq'} z^\star_{j\ell} z_{j\ell'} \\
&= n^2 \rho \sum_{q,q'} \sum_{\ell,\ell'} \mathbb{R}(\mathbf{z})_{q,q'} \mathbb{R}(\mathbf{z})_{\ell,\ell'} \left[\psi'(\pi^\star_{q\ell})(\pi_{q'\ell'} - \pi^\star_{q\ell}) + \psi(\pi^\star_{q\ell}) - \psi(\pi_{q'\ell'})\right] \\
&= -n^2 \rho \sum_{q,q'} \sum_{\ell,\ell'} \mathbb{R}(\mathbf{z})_{q,q'} \mathbb{R}(\mathbf{z})_{\ell,\ell'} \mathcal{KL}(\pi^{\star q\ell}, \pi_{q'\ell'})
\end{aligned}$$

If $\mathbf{z}^\star$ is regular, and for $n > 2/c$, all the rows of $\mathbb{R}(\mathbf{z})$ have at least one positive element and we can apply Lemma 3.2 of Bickel et al. (2013) to characterize the maximum for ELR.

The maximality of $\tilde{\Lambda}(\mathbf{z}^\star)$ results from the fact that $\tilde{\Lambda}(\mathbf{z}) = ELR(\bar{\boldsymbol{\pi}}(\mathbf{z}), \mathbf{z})$ where $\bar{\boldsymbol{\pi}}(\mathbf{z})$ is a particular value of $\boldsymbol{\pi}$, $\tilde{\Lambda}$ is immediately maximum at $\mathbf{z} \sim \mathbf{z}^\star$, and for those, we have $\bar{\boldsymbol{\pi}}(\mathbf{z}) \sim \boldsymbol{\pi}^\star$.

The separation and local behavior of G around $\mathbf{z}^\star$ is a direct consequence of the proposition 21. □

11.4 Proof of Proposition 21 (Local upper bound for $\tilde{\Lambda}$)

Proof. We work conditionally on $\mathbf{z}^\star$. The principle of the proof relies on the extension of $\tilde{\Lambda}$ to a continuous subspace of $\mathcal{M}_\mathcal{Q}([0,1])$, in which the confusion matrix is naturally embedded. The regularity assumption allows us to work on a subspace that is bounded away from the borders of $\mathcal{M}_\mathcal{Q}([0,1])$. The proof then proceeds by computing the gradient of $\tilde{\Lambda}$ at and around its argmax and using those gradients to control the local behavior of $\tilde{\Lambda}$ around its argmax. The local behavior allows us in turn to show that $\tilde{\Lambda}$ is well-separated.

Note that $\tilde{\Lambda}$ only depends on $\mathbf{z}$ through $\mathbb{R}(\mathbf{z})$. We can therefore extend it to matrix $U \in \mathcal{U}_c$ where $\mathcal{U}$ is the subset of matrices $\mathcal{M}_\mathcal{Q}([0,1])$ with each row sum higher than $c/2$.

$$\tilde{\Lambda}(U) = -\rho n^2 \sum_{q,q'} \sum_{\ell,\ell'} U_{qq'} U_{\ell\ell'} \mathcal{KL}\left(\pi^\star_{q\ell}, \bar{\pi}_{q'\ell'}\right)$$

where

$$\bar{\pi}_{q\ell} = \bar{\pi}_{q\ell}(U) = (\psi')^{-1}\left(\frac{\left[U^T \boldsymbol{S}^\star U\right]_{q\ell}}{\left[U^T \mathbf{1} U\right]_{q\ell}}\right)$$

and $\mathbf{1}$ is the $\mathcal{Q} \times \mathcal{Q}$ matrix filled with 1. Confusion matrix $\mathbb{R}(\mathbf{z})$ satisfy $\mathbb{R}(\mathbf{z})\mathbb{1} = \boldsymbol{\alpha}(\mathbf{z}^\star)$, with $\mathbb{1} = (1, \ldots, 1)^T$ a vector only containing 1 values, and are obviously in $\mathcal{U}_c$ as soon as $\mathbf{z}^\star$ is $c/2$ regular.

The maps $f_{q,q',\ell,\ell'} : (U) \mapsto KL(\pi^\star_{q\ell}, \bar{\pi}_{q\ell}(U))$ are twice differentiable with second derivatives bounded over $\mathcal{U}_c$ and therefore so is $\tilde{\Lambda}(U)$. Tedious but straightforward computations show that the derivative of $\tilde{\Lambda}$ at $D_\alpha := \mathrm{Diag}(\boldsymbol{\alpha}(\mathbf{z}^\star))$ is:

$$A_{qq'}(\mathbf{z}^\star) := \frac{-1}{n^2} \frac{\partial \tilde{\Lambda}}{\partial U_{qq'}}(D_\alpha) = 2\rho \sum_\ell \alpha_\ell(\mathbf{z}^\star) \mathcal{KL}\left(\pi^\star_{q\ell}, \pi^\star_{q'\ell}\right)$$

$A(\mathbf{z}^\star)$ is the matrix-derivative of $-\tilde{\Lambda}/n^2$ at D_α. Since $\mathbf{z}^\star$ is $c/2$-regular and by definition of $\delta(\boldsymbol{\pi}^\star)$, $A(\mathbf{z}^\star)_{qq'} \geq c\rho\delta(\boldsymbol{\pi}^\star)$ if $q \neq q'$ and $A(\mathbf{z}^\star)_{qq} = 0$ for all q. By boundedness of the second derivative, there exists $C > 0$ such that for all D_α and all $H \in B(D_\alpha, C)$, we have:

$$\frac{-1}{n^2} \frac{\partial \tilde{\Lambda}}{\partial U_{qq'}}(H) \begin{cases} \geq \rho \frac{7c\delta(\boldsymbol{\pi}^\star)}{8} & \text{if } q \neq q' \\ \leq \rho \frac{c\delta(\boldsymbol{\pi}^\star)}{8} & \text{if } q = q' \end{cases}$$

Choose U in $\mathcal{U}_c \cap B(D_\alpha, C)$ satisfying $U\mathbb{1} = \boldsymbol{\alpha}(\mathbf{z}^\star)$. $U - D_\alpha$ have non-negative off diagonal coefficients and negative diagonal coefficients. Furthermore, the coefficients of U, D_α sum up to 1 and $\mathrm{Tr}(D_\alpha) = 1$. By Taylor expansion, there exists H also in $\mathcal{U}_c \cap B(D_\alpha, C)$ such that

$$\begin{aligned}\frac{-1}{n^2}\tilde{\Lambda}(U) &= \frac{-1}{n^2}\tilde{\Lambda}(D_\alpha) + \mathrm{Tr}\left((U - D_\alpha)\frac{-1}{n^2}\frac{\partial\tilde{\Lambda}}{\partial U}(H)\right) \\ &\geq \rho\frac{c\delta(\boldsymbol{\pi}^\star)}{8}[7\sum_{q\neq q'}(U - D_\alpha)_{qq'} - \sum_q (U - D_\alpha)_{qq} \\ &= c\rho\frac{3\delta(\boldsymbol{\pi}^\star)}{4}(1 - \mathrm{Tr}(U))\end{aligned}$$

To conclude the proof, assume without loss of generality that $\mathbf{z} \in S(\mathbf{z}^\star, C)$ achieves the $\|.\|_{0,\sim}$ norm (i.e. it is the closest to $\mathbf{z}^\star$ in its representative class). Then $U = \mathbb{R}(\mathbf{z})$ is in $(\mathcal{U}_c \cap B(D_\alpha, C)$ and satisfy $U\mathbb{1} = \boldsymbol{\alpha}(\mathbf{z}^\star)$. We just need to note $n(1 - \mathrm{Tr}(\mathbb{R}(\mathbf{z}))) = \|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}$ to end the proof.

□

11.5 Proof of Proposition 23 (global convergence LR)

Proof. Conditionally upon $\mathbf{z}^\star$,

$$\begin{aligned}LR(\boldsymbol{\theta}, \mathbf{z}) - \tilde{\Lambda}(\mathbf{z}) &\leq LR(\boldsymbol{\theta}, \mathbf{z}) - ELR(\boldsymbol{\theta}, \mathbf{z}) \\ &= \sum_i\sum_j (\pi_{z_i z_j} - \pi^{\star_{z_i^\star z_j^\star}})\left(y_{ij}r_{ij} - \psi'(\pi^{\star_{z_i^\star z_j^\star}})\rho\right) \\ &\quad + \sum_i\sum_j (\psi(\pi_{z_i z_j}) - \psi(\pi^{\star_{z_i^\star z_j^\star}}))(r_{ij} - \rho) \\ &= \sum_{qq'}\sum_{\ell\ell'} (\pi_{q'\ell'} - \pi^{\star_{q\ell}}) W_{qq'\ell\ell'} \\ &\leq \sup_{\substack{\Gamma\in\mathbb{R}^{\mathcal{Q}^2\times\mathcal{Q}^2} \\ \|\Gamma\|_\infty \leq \mathrm{Diam}(\boldsymbol{\Theta})}} \sum_{qq'}\sum_{\ell\ell'} \Gamma_{qq'\ell\ell'} W_{qq'\ell\ell'} := Z\end{aligned}$$

uniformly in $\boldsymbol{\theta}$, where the $W_{qq'\ell\ell'}$ are independent and by Taylor expansion defined by:

$$W_{qq'\ell\ell'} = \sum_i\sum_j z_{iq}^\star z_{j\ell}^\star z_{i,q'} z_{j\ell'} \left(y_{ij}r_{ij} - \psi'(\pi^{\star_{q\ell}})\rho - (r_{ij} - \rho)C_{qq'\ell\ell'}\right), \quad C_{qq'\ell\ell'} \in \psi'(\Theta)$$

is the sum of $n^2\mathbb{R}(\mathbf{z})_{qq'}\mathbb{R}(\mathbf{z})_{\ell\ell'}$ sub-exponential variables with parameters $(\nu^2, 1/b)$ and is therefore itself sub-exponential with parameters $(n^2\mathbb{R}(\mathbf{z})_{qq'}\mathbb{R}(\mathbf{z})_{\ell\ell'}\nu^2, 1/b)$. According to Proposition B.3 of Brault et al. (2017) , $\mathbb{E}_{\boldsymbol{\theta}^\star}[Z|\mathbf{z}^\star] \leq Q^2\,\mathrm{Diam}(\boldsymbol{\Theta})\sqrt{n^2\nu^2}$ and Z is sub-exponential with parameters $(n^2\,\mathrm{Diam}(\boldsymbol{\Theta})^2(2\sqrt{2})^2\nu^2, 2\sqrt{2}\,\mathrm{Diam}(\boldsymbol{\Theta})/b)$. In particular, for all $\varepsilon_n < \nu b$

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}^\star}\left(Z \geq \nu\mathcal{Q}^2\,\mathrm{Diam}(\boldsymbol{\Theta})n\left\{1 + \frac{\sqrt{8n^2}\varepsilon_n}{\mathcal{Q}^2}\right\}\middle|\mathbf{z}^\star\right) \\ &\qquad \leq \mathbb{P}_{\boldsymbol{\theta}^\star}\left(Z \geq \mathbb{E}_{\boldsymbol{\theta}^\star}[Z|\mathbf{z}^\star] + \nu\,\mathrm{Diam}(\boldsymbol{\Theta})n^2 2\sqrt{2}\varepsilon_n\middle|\mathbf{z}^\star\right) \\ &\qquad \leq \exp\left(-\frac{n^2\varepsilon_n^2}{2}\right)\end{aligned}$$

We can then remove the conditioning and take a union bound. □

11.6 Proof of Proposition 24 (contribution of far away assignments)

Proof. Conditionally on $\mathbf{z}^\star$, we know from proposition 22 that $\tilde{\Lambda}$ is maximal in $\mathbf{z}^\star$ and its equivalence class. Choose $0 < t_n$ decreasing to 0 but satisfying $\frac{n\rho t_n}{\sqrt{\log(n)}} \to +\infty$. According to 22 (iii), for all $\mathbf{z} \notin S(\mathbf{z}^\star, t_n)$

$$\tilde{\Lambda}(\mathbf{z}) \leq -c\rho n \frac{3\delta(\boldsymbol{\pi}^\star)}{4} \|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \leq -c\rho \frac{3\delta(\boldsymbol{\pi}^\star)}{4} n^2 t_n \tag{11.1}$$

since $\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \geq n t_n$.

Set $\varepsilon_n = \inf(5c\rho\delta(\boldsymbol{\pi}^\star)t_n/(\sqrt{2}\nu \operatorname{Diam}(\boldsymbol{\Theta})), \nu b)$ and n large enough that $\epsilon_n \geq \frac{Q^2}{n\sqrt{8}}$. By proposition 23, and with our choice of ε_n, with probability higher than $1 - \Delta_n^1(\varepsilon_n)$,

$$\begin{aligned}
&\sum_{\mathbf{z} \notin S(\mathbf{z}^\star, t_n)} p(\mathbf{y}^\mathbf{o}, \mathbf{z}; \boldsymbol{\theta}) \\
&= p(\mathbf{y}^\mathbf{o} | \mathbf{z}^\star, \boldsymbol{\theta}^\star) \sum_{\mathbf{z} \notin S(\mathbf{z}^\star, t_n)} p(\mathbf{z}; \boldsymbol{\theta}) e^{LR(\boldsymbol{\theta}, \mathbf{z}) - \tilde{\Lambda}(\mathbf{z}) + \tilde{\Lambda}(\mathbf{z})} \\
&\leq p(\mathbf{y}^\mathbf{o} | \mathbf{z}^\star, \boldsymbol{\theta}^\star) \sum_{\mathbf{z}} p(\mathbf{z}; \boldsymbol{\theta}) e^{LR(\boldsymbol{\theta}, \mathbf{z}) - \tilde{\Lambda}(\mathbf{z}) - 3n^2 t_n c\rho\delta(\boldsymbol{\pi}^\star)/4} \\
&\leq p(\mathbf{y}^\mathbf{o} | \mathbf{z}^\star, \boldsymbol{\theta}^\star) \sum_{\mathbf{z}} p(\mathbf{z}; \boldsymbol{\theta}) e^{-n^2 t_n c\rho\delta(\boldsymbol{\pi}^\star)/8} \\
&= \frac{p(\mathbf{y}^\mathbf{o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)}{p(\mathbf{z}^\star; \boldsymbol{\theta}^\star)} e^{-n^2 t_n c\rho\delta(\boldsymbol{\pi}^\star)/8} \\
&\leq p(\mathbf{y}^\mathbf{o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star) \exp\left(-n^2 t_n \frac{c\rho\delta(\boldsymbol{\pi}^\star)}{8} + n \log \frac{1}{c}\right) \\
&= p(\mathbf{y}^\mathbf{o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star) o(1)
\end{aligned}$$

where the second line comes from inequality (11.1), the third from the global control studied in Proposition 23 and the definition of ε_n, the fourth from the definition of $p(\mathbf{y}^\mathbf{o}, \mathbf{z}^\star; \boldsymbol{\theta}^\star)$, the fifth from the bounds on $\boldsymbol{\alpha}^\star$ and the last from $\frac{n\rho t_n}{\sqrt{\log(n)}} \to +\infty$.

In addition, with our choice of t_n, we have $\varepsilon_n \gg \sqrt{\log(n)}/n$ so that the series $\sum_n \Delta_n^1(\varepsilon_n)$ converges and:

$$\sum_{\mathbf{z} \notin S(\mathbf{z}^\star, t_{nd})} p(\mathbf{y}^\mathbf{o}, \mathbf{z}; \boldsymbol{\theta}) = p(\mathbf{y}^\mathbf{o}; \mathbf{z}^\star, \boldsymbol{\theta}^\star) o_P(1)$$

□

11.7 Proof of Proposition 25 (local convergence LR)

Proof. We work conditionally on $\mathbf{z}^\star \in \mathcal{Z}_1$. Choose $\varepsilon \leq \kappa\underline{\sigma}^2$ small. Assignments $\mathbf{z}$ at $\|.\|_{0,\sim}$-distance less than $c/4$ of $\mathbf{z}^\star$ are $c/4$-regular. According to Proposition B.1 of Brault et al. (2017) , $\widehat{y}_{q\ell}$ and $\bar{y}_{q\ell}$ are at distance at most ε with probability higher than $1 - 2\exp\left(-\frac{n^2c^2\varepsilon^2}{32(\nu^2 + b^{-1}\varepsilon)}\right)$. Defining

$$\tilde{\tilde{\Lambda}}(\mathbf{z}) = \sum_q \sum_\ell N_{q\ell}^o(\mathbf{z}) \nu(\bar{y}_{q\ell}(\mathbf{z}), \pi_{q\ell}) - \sum_q \sum_\ell N_{q\ell}^o(\mathbf{z}^\star) \nu(\psi'(\pi_{q\ell}^\star), \pi_{q\ell}^\star),$$

where $\tilde{\Lambda}(\mathbf{z}) = \mathbb{E}\left[\tilde{\tilde{\Lambda}}(\mathbf{z})|\mathbf{z}^\star\right]$. Manipulation of Λ, $\tilde{\Lambda}$ and $\tilde{\tilde{\Lambda}}$ yield

$$\begin{aligned}\frac{\Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z})}{n^2} \leq & \frac{\Lambda(\mathbf{z}) - \tilde{\tilde{\Lambda}}(\mathbf{z})}{n^2} + \frac{\tilde{\tilde{\Lambda}}(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z})}{n^2} \\ = & \frac{1}{n^2}\sum_q\sum_\ell \left(N^o_{q\ell}(\mathbf{z})\left[f(\widehat{y}_{q\ell}) - f(\bar{y}_{q\ell})\right] - N^o_{q\ell}(\mathbf{z}^\star)\pi^\star_{q\ell}(\widehat{y}^\star_{q\ell} - \bar{y}^\star_{q\ell})\right) \\ & + \frac{1}{n^2}\sum_q\sum_\ell f(\bar{y}^\star_{q\ell})\underbrace{\left[N^o_{q\ell}(\mathbf{z}) - N^o_{q\ell}(\mathbf{z}^\star) - \rho(N_{q\ell}(\mathbf{z}) - N_{q\ell}(\mathbf{z}^\star))\right]}_{=A_{q\ell}} \\ & + \frac{1}{n^2}\sum_q\sum_\ell [N^o_{q\ell}(\mathbf{z}) - \rho N_{q\ell}(\mathbf{z})](f(\bar{y}^\star_{q\ell}) - f(\bar{y}_{q\ell}))\end{aligned}$$

where $f(x) = x(\psi')^{-1}(x) - \psi \circ (\psi')^{-1}(x)$, $\widehat{y}^\star_{q\ell} = \widehat{y}_{q\ell}(\mathbf{z}^\star)$ and $\bar{y}^\star_{q\ell} = \psi'(\pi^\star_{q\ell})$.

Concerning the first term. The function f is twice differentiable on $\mathring{\mathcal{A}}$ with $f'(x) = (\psi')^{-1}(x)$ and $f''(x) = 1/\psi'' \circ (\psi')^{-1}(x)$. f' (resp. f'') are bounded over $I = \psi'(C_\pi)$ by C_π (resp. $1/\underline{\sigma}^2$) so that:

$$f(\widehat{y}_{q\ell}) - f(\bar{y}_{q\ell}) = f'(\bar{y}_{q\ell})\left(\widehat{y}_{q\ell} - \bar{y}_{q\ell}\right) + \Omega\left((\widehat{y}_{q\ell} - \bar{y}_{q\ell})^2\right)$$

By Proposition B.1 (adapted for SBM) of Brault et al. (2017) , $(\widehat{y}_{q\ell} - \bar{y}_{q\ell})^2 = \mathcal{O}_P(1/n^2)$ where the $\mathcal{O}_P$ is uniform in $\mathbf{z}$ and does not depend on $\mathbf{z}^\star$. Similarly,

$$f'(\bar{y}_{q\ell}) = f'(\bar{y}^\star_{q\ell}) + \Omega(\bar{y}_{q\ell} - \bar{y}^\star_{q\ell}) = \pi^\star_{q\ell} + \Omega(\bar{y}_{q\ell} - \bar{y}^\star_{q\ell})$$

$\bar{y}_{q\ell}$ is a convex combination of the $S^\star_{q\ell} = \psi'(\pi^\star_{q\ell})$ therefore,

$$\begin{aligned}|\bar{y}_{q\ell} - \bar{y}^\star_{q\ell}| &= \left|\frac{\left[\mathbb{R}(\mathbf{z})^T \boldsymbol{S}^\star \mathbb{R}(\mathbf{z})\right]_{q\ell}}{\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})} - \bar{y}^\star_{q\ell}\right| \\ &\leq \left(1 - \frac{\mathbb{R}(\mathbf{z})_{qq}\mathbb{R}(\mathbf{z})_{\ell\ell}}{\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})}\right)(S^\star_{\max} - S^\star_{\min})\end{aligned}$$

Note that:

$$\begin{aligned}\sum_{q,\ell} N^o_{q\ell}(\mathbf{z})\left(1 - \frac{\mathbb{R}(\mathbf{z})_{qq}\mathbb{R}(\mathbf{z})_{\ell\ell}}{\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})}\right) &= n^2\rho(1 + o_P(1))\sum_{q,\ell}[1 - \mathbb{R}(\mathbf{z})_{qq}\mathbb{R}(\mathbf{z})_{\ell\ell}] \\ &= n^2\rho(1 + o_P(1))[1 - \mathrm{Tr}(\mathbb{R}(\mathbf{z}))^2] \\ &\leq n\rho(1 + o_P(1))2\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}\end{aligned}$$

and $\widehat{y}_{q\ell} - \bar{y}_{q\ell} = o_P(1)$. Therefore

$$\frac{1}{n^2}\sum_{q,\ell} N^o_{q\ell}(\mathbf{z})\Omega(\bar{y}_{q\ell} - \bar{y}^\star_{q\ell}) \times (\widehat{y}_{q\ell} - \bar{y}_{q\ell}) = o_P\left(\frac{\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}}{n}\right)$$

The remaining term writes

$$\frac{1}{n^2}\sum_{q,\ell}\pi^\star_{q\ell}\left[N^o_{q\ell}(\mathbf{z})(\widehat{y}_{q\ell} - \bar{y}_{q\ell}) - N^o_{q\ell}(\mathbf{z}^\star)(\widehat{y}^\star_{q\ell} - \bar{y}^\star_{q\ell})\right]$$

and is also $o_P((\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}/n)$ uniformly in $\mathbf{z}$ and $\mathbf{z}^\star \in \Omega_1$ by Proposition 28.

Concerning the second term. For all q, ℓ, defining

$$\begin{cases} N^+_{q\ell}(\mathbf{z},\mathbf{z}^\star) = n^2 \sum_{q'} \mathbb{R}(\mathbf{z})_{qq'}(\mathbf{z}) \sum_{\ell'} \mathbb{R}(\mathbf{z})_{\ell\ell'}(\mathbf{z}) - n^2 \mathbb{R}(\mathbf{z})_{qq}\mathbb{R}(\mathbf{z})_{\ell\ell} \\ N^-_{q\ell}(\mathbf{z},\mathbf{z}^\star) = n^2 \sum_{q} \mathbb{R}(\mathbf{z})_{qq'}(\mathbf{z}) \sum_{\ell} \mathbb{R}(\mathbf{z})_{\ell\ell'}(\mathbf{z}) - n^2 \mathbb{R}(\mathbf{z})_{qq}\mathbb{R}(\mathbf{z})_{\ell\ell} \end{cases}$$

and noticing that $N^+_{q\ell}(\mathbf{z},\mathbf{z}^\star) = \#\{(i,j) : z_{iq} = 1, z_{j\ell} = 1, (z_{q\ell}, z_{j\ell}) \neq (z^\star_{q\ell}, z^\star_{j\ell})\}$ and $N^-_{q\ell}(\mathbf{z},\mathbf{z}^\star) = \#\{(i,j) : z^\star_{iq} = 1, z^\star_{j\ell} = 1, (z_{q\ell}, z_{j\ell}) \neq (z^\star_{q\ell}, z^\star_{j\ell})\}$. Using the following notations

$$\hat{\rho}^+_{q\ell} = \frac{1}{N^+_{q\ell}(\mathbf{z},\mathbf{z}^\star)} \sum_{(i,j)\in N^+_{q\ell}(\mathbf{z},\mathbf{z}^\star)} R_{ij}, \quad \hat{\rho}^-_{q\ell} = \frac{1}{N^-_{q\ell}(\mathbf{z},\mathbf{z}^\star)} \sum_{(i,j)\in N^-_{q\ell}(\mathbf{z},\mathbf{z}^\star)} R_{ij}$$

we are able to write

$$\begin{aligned} A_{q\ell} = & \sum_{\substack{i<j \\ z_{iq}=1, z_{j\ell}=1}} (R_{ij} - \rho) - \sum_{\substack{i<j \\ z^\star_{iq}=1, z^\star_{j\ell}=1}} (R_{ij} - \rho) \\ = & N^+_{q\ell}(\mathbf{z},\mathbf{z}^\star)(\hat{\rho}^+_{q\ell} - \rho) - N^-_{q\ell}(\mathbf{z},\mathbf{z}^\star)(\hat{\rho}^-_{q\ell} - \rho). \end{aligned}$$

Where the second equality is the sum of independent random variables.
Note that :

$$\begin{aligned} \sum_{q\ell} N^+_{q\ell}(\mathbf{z},\mathbf{z}^\star) &= \sum_{q\ell} N^-_{q\ell}(\mathbf{z},\mathbf{z}^\star) \\ &= n^2 \sum_{q,\ell} [1 - \mathbb{R}(\mathbf{z})_{qq}\mathbb{R}(\mathbf{z})_{\ell\ell}] \\ &= n^2 [1 - \mathrm{Tr}(\mathbb{R}(\mathbf{z}))^2] \\ &\leq n2\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim} \end{aligned}$$

also that $\hat{\rho}^+_{q\ell} - \rho = o_P(1)$ and $\hat{\rho}^-_{q\ell} - \rho = o_P(1)$. Therefore

$$\frac{1}{n^2} \sum_q \sum_\ell f(\bar{y}^\star_{q\ell}) A_{q\ell} = o_P\left(\frac{\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}}{n}\right).$$

Concerning the third term. Using arguments developed previously leads to the same conclusion than before :

$$\frac{1}{n^2} \sum_q \sum_\ell [N^o_{q\ell}(\mathbf{z}) - \rho N_{q\ell}(\mathbf{z})](f(\bar{y}^\star_{q\ell}) - f(\bar{y}_{q\ell})) = o_P\left(\frac{\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}}{n}\right).$$

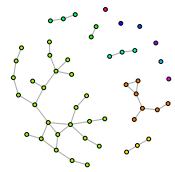
As a conclusion, writing

$$\sup_{\mathbf{z}\nsim\mathbf{z}^\star} \frac{\Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}^\star)}{n\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}} = \sup_{\mathbf{z}\nsim\mathbf{z}^\star} \left(\frac{\Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z})}{n\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}} + \frac{\tilde{\Lambda}(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}^\star)}{n\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}} \right)$$

and noticing that $\frac{\tilde{\Lambda}(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}^\star)}{n\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}} \leq 0$ since $\tilde{\Lambda}$ is maximized in $\mathbf{z}^\star$ (see 22). We have

$$\sup_{\mathbf{z}\nsim\mathbf{z}^\star} \frac{\Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}^\star)}{n\|\mathbf{z} - \mathbf{z}^\star\|_{0,\sim}} = o_P(1).$$

□



11.8 Proof of Proposition 26 (contribution of local assignments)

Proof. By Proposition 18, it is enough to prove that the sum is small compared to $p(\mathbf{z}^\star, \mathbf{y}^\mathbf{o}; \boldsymbol{\theta}^\star)$ on Ω_1. We work conditionally on $\mathbf{z}^\star \in \mathcal{Z}_1$. Choose $\mathbf{z}$ in $S(\mathbf{z}^\star, C)$ with C defined in proposition 24.

$$\log\left(\frac{p(\mathbf{z}, \mathbf{y}^\mathbf{o}; \boldsymbol{\theta})}{p(\mathbf{z}^\star, \mathbf{y}^\mathbf{o}; \boldsymbol{\theta}^\star)}\right) = \log\left(\frac{p(\mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{z}^\star; \boldsymbol{\theta}^\star)}\right) + LR(\boldsymbol{\theta}, \mathbf{z})$$

For C small enough, we can assume without loss of generality that $\mathbf{z}$ is the representative closest to $\mathbf{z}^\star$ and note $r = \|\mathbf{z} - \mathbf{z}^\star\|_0$. Then:

$$\begin{aligned} LR(\boldsymbol{\theta}, \mathbf{z}) &\leq \Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}) + \tilde{\Lambda}(\mathbf{z}) \\ &\leq \Lambda(\mathbf{z}) - \tilde{\Lambda}(\mathbf{z}) - c\rho\frac{3\delta(\boldsymbol{\pi}^\star)}{4}nr \\ &\leq c\rho\frac{3\delta(\boldsymbol{\pi}^\star)}{4}nr(1 + o_P(1)) \end{aligned}$$

where the first line comes from the definition of Λ, the second line from Proposition 22 and the third from Proposition 25. Thanks to proposition 29, we also know that:

$$\log\left(\frac{p(\mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{z}^\star; \boldsymbol{\theta}^\star)}\right) \leq \mathcal{O}_P(1)\exp\left\{M_{c/4}r\right\}$$

There are at most $\binom{n}{r}Q^r$ assignments $\mathbf{z}$ at distance r of $\mathbf{z}^\star$ and each of them has at most Q^Q equivalent configurations. Therefore,

$$\begin{aligned} &\frac{\sum_{\substack{\mathbf{z}\in S(\mathbf{z}^\star, \tilde{c}) \\ \mathbf{z}\not\sim\mathbf{z}^\star}} p(\mathbf{z}, \mathbf{y}^\mathbf{o}; \boldsymbol{\theta})}{p(\mathbf{z}^\star, \mathbf{y}^\mathbf{o}; \boldsymbol{\theta}^\star)} \\ &\leq \mathcal{O}_P(1)\sum_{r\geq 1}\binom{n}{r}\mathcal{Q}^{\mathcal{Q}+r}\exp\left(rM_{c/4} - c\rho\frac{3\delta(\boldsymbol{\pi}^\star)}{4}nr(1 + o_P(1))\right) \\ &= \mathcal{O}_P(1)\left(1 + e^{(\mathcal{Q}+1)\log\mathcal{Q} + M_{c/4} - c\rho n\frac{3\delta(\boldsymbol{\pi}^\star)(1+o_P(1))}{4}}\right)^n - 1 \\ &\leq \mathcal{O}_P(1)a_n\exp(a_n) \end{aligned}$$

where $a_n = ne^{(\mathcal{Q}+1)\log\mathcal{Q} + M_{c/4} - c\rho n\frac{3\delta(\boldsymbol{\pi}^\star)(1+o_P(1))}{4}} = o_P(1)$.

□

11.9 Proof of Proposition 27 (contribution of equivalent assignments)

Proof. Choose s permutations of $\{1, \ldots, \mathcal{Q}\}$ and assume that $\mathbf{z} = \mathbf{z}^{\star,s}$. Then $p(\mathbf{y}^\mathbf{o}, \mathbf{z}; \boldsymbol{\theta}) = p(\mathbf{y}^\mathbf{o}, \mathbf{z}^{\star,s}; \boldsymbol{\theta}) = p(\mathbf{y}^\mathbf{o}, \mathbf{z}^\star; \boldsymbol{\theta}^s)$. If furthermore $s \in \mathrm{Sym}(\boldsymbol{\theta})$, $\boldsymbol{\theta}^s = \boldsymbol{\theta}$ and immediately $p(\mathbf{y}^\mathbf{o}, \mathbf{z}; \boldsymbol{\theta}) = p(\mathbf{y}^\mathbf{o}, \mathbf{z}^\star; \boldsymbol{\theta})$. We can therefore partition the sum as

$$\begin{aligned}\sum_{\mathbf{z}\sim\mathbf{z}^\star} p(\mathbf{y}^\mathbf{o},\mathbf{z};\boldsymbol{\theta}) &= \sum_s p(\mathbf{y}^\mathbf{o},\mathbf{z}^{\star,s};\boldsymbol{\theta}) \\ &= \sum_s p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^s) \\ &= \sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \#\,\mathrm{Sym}(\boldsymbol{\theta}')p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}') \\ &= \#\,\mathrm{Sym}(\boldsymbol{\theta}) \sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}')\end{aligned}$$

$p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta})$ uni-modal in $\boldsymbol{\theta}$, with a mode in $\widehat{\boldsymbol{\theta}}_{MC}$. By consistency of $\widehat{\boldsymbol{\theta}}_{MC}$, either $p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}) = o_P(p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star))$ or $p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}) = \mathcal{O}_P(p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star))$ and $\boldsymbol{\theta}\to\boldsymbol{\theta}^\star$. In the latter case, any $\boldsymbol{\theta}'\sim\boldsymbol{\theta}$ other than $\boldsymbol{\theta}$ is bounded away from $\boldsymbol{\theta}^\star$ and thus $p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}') = o_P(p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star))$. In summary,

$$\sum_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}')}{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star)} = \max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}} \frac{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}')}{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star)}(1+o_P(1))$$

□

11.10 Proof of Corollary 18: Behavior of $\widehat{\theta}_{MLE}$

We may prove the corollary by contradiction. Note first that unless $\boldsymbol{\Theta}$ is constrained and with high probability, $\widehat{\boldsymbol{\theta}}_{MLE}$ and $\widehat{\boldsymbol{\theta}}(\mathbf{z}^\star)$ exhibit no symmetries. Indeed, equalities like $\widehat{y}_{q\ell} = \widehat{y}_{q',\ell'}$ have vanishingly small probabilities of being simultaneously true when y_{ij} is discrete, and even null when y_{ij} is continuous. Assume then $\min_s(\widehat{\boldsymbol{\alpha}}^s_{MLE} - \hat{\boldsymbol{\alpha}}(\mathbf{z}^\star)) \neq o_P(1/\sqrt{n})$ or $\min_s(\widehat{\boldsymbol{\pi}}^s_{MLE} - \hat{\boldsymbol{\pi}}(\mathbf{z}^\star)) \neq o_P(1/n)$ where s is a permutation of $\{1,\ldots,\mathcal{Q}\}$. Then, by Proposition 17 and the consistency of $\hat{\boldsymbol{\theta}}(\mathbf{z}^\star)$

$$\min_s \mathcal{L}^\star_{co}\left(\hat{\boldsymbol{\theta}}(\mathbf{z}^\star)\right) - \mathcal{L}^\star_{co}\left(\widehat{\boldsymbol{\theta}}^s_{MLE}\right) = \Omega_P(1). \tag{11.2}$$

But, since $\hat{\boldsymbol{\theta}}(\mathbf{z}^\star)$ and $\widehat{\boldsymbol{\theta}}_{MLE}$ maximize respectively $\frac{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}')}{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star)}$ and $\frac{p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta})}{p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta}^\star)}$ and have no symmetries, it follows by Theorem 15 that

$$\left|\frac{p\left(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\hat{\boldsymbol{\theta}}(\mathbf{z}^\star)\right)}{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star)} - \max_s \frac{p\left(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\widehat{\boldsymbol{\theta}}^s_{MLE}\right)}{p(\mathbf{y}^\mathbf{o},\mathbf{z}^\star;\boldsymbol{\theta}^\star)}\right| = o_P(1)$$

which contradicts Eq (11.2) and concludes the proof.

11.11 Proof of Corollary 19: Behavior of $J(\mathbb{Q},\boldsymbol{\theta})$

Remark first that for every $\boldsymbol{\theta}$ and for every $\mathbf{z}$,

$$p(\mathbf{y}^\mathbf{o},\mathbf{z};\boldsymbol{\theta}) \leq \exp\left[J(\delta_\mathbf{z},\boldsymbol{\theta})\right] \leq \max_{\mathbb{Q}\in\mathcal{Q}} \exp\left[J(\mathbb{Q},\boldsymbol{\theta})\right] \leq p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta})$$

where $\delta_\mathbf{z}$ denotes the dirac mass on $\mathbf{z}$. By dividing by $p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta}^\star)$, we obtain

$$\frac{p(\mathbf{y}^\mathbf{o},\mathbf{z};\boldsymbol{\theta})}{p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta}^\star)} \leq \frac{\max_{\mathbb{Q}\in\mathcal{Q}} \exp\left[J(\mathbb{Q},\boldsymbol{\theta})\right]}{p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta}^\star)} \leq \frac{p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta})}{p(\mathbf{y}^\mathbf{o};\boldsymbol{\theta}^\star)}.$$

As this inequality is true for every couple $\mathbf{z}$, we have in particular:

$$\max_{\mathbf{z}\sim\mathbf{z}^\star}\frac{p\left(\mathbf{y^o},\mathbf{z};\boldsymbol{\theta}\right)}{p\left(\mathbf{y^o};\boldsymbol{\theta}^\star\right)}=\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}'\right)}{p\left(\mathbf{y^o};\boldsymbol{\theta}^\star\right)}\leq\frac{\max\limits_{\mathbb{Q}\in\mathcal{Q}}\ \exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]}{p\left(\mathbf{y^o};\boldsymbol{\theta}^\star\right)}.$$

Noticing that $p\left(\mathbf{y^o};\boldsymbol{\theta}^\star\right)=\#\operatorname{Sym}(\boldsymbol{\theta}^\star)p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}^\star\right)(1+o_p(1))$, Theorem 15 therefore leads to the following bounds:

$$\begin{aligned}\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}'\right)}{p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}^\star\right)}(1+o_P(1))&\leq\frac{\max\limits_{\mathbb{Q}\in\mathcal{Q}}\ \exp\left[J\left(\mathbb{Q},\boldsymbol{\theta}\right)\right]}{p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}^\star\right)}\\&\leq\#\operatorname{Sym}(\boldsymbol{\theta})\max_{\boldsymbol{\theta}'\sim\boldsymbol{\theta}}\frac{p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}'\right)}{p\left(\mathbf{y^o},\mathbf{z}^\star;\boldsymbol{\theta}^\star\right)}(1+o_P(1))+o_P(1).\end{aligned}$$

Again unless $\boldsymbol{\Theta}$ is constrained, $\widehat{\boldsymbol{\theta}}_{VAR}$ exhibits no symmetries with high probability and the same proof by contradiction as in appendix 11.10 gives the result.

12 Sub-exponential random variables

We now prove two propositions regarding sub-exponential variables. Recall first that a random variable X is sub-exponential with parameters (τ^2,b) if for all λ such that $|\lambda|\leq 1/b$,

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}(X))}]\leq\exp\left(\frac{\lambda^2\tau^2}{2}\right).$$

In particular, all distributions coming from a natural exponential family are sub-exponential. Sub-exponential variables satisfy a large deviation Bernstein-type inequality:

$$\mathbb{P}(X-\mathbb{E}[X]\geq t)\leq\begin{cases}\exp\left(-\frac{t^2}{2\tau^2}\right) & \text{if}\quad 0\leq t\leq\frac{\tau^2}{b}\\ \exp\left(-\frac{t}{2b}\right) & \text{if}\quad t\geq\frac{\tau^2}{b}\end{cases}\tag{12.1}$$

So that

$$\mathbb{P}(X-\mathbb{E}[X]\geq t)\leq\exp\left(-\frac{t^2}{2(\tau^2+bt)}\right)$$

The sub-exponential property is preserved by summation and multiplication.

- If X is sub-exponential with parameters (τ^2,b) and $\alpha\in\mathbb{R}$, then so is αX with parameters $(\alpha^2\tau^2,\alpha b)$
- If the X_i, $i=1,\ldots,n$ are sub-exponential with parameters (τ_i^2,b_i) and independent, then so is $X=X_1+\cdots+X_n$ with parameters $(\sum_i\tau_i^2,\max_i b_i)$

Theorem 20 (Equivalent characterizations of sub-exponential variables)**.** *For a zero-mean random variable X, the following statements are equivalent:*

1. *There are non-negative numbers (ν,b^{-1}) such that*

$$\mathbb{E}[e^{\lambda X}]\leq\exp\left(\frac{\lambda^2\nu^2}{2}\right)\quad\textit{for all } |\lambda|<b.$$

2. *There is a positive number $c_0>0$ such that $\mathbb{E}[e^{\lambda X}]<\infty$ for all $|\lambda|<c_0$.*
3. *There are constants $c_1,c_2>0$ such that*

$$\mathbb{P}(|X|\geq t)\leq c_1e^{-c_2t}\quad\textit{for all } t>0.$$

4. *The quantity* $\gamma := \sup_{k\geq 2}\left[\frac{\mathbb{E}[X^k]}{k!}\right]^{1/k}$ *is finite.*

Proof. A proof of this theorem can be found in Wainwright (2015). □

Proposition 28 (Maximum in $\mathbf{z}$)**.** *Let* ($\bar{\mathbf{z}}$ *be any configuration and* $\mathbf{z}$ *the* $\sim$*-equivalent configuration that achieves* $\|\mathbf{z}-\mathbf{z}^\star\|_0 = \|\bar{\mathbf{z}}-\mathbf{z}^\star\|_{0,\sim}$ *let* $\widehat{y}_{q\ell} = \hat{y}_{q,\ell}(\mathbf{z})$ *(resp.* $\bar{y}_{q\ell}(\mathbf{z})$*) and* $\widehat{y}^\star_{q\ell} = \hat{y}_{q,\ell}(\mathbf{z}^\star)$ *(resp.* $\bar{y}^\star_{q\ell} = \bar{y}_{q\ell}(\mathbf{z}^\star) = \psi'(\pi^\star_{q\ell})$*) be as defined in Equations (3.1) and (5.3). Under the assumptions of the section 2.5, for all* $\varepsilon \leq \kappa\bar{\sigma}^2$*,*

$$\mathbb{P}\left(\max_{\bar{\mathbf{z}}\sim\mathbf{z}^\star}\max_{k,l}\frac{N^o_{q\ell}(\mathbf{z})(\hat{y}_{q,\ell}-\bar{y}_{q\ell}) - N^o_{q\ell}(\mathbf{z}^\star)(\widehat{y}^\star_{q\ell}-\bar{y}^\star_{q\ell})}{n\|\mathbf{z}-\mathbf{z}^\star\|_0} > \varepsilon\right) \qquad = \qquad o(1)$$

Proof. Note $r = \|\mathbf{z}-\mathbf{z}^\star\|_0$. The numerator within the max in the fraction can be expanded to

$$Z_{q\ell}(\mathbf{z}) = \sum_{i,j}(z_{iq}z_{j\ell} - z^\star_{iq}z^\star_{j\ell})(y_{ij}r_{ij} - \pi^\star_{z^\star_{iq}z^\star_{j\ell}}\rho)$$

and is thus a sum of at most $N = nr$ non-null centered sub-exponential random variables with parameters $(a^2, 1/w)$. It is therefore a centered sub-exponential with parameters $(Na^2, 1/w)$. By Bernstein inequality, for all $\varepsilon \leq \kappa a^2$ we have

$$\mathbb{P}(Z \geq \varepsilon nr) \leq \exp\left(-\frac{nr\varepsilon^2}{2a^2}\right).$$

There are at most $n^r\mathcal{Q}^r\mathcal{Q}^{\mathcal{Q}}$ $\mathbf{z}$ at $\|.\|_{0,\sim}$ distance r of $\mathbf{z}^\star$. An union bound shows that:

$$\begin{aligned}\mathbb{P}\left(\max_{\bar{\mathbf{z}}\sim\mathbf{z}^\star}\max_{q,\ell}\frac{Z_{q\ell}(\mathbf{z})}{n\|\mathbf{z}-\mathbf{z}^\star\|_0} \geq \varepsilon\right)& \\ &\leq \sum_{r\geq 1}\sum_{r=\|\bar{\mathbf{z}}-\mathbf{z}^\star\|_{0,\sim}}\mathcal{Q}^2\mathbb{P}(Z_{q\ell}(\mathbf{z}) \geq \varepsilon nr) \\ &\leq \sum_{r\geq 1}\mathcal{Q}^{\mathcal{Q}}\exp\left(-nr\varepsilon^2/2a^2 + r\log(n\mathcal{Q}) + 2\log(\mathcal{Q})\right) = o(1)\end{aligned}$$

where the last equality is true as soon as $n\varepsilon_n \gg \log n$.

□

13 Likelihood ratio of assignments

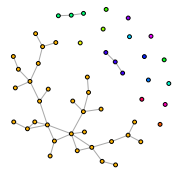
Proposition 29. *Let* $\mathbf{z}^\star$ *be* $c/2$*-regular and* $\mathbf{z}$ *at* $\|.\|_0$*-distance* $c/4$ *of* $\mathbf{z}^\star$*. Then, for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\log\frac{p(\mathbf{z};\boldsymbol{\theta})}{p(\mathbf{z}^\star;\boldsymbol{\theta}^\star)} \leq \mathcal{O}_P(1)\exp\left\{M_{c/4}\|\mathbf{z}-\mathbf{z}^\star\|_0\right\}$$

Proof. Note then that:

$$\begin{aligned}\frac{p(\mathbf{z};\boldsymbol{\theta})}{p(\mathbf{z}^\star;\boldsymbol{\theta}^\star)} = \frac{p(\mathbf{z};\boldsymbol{\alpha})}{p(\mathbf{z}^\star;\boldsymbol{\alpha}^\star)} &= \frac{p(\mathbf{z};\boldsymbol{\alpha})}{p(\mathbf{z}^\star;\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))}\frac{p(\mathbf{z}^\star;\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))}{p(\mathbf{z}^\star;\boldsymbol{\alpha}^\star)} \\ \leq &\quad \frac{p(\mathbf{z};\hat{\boldsymbol{\alpha}}(\mathbf{z}))}{p(\mathbf{z}^\star;\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))}\frac{p(\mathbf{z}^\star;\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))}{p(\mathbf{z}^\star;\boldsymbol{\alpha}^\star)} \\ \leq &\quad \exp\left\{M_{c/4}\|\mathbf{z}-\mathbf{z}^\star\|_0\right\} \times \frac{p(\mathbf{z}^\star;\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))}{p(\mathbf{z}^\star;\boldsymbol{\alpha}^\star)} \\ \leq &\quad \mathcal{O}_P(1)\exp\left\{M_{c/4}\|\mathbf{z}-\mathbf{z}^\star\|_0\right\}\end{aligned}$$

where the first inequality comes from the definition of $\hat{\boldsymbol{\alpha}}(\mathbf{z})$ and the second from Lemma B.6 of Brault et al. (2017) and the fact that $\mathbf{z}^\star$ and $\mathbf{z}$ are $c/4$-regular. Finally, local asymptotic normality of the MLE for multinomial proportions ensures that $\frac{p(\mathbf{z}^\star;\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))}{p(\mathbf{z}^\star;\boldsymbol{\alpha}^\star)} = \mathcal{O}_P(1)$. □



14 General technical results

Proposition 30 (Maximum in $\mathbf{z}$). *Let $\mathbf{z}$ be a configuration and $\hat{y}_{q,\ell}(\mathbf{z})$ resp. $\bar{y}_{q\ell}(\mathbf{z})$ be as defined in Equations (3.1) and (5.3). Under the assumptions of the section 2.5, for all $\varepsilon > 0$*

$$\mathbb{P}\left(\max_{\mathbf{z}} \max_{k,l} \widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})|\hat{y}_{q,\ell} - \bar{y}_{q\ell}| > \varepsilon\right) \leq \mathcal{Q}^{n+2} \exp\left(-\frac{n^2\varepsilon^2}{2(\bar{\sigma}^2 + \kappa^{-1}\varepsilon)}\right). \tag{14.1}$$

Additionally, the suprema over all $c/2$-regular assignments satisfies:

$$\mathbb{P}\left(\max_{\mathbf{z}\in\mathcal{Z}_1} \max_{k,l} |\hat{y}_{q,\ell} - \bar{y}_{q\ell}| > \varepsilon\right) \leq \mathcal{Q}^{n+2} \exp\left(-\frac{n^2c^2\varepsilon^2}{8(\bar{\sigma}^2 + \kappa^{-1}\varepsilon)}\right). \tag{14.2}$$

Note that equations 14.1 and 14.2 remain valid when replacing $c/2$ by any $\tilde{c} < c/2$.

Proof. The random variables $Y_{ij}R_{ij}$ are sub-exponential with parameters $(\nu^2, 1/b)$. Conditionally to $\mathbf{z}^\star$, $z_{+q}z_{+\ell}(\hat{y}_{q,\ell} - \bar{y}_{q\ell})$ is a sum of $z_{+q}z_{+\ell}$ centered sub-exponential random variables. By Bernstein's inequality Boucheron et al. (2013), we therefore have for all $t > 0$

$$\mathbb{P}(z_{+q}z_{+\ell}|\hat{y}_{q,\ell} - \bar{y}_{q\ell}| \geq t) \leq 2\exp\left(-\frac{t^2}{2(z_{+q}z_{+\ell}\nu^2 + b^{-1}t)}\right)$$

In particular, if $t = n^2y$,

$$\mathbb{P}\left(\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})|\hat{y}_{q,\ell} - \bar{y}_{q\ell}| \geq x\right) \leq 2\exp\left(-\frac{n^2y^2}{2(\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})\nu^2 + b^{-1}y)}\right) \leq 2\exp\left(-\frac{n^2y^2}{2(\nu^2 + b^{-1}y)}\right)$$

uniformly over $\mathbf{z}$. Equation (14.1) then results from a union bound. Similarly,

$$\begin{aligned}
\mathbb{P}\left(|\hat{y}_{q,\ell} - \bar{y}_{q\ell}| \geq x\right) &= \mathbb{P}\left(\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})|\hat{y}_{q,\ell} - \bar{y}_{q\ell}| \geq \widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})y\right) \\
&\leq 2\exp\left(-\frac{n^2y^2\widehat{\alpha}_q(\mathbf{z})^2\widehat{\alpha}_\ell(\mathbf{z})^2}{2(\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z})\nu^2 + b^{-1}y\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z}))}\right) \\
&\leq 2\exp\left(-\frac{n^2c^2y^2}{8(\nu^2 + b^{-1}y)}\right)
\end{aligned}$$

Where the last inequality comes from the fact that $c/2$-regular assignments satisfy $\widehat{\alpha}_q(\mathbf{z})\widehat{\alpha}_\ell(\mathbf{z}) \geq c^2/4$. Equation (14.2) then results from a union bound over $\mathcal{Z}_1 \subset \mathcal{Z}$.

□

Lemme 3. *If X is a zero mean random variable, sub-exponential with parameters (σ^2, b), then $|X|$ is sub-exponential with parameters $(8\sigma^2, 2\sqrt{2}b)$.*

Proof. Note $\mu = \mathbb{E}|X|$ and consider $Y = |X| - \mu$. Choose λ such that $|\lambda| < (2\sqrt{2}b)^{-1}$. We need to bound $\mathbb{E}[e^{\lambda Y}]$. Note first that $\mathbb{E}[e^{\lambda Y}] \leq \mathbb{E}[e^{\lambda X}] + \mathbb{E}[e^{-\lambda X}] < +\infty$ is properly defined by sub-exponential property of X and we have

$$\mathbb{E}[e^{\lambda Y}] \leq 1 + \sum_{k=2} \frac{|\lambda|^k\mathbb{E}[|Y|^k]}{k!}$$

where we used the fact that $\mathbb{E}[Y] = 0$. We know bound odd moments of $|\lambda Y|$.

$$\mathbb{E}[|\lambda Y|^{2k+1}] \leq (\mathbb{E}[|\lambda Y|^{2k}]\mathbb{E}[|\lambda Y|^{2k+2}])^{1/2} \leq \frac{1}{2}(\lambda^{2k}\mathbb{E}[Y^{2k}] + \lambda^{2k+2}\mathbb{E}[Y^{2k+2}])$$

where we used first Cauchy-Schwarz and then the arithmetic-geometric mean inequality. The Taylor series expansion can thus be reduced to

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y}] &\leq 1 + \left(\frac{1}{2} + \frac{1}{2.3!}\right)\mathbb{E}[Y^2]\lambda^2 + \sum_{k=2}^{+\infty}\left(\frac{1}{(2k)!} + \frac{1}{2}\left[\frac{1}{(2k-1)!} + \frac{1}{(2k+1)!}\right]\right)\lambda^{2k}\mathbb{E}[Y^{2k}] \\
&\leq \sum_{k=0}^{+\infty} 2^k \frac{\lambda^{2k}\mathbb{E}[Y^{2k}]}{(2k)!} \\
&\leq \sum_{k=0}^{+\infty} 2^{3k} \frac{\lambda^{2k}\mathbb{E}[X^{2k}]}{(2k)!} = \mathbb{E}\left[\cosh\left(2\sqrt{2}\lambda X\right)\right] = \mathbb{E}\left[\frac{e^{2\sqrt{2}\lambda X} + e^{-2\sqrt{2}\lambda X}}{2}\right] \\
&\leq e^{\frac{8\lambda^2\sigma^2}{2}}
\end{aligned}$$

where we used the well-known inequality $\mathbb{E}[|X - \mathbb{E}[X]|^k] \leq 2^k\mathbb{E}[|X|^k]$ to substitute $2^{2k}\mathbb{E}[X^{2k}]$ to $\mathbb{E}[Y^{2k}]$.

□

Proposition 31 (concentration for sub-exponential)**.** *Let $X_1, \ldots, X_n$ be independent zero mean random variables, sub-exponential with parameters (σ_i^2, b_i). Note $V_0^2 = \sum_i \sigma_i^2$ and $b = \max_i b_i$. Then the random variable Z defined by:*

$$Z = \sup_{\substack{\Gamma \in \mathbb{R}^n \\ \|\Gamma\|_\infty \leq M}} \sum_i \Gamma_i X_i$$

is also sub-exponential with parameters $(8M^2V_0^2, 2\sqrt{2}Mb)$. Moreover $\mathbb{E}[Z] \leq MV_0\sqrt{n}$ so that for all $t > 0$,

$$\mathbb{P}(Z - MV_0\sqrt{n} \geq t) \leq \exp\left(-\frac{t^2}{2(8M^2V_0^2 + 2\sqrt{2}Mbt)}\right) \tag{14.3}$$

Proof. Note first that Z can be simplified to $Z = M\sum_i |X_i|$. We just need to bound $\mathbb{E}[Z]$. The rest of the proposition results from the fact that the $|X_i|$ are sub-exponential $(8\sigma_i^2, 2\sqrt{2}b_i)$ by Lemma 3 and standard properties of sums of independent re-scaled sub-exponential variables.

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\sup_{\substack{\Gamma \in \mathbb{R}^n \\ \|\Gamma\|_\infty \leq M}} \sum_i \Gamma_i X_i\right] = \mathbb{E}\left[\sum_i M|X_i|\right] \leq M\sum_i \sqrt{\mathbb{E}[X_i^2]} \\
&= M\sum_i \sigma_i \leq M\left(\sum_i 1\right)^{1/2}\left(\sum_i \sigma_i^2\right)^{1/2} = MV_0\sqrt{n}
\end{aligned}$$

using Cauchy-Schwarz.

□

Lemme 4.

Let $\mathcal{Z}_1$ be the subset of $\mathcal{Z}$ of c-regular configurations, as defined in Definition 11. Let $\mathbb{S}^Q = \{\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_Q) \in [0,1]^Q : \sum_{k=1}^Q \alpha_k = 1\}$ be the Q-dimensional simplex and note $\mathbb{S}_c^Q = \mathbb{S}^Q \cap [c, 1-c]^Q$. Then there exists two positive constants M_c and M_c' such that for all $\mathbf{z}$, $\mathbf{z}^\star$ in $\mathcal{Z}_1$ and all $\boldsymbol{\alpha} \in \mathbb{S}_c^Q$

$$|\log p(\mathbf{z}; \hat{\boldsymbol{\alpha}}(\mathbf{z})) - \log p(\mathbf{z}^\star; \hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))| \quad \leq \quad M_c\|\mathbf{z} - \mathbf{z}^\star\|_0$$

Proof. Consider the entropy map $H : \mathbb{S}^Q \to \mathbb{R}$ defined as $H(\boldsymbol{\alpha}) = -\sum_{k=1}^{Q} \alpha_k \log(\alpha_k)$. The gradient ∇H is uniformly bounded by $\frac{M_c}{2} = \log \frac{1-c}{c}$ in $\|.\|_\infty$-norm over $\mathbb{S}^Q \cap [c, 1-c]^Q$. Therefore, for all $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^\star \in \mathbb{S}^Q \cap [c, 1-c]^Q$, we have

$$|H(\boldsymbol{\alpha}) - H(\boldsymbol{\alpha}^\star)| \leq \frac{M_c}{2} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^\star\|_1$$

To prove the inequality, we remark that $\mathbf{z} \in \mathcal{Z}_1$ translates to $\hat{\boldsymbol{\alpha}}(\mathbf{z}) \in \mathbb{S}^Q \cap [c, 1-c]^Q$, that $\log p(\mathbf{z}; \hat{\boldsymbol{\alpha}}(\mathbf{z})) - \log p(\mathbf{z}^\star; \hat{\boldsymbol{\alpha}}(\mathbf{z}^\star)) = n[H(\hat{\boldsymbol{\alpha}}(\mathbf{z})) - H(\hat{\boldsymbol{\alpha}}(\mathbf{z}^\star))]$ and finally that $\|\hat{\boldsymbol{\alpha}}(\mathbf{z}) - \hat{\boldsymbol{\alpha}}(\mathbf{z}^\star)\|_1 \leq \frac{2}{n} \|\mathbf{z} - \mathbf{z}^\star\|_0$. □

Chapter 4

SBM with covariates and missing values

Contents

1 Introduction

This chapter deals with missing data in the Stochastic Block Model (in short SBM) with covariates. The case with no covariates and missing data in the SBM has already been treated in Chapter 2. We follow this work and introduce three different models of SBM with covariates on nodes or dyads. In order to explore the link between the sampling design and the covariates we define one sampling design centered on dyads and another one centered on nodes, both depending on covariates. We exhibit a specific case where a sampling design which is MAR conditionally on covariates becomes NMAR when the covariates are not observed. We show explicitly the equivalence between the two models and run simulations to compare the inference with a MAR model conditionally on covariates, NMAR model and a MAR model. We also propose another case where there is no explicit equivalence but where the NMAR model leads to better inference than a MAR model when the covariates are missing.

Related works. Stochastic Block Models with the adding of covariates are popular to model heterogeneity in networks. The main purpose of accounting for covariates in the model is to remove covariates effects on clustering. In Tallberg (2004), the authors proposed a Bayesian approach to model an SBM with covariates with a multinomial probit model where distribution of the block given by the latent variable $\mathbf{Z}_i$ depends on the vector of covariates $\mathbf{X}_i$. As a consequence, in this model covariates act on dyads through nodes memberships. When covariates are associated to dyads, Mariadassou et al. (2010) combined these covariates with the hidden structure using the framework of the generalized linear model, for example in the binary SBM with covariates

$$Y_{ij}|z_i = q, z_j = \ell, \mathbf{X}_{ij} \sim^{ind} \mathcal{B}(\text{logistic}(\gamma_{q\ell} + \beta^T \mathbf{X}_{ij})). \tag{1.1}$$

A very close model is proposed in Zanghi et al. (2010) with Gaussian distribution on dyads. In these frameworks, heterogeneity is modeled by parameters $(\gamma_{q\ell})_{q\ell}$ corresponding when $\mathbf{X} = 0$ to inter and intra block probabilities (or connectivity matrix) in the binary SBM. It accounts for heterogeneity that is not explained by the regression term $\beta^T \mathbf{X}_{ij}$. A similar interpretation is given in Choi et al. (2012b) who first apply a regression step and then perform SBM inference on the residuals. SBM with covariates has also been applied in Miele et al. (2014) in a spatial data context where nodes correspond to entities that have explicit geographic locations. If no hidden structure is supposed, Robins et al. (2007) show how to incorporate covariates into a graph model. Alternative models are used to model covariates in SBM, like in Sweet (2015) where authors use a hierarchical Bayesian model. When the principal purpose is community detection, the previous models enable to detect structure in the network beyond the effect of covariates while the following references focus on detecting structure by using both covariates and the network. We cite Zhang et al. (2016) which is based on a joint community detection criterion and Binkiewicz et al. (2017) using spectral properties of the adjacency matrix of a network.

2 Stochastic Block Models with covariates and sampling models

We will consider three different structures in which covariates and missing data impact the SBM. Conditional dependencies of these models are represented with a DAG in Figure 4.1. We choose to study only cases where covariates represented by $\mathbf{X}$ impact either the sampling design or the network model directly. Furthermore, note that the systematic edge between $\mathbf{Z}$ and $\mathbf{Y}$ is part of the SBM. Finally, we do not consider any edge between nodes $\mathbf{Y}$, $\mathbf{Z}$ and node $\mathbf{R}$ since we require that conditionally on $\mathbf{X}$ the missing data are MAR. In the first model, covariates influence directly latent variables and the sampling. In the second model covariates influence also the sampling and the distribution of edges. Finally, in the last model, covariates influence latent variables, the sampling and the distribution of edges.

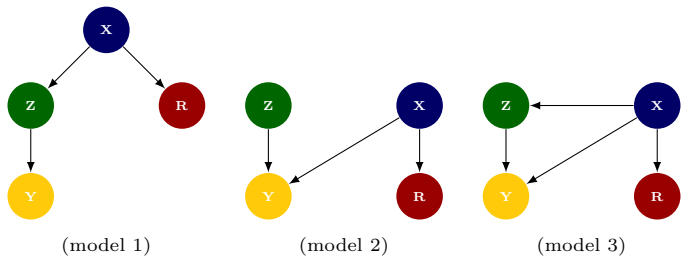


Figure 4.1 – DAGs of relationships between $\mathbf{Y}, \mathbf{Z}, \mathbf{R}$ and $\mathbf{X}$ considered in the framework of missing data for SBM with covariates.

Model 1 Writing $\boldsymbol{\alpha}_{\cdot} = (\boldsymbol{\alpha}_{\cdot 1}, ..., \boldsymbol{\alpha}_{\cdot Q}) \in [0,1]^Q$ we define

$$\alpha_{iq} = \frac{e^{\boldsymbol{\beta}_q^t \mathbf{X}_i \mathbb{1}_{\{q \neq Q\}}}}{1 + \sum_{k=1}^{Q-1} \boldsymbol{\beta}_k^t \mathbf{X}_i}, \quad \forall (i,q) \in \mathcal{N} \times \mathcal{Q},$$

with $\boldsymbol{\beta}_q \in \mathbb{R}^N$ for all $q \in [\![1, Q-1]\!]$ and $\boldsymbol{\beta}_Q = 0$. Then we have

$$\begin{aligned} \mathbf{Z}_i \mid \mathbf{X}_i &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}_i), \quad \forall i \in \mathcal{N}, \\ Y_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j &\sim^{\text{ind}} \mathcal{B}(\pi_{z_i z_j}), \quad \forall (i,j) \in \mathcal{N}^2, \end{aligned}$$

with $\pi_{q\ell} \in [0,1]$, $(q,\ell) \in \mathcal{Q}^2$.

Model 2

$$\begin{aligned} \mathbf{Z}_i &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}), \quad \forall i \in \mathcal{N}, \\ Y_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, \mathbf{X}_{ij} &\sim^{\text{ind}} \mathcal{B}(g(\gamma_{z_i z_j} + \beta^t \mathbf{X}_{ij})), \quad \forall (i,j) \in \mathcal{N}^2, \end{aligned}$$

where $\boldsymbol{\gamma} \in \mathcal{M}_Q([0,1])$, $\boldsymbol{\beta} \in \mathbb{R}^m$, $\boldsymbol{\alpha} \in [0,1]^Q$, $g(x) = (1+e^{-x})^{-1}$. Note that if the covariates are linked with nodes, the covariates are transferred to dyads by using a "similarity" symmetric function $\phi(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^m$ such that $\mathbf{X}_{ij} = \phi(\mathbf{X}_i, \mathbf{X}_j)$ for all (i,j) in $\mathcal{D}$.

Model 3 Writing $\boldsymbol{\alpha}_{\cdot} = (\boldsymbol{\alpha}_{\cdot 1}, ..., \boldsymbol{\alpha}_{\cdot Q}) \in [0,1]^Q$ we define

$$\alpha_{iq} = \frac{e^{\boldsymbol{\beta}_q^t \mathbf{X}_i \mathbb{1}_{\{q \neq Q\}}}}{1 + \sum_{k=1}^{Q-1} \boldsymbol{\beta}_k^t \mathbf{X}_i}, \quad \forall (i,q) \in \mathcal{N} \times \mathcal{Q},$$

with $\boldsymbol{\beta}_q \in \mathbb{R}^N$ for all $q \in [\![1, Q-1]\!]$ and $\boldsymbol{\beta}_Q = 0$. Then we have

$$\begin{aligned} \mathbf{Z}_i \mid \mathbf{X}_i &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}_i), \quad \forall i \in \mathcal{N}, \\ Y_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, \mathbf{X}_{ij} &\sim^{\text{ind}} \mathcal{B}(g(\gamma_{z_i z_j} + \beta^t \mathbf{X}_{ij})), \quad \forall (i,j) \in \mathcal{N}^2, \end{aligned}$$

with parameters and notations defined as in Model 2.

In the three previous models, we need to specify the distribution $\mathbf{R}|\mathbf{X}$. This distribution can be naturally one of the two following distributions depending whether $\mathbf{X}$ is a set of covariates on nodes or on dyads. We then define two sampling designs, one dyad-centered and the other node-centered. In both sampling, the probability to observe a dyad (resp. node) depends on the value of the covariate. If the covariates have no effect, then the probability to observe a dyad (resp. node) is independent of the dyad (resp. node).

Definition 6 (Dyad sampling)**.** *Let $\delta \in \mathbb{R}$, $\kappa \in \mathbb{R}^m$. The probability to observe a dyad is*

$$\mathbb{P}(R_{ij} = 1 | \boldsymbol{X}_{ij}) = g(\delta + \kappa^t \boldsymbol{X}_{ij}),$$

with $\boldsymbol{\psi} = \{\delta, \kappa\}$.

For all $i \in \mathcal{N}$, we denote by $V_i = 1$ if the node i is observed, $V_i = 0$ otherwise. Then, if $V_i = 1$ we have $R_{ij} = 1$ for all $j \in \mathcal{N}$.

Definition 7 (Node sampling)**.** *Let $\nu \in \mathbb{R}$ and $\eta \in \mathbb{R}^N$. The probability to observe all dyads corresponding to a node is*

$$\mathbb{P}(V_i = 1 | \boldsymbol{X}_i) = g(\nu + \eta^t \boldsymbol{X}_i),$$

with $\boldsymbol{\psi} = \{\nu, \eta\}$.

In the next paragraph, we exhibit a particular case where we can provide an equivalence between a MAR model conditionally on covariates and a NMAR model when covariates are not observed. Indeed, when sampling a network, some covariates may not be available. In this case, a NMAR model may compensate for the absence of covariates.

A case of equivalence. In this paragraph, we demonstrate on a specific case that an SBM with covariates where the missing data are MAR conditionally on covariates can be equivalent to an SBM without covariates with missing data NMAR. We first recall the definition of the block-node sampling design where nodes are sampled depending on the group they belong to.

Definition 8 (Block-node sampling)**.** *Let* $\rho = (\rho_q)_{1\leq q\leq Q} \in [0,1]^Q$*. In block-node sampling, the conditional distribution of* $\mathbf{V}|\mathbf{Z}$ *is given by*

$$V_i|\mathbf{Z}_i \sim^{ind} \mathcal{B}(\rho_{z_i}). \tag{2.1}$$

Proposition 32. *Let's define the covariates matrix* $\boldsymbol{X} := [\boldsymbol{X}_1, ..., \boldsymbol{X}_n] \in \mathcal{M}_{Q\times n}(\{0,1\})$ *for* $Q > 1$*. The model is generated as follow*

$$\begin{aligned} \boldsymbol{X}_i \quad &\sim^{iid} \quad \mathcal{M}(1,\boldsymbol{\alpha}), \quad \forall i \in \mathcal{N}, \\ \mathbb{P}(Z_{ia} = 1|X_{ia} = 1) = \delta \quad &\textit{and} \quad \mathbb{P}(Z_{ib} = 1|X_{ia} = 1) = \frac{1-\delta}{Q-1}, \quad b \neq a, \ \forall i \in \mathcal{N}. \end{aligned}$$

The probabilities to observe a dyad are given by

$$p_q = \mathbb{P}(V_i = 1|X_{iq} = 1).$$

Then, this model corresponds to a conventional SBM under block-node sampling, with parameters

$$\begin{aligned} \rho_q &= \mathbb{P}(V_i = 1|Z_{iq} = 1), \\ &= \frac{\delta p_q \alpha_q + \frac{1-\delta}{Q-1}\sum_{\ell\neq q} p_\ell \alpha_\ell}{\delta\alpha_q + \frac{1-\delta}{Q-1}(1-\alpha_q)}. \end{aligned}$$

Remark 1. $\delta = 1 \Rightarrow \rho_q = p_q$.

This result illustrate the article Molenberghs et al. (2008) which states that any model generating missing data that are MAR has a dual model generating missing data NMAR and fitting equally the data.

3 Statistical inference

On the basis of Figure 4.1, the type of missingness for SBM is defined as follows:

$$\text{Sampling design for SBM is} \begin{cases} \text{MCAR} & \text{if } \mathbf{R} \perp\!\!\!\perp (\mathbf{Y}^{\text{m}}, \mathbf{Z}, \mathbf{Y}^{\text{o}}) \mid \mathbf{X}, \\ \text{MAR} & \text{if } \mathbf{R} \perp\!\!\!\perp (\mathbf{Y}^{\text{m}}, \mathbf{Z}) \mid (\mathbf{Y}^{\text{o}}, \mathbf{X}), \\ \text{NMAR} & \text{otherwise.} \end{cases} \tag{3.1}$$

Proposition 33. *From (3.1), if the sampling is MAR or MCAR then maximizing* $p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\text{o}}, \mathbf{R})$ *or* $p_{\boldsymbol{\theta}}(\mathbf{Y}^{\text{o}})$ *in* $\boldsymbol{\theta}$ *is equivalent.*

In the light of (3.1), sampling designs defined in definitions 6 and 7 are missing completely at random (MCAR) conditionally on covariates.

3.1 Inference of Model 1

In the MAR case, inference can be conducted directly and without bias on the observed part of the adjacency matrix. We start by recalling the complete likelihood $p_{\boldsymbol{\theta}}(\mathbf{Y}^o, \mathbf{Z}|\mathbf{X})$ which has an explicit form contrary to the likelihood of the observed data $p_{\boldsymbol{\theta}}(\mathbf{Y}^o|\mathbf{X})$.

The complete log-likelihood restricted to the observed variables is

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}^o, \mathbf{Z}|\mathbf{X}) = \sum_{(i,j)\in\mathcal{D}^{\text{o}}} \sum_{q,\ell} Z_{iq} Z_{j\ell} \log b(Y_{ij}, \pi_{q\ell}) + \sum_{i\in\mathcal{N}^{\text{o}}} \sum_{q} Z_{iq} \log(\alpha_{iq}), \tag{3.2}$$

with $b(x,\pi) = \pi^x(1-\pi)^{1-x}$ the Bernoulli probability density function.

By Equation (3.2) and "mean-field" approximation, the close form of the lower bound is

$$J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\mathbf{X}) = \sum_{(i,j)\in\mathcal{D}^o}\sum_{q,\ell}\tau_{iq}\tau_{j\ell}\log b(Y_{ij},\pi_{q\ell}) + \sum_{i\in\mathcal{N}^o}\sum_{q}\tau_{iq}\log(\alpha_{iq}/\tau_{iq}). \tag{3.3}$$

We refer to the notations defined page 5 for $\mathcal{D}^o$ and $\mathcal{N}^o$. The two maximization problems are solved as stated in the following proposition, straightforwardly derived from Daudin et al. (2008).

Proposition 34. *Consider the lower bound $J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ given by (3.3).*

1. *The parameter $\boldsymbol{\pi}$ maximizing $J_{\boldsymbol{\pi}}(\mathbf{Y}^o|\boldsymbol{X})$ when $\boldsymbol{\tau}$ is held fixed is*

$$\hat{\pi}_{q\ell} = \frac{\sum_{(i,j)\in\mathcal{D}^o}\hat{\tau}_{iq}\hat{\tau}_{j\ell}Y_{ij}}{\sum_{(i,j)\in\mathcal{D}^o}\hat{\tau}_{iq}\hat{\tau}_{j\ell}},$$

   $\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ has no explicit form and must be estimated with an optimization algorithm. Corresponding gradient is

$$\frac{\partial J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})}{\partial\alpha_{iq}} = \frac{\tau_{iq}}{\alpha_{iq}}. \tag{3.4}$$

2. *The variational parameters $\boldsymbol{\tau}$ maximizing $J_{\boldsymbol{\tau}}(\mathbf{Y}^o|\boldsymbol{X})$ when $\boldsymbol{\theta}$ is held fixed are obtained thanks to the following fixed point relation:*

$$\hat{\tau}_{iq} \propto \alpha_{iq}\left(\prod_{(i,j)\in\mathcal{D}^o}\prod_{\ell\in\mathcal{Q}} b(Y_{ij};\pi_{q\ell})^{\hat{\tau}_{j\ell}}\right).$$

3.2 Inference of Model 2

The complete log-likelihood for Model 2 is

$$\begin{aligned}\log p_\theta(\mathbf{Y}^o,\mathbf{Z}|\mathbf{X}) &= \sum_{(i,j)\in\mathcal{D}^o} Y_{ij}g\left(\mathbf{Z}_i{}^t\boldsymbol{\gamma}\mathbf{Z}_j+\boldsymbol{\beta}^t\mathbf{X}_{ij}\right)+(1-Y_{ij})\log(1-g\left(\mathbf{Z}_i{}^t\boldsymbol{\gamma}\mathbf{Z}_j+\boldsymbol{\beta}^t\mathbf{X}_{ij}\right))\\ &\quad+\sum_{i\in\mathcal{N}^o}\mathbf{Z}_i^t\log(\boldsymbol{\alpha}),\end{aligned} \tag{3.5}$$

with $h(x) = \log(g(x)) = -\log(1+e^{-x})$. As a consequences the variational lower bound is given by

$$J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\mathbf{X}) = \sum_{(i,j)\in\mathcal{D}^o}\sum_{q,\ell}\tau_{iq}\tau_{j\ell}\Big\{(Y_{ij}-1)(\gamma_{q\ell}+\boldsymbol{\beta}^t\mathbf{X}_{ij})+h(\gamma_{q\ell}+\boldsymbol{\beta}^t\mathbf{X}_{ij})\Big\} \tag{3.6}$$

$$+\sum_{i,q}\tau_{iq}\log\left(\frac{\alpha_q}{\tau_{iq}}\right). \tag{3.7}$$

Proposition 35. *Consider the maximization of the lower bound (3.7).*

1. *The parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ maximizing $J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ when all other parameters are held fixed have no explicit forms and must be estimated with an optimization algorithm. Corresponding gradients are*

$$\frac{\partial J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})}{\partial\gamma_{q\ell}} = \sum_{(i,j)\in\mathcal{D}^o}\tau_{iq}\tau_{j\ell}\left\{Y_{ij}-1+\frac{e^{-(\gamma_{q\ell}+\boldsymbol{\beta}^t\boldsymbol{X}_{ij})}}{1+e^{-(\gamma_{q\ell}+\boldsymbol{\beta}^t\boldsymbol{X}_{ij})}}\right\}, \tag{3.8}$$

$$\frac{\partial J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})}{\partial\beta_k} = \sum_{(i,j)\in\mathcal{D}^o}\sum_{q,\ell}\tau_{iq}\tau_{j\ell}\left(\boldsymbol{X}_{ij}\right)_k\left\{Y_{ij}-1+\frac{e^{-(\gamma_{q\ell}+\boldsymbol{\beta}^t\boldsymbol{X}_{ij})}}{1+e^{-(\gamma_{q\ell}+\boldsymbol{\beta}^t\boldsymbol{X}_{ij})}}\right\}. \tag{3.9}$$

Concerning $\boldsymbol{\alpha}$,

$$\hat{\alpha}_q = \frac{\sum_{i\in\mathcal{N}^o} \hat{\tau}_{iq}}{\text{card}\,(\mathcal{N}^o)}.$$

2. *The optimal* $\boldsymbol{\tau}$ *in* $J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ *when all other parameters are held fixed verify*

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{(i,j)\in\mathcal{D}^o} \prod_{\ell} \exp\left\{(Y_{ij}-1)(\gamma_{q\ell}+\beta^t\boldsymbol{X}_{ij}) + h(\gamma_{q\ell}+\beta^t\boldsymbol{X}_{ij})\right\}^{\hat{\tau}_{j\ell}}. \tag{3.10}$$

3.3 Inference of Model 3

The inference of Model 3 can be done directly by combination of Model 1 and Model 2 estimation procedures. The complete log-likelihood for Model 3 is

$$\begin{aligned}\log p_\theta(\mathbf{Y}^o,\mathbf{Z}|\mathbf{X}) &= \sum_{(i,j)\in\mathcal{D}^o} Y_{ij} g\left(\mathbf{Z}_i{}^t\boldsymbol{\gamma}\mathbf{Z}_j + \boldsymbol{\beta}^t\mathbf{X}_{ij}\right) + (1-Y_{ij})\log(1-g\left(\mathbf{Z}_i{}^t\boldsymbol{\gamma}\mathbf{Z}_j+\boldsymbol{\beta}^t\mathbf{X}_{ij}\right)) \\ &\quad + \sum_{i\in\mathcal{N}^o} \mathbf{Z}_i^t \log\left(\boldsymbol{\alpha}_i\right), \end{aligned} \tag{3.11}$$

with $h(x) = \log(g(x)) = -\log(1+e^{-x})$. As a consequences the variational lower bound is given by

$$J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\mathbf{X}) = \sum_{(i,j)\in\mathcal{D}^o}\sum_{q,\ell} \tau_{iq}\tau_{j\ell}\left\{(Y_{ij}-1)(\gamma_{q\ell}+\boldsymbol{\beta}^t\mathbf{X}_{ij}) + h(\gamma_{q\ell}+\boldsymbol{\beta}^t\mathbf{X}_{ij})\right\} \tag{3.12}$$

$$+\sum_{i,q}\tau_{iq}\log\left(\frac{\alpha_{iq}}{\tau_{iq}}\right). \tag{3.13}$$

Proposition 36. *Consider the lower bound* $J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ *given by* (3.13).

1. *The parameters* $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ *and* $\boldsymbol{\beta}$ *maximizing* $J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ *when all other parameters are held fixed have no explicit forms and must be estimated with an optimization algorithm. Corresponding gradients are the same as in Equations* (3.4), (3.8) *and* (3.9).

2. *The variational parameters* $\boldsymbol{\tau}$ *maximizing* $J_{\boldsymbol{\tau},\boldsymbol{\theta}}(\mathbf{Y}^o|\boldsymbol{X})$ *when* $\boldsymbol{\theta}$ *is held fixed are obtained thanks to the following fixed point relation:*

$$\hat{\tau}_{iq} \propto \alpha_{iq} \prod_{(i,j)\in\mathcal{D}^o} \prod_{\ell} \exp\left\{(Y_{ij}-1)(\gamma_{q\ell}+\beta^t\boldsymbol{X}_{ij}) + h(\gamma_{q\ell}+\beta^t\boldsymbol{X}_{ij})\right\}^{\hat{\tau}_{j\ell}}. \tag{3.14}$$

3.4 Model selection

The Integrated Classification Likelihood criterion was introduced by Biernacki et al. (2000) for mixture models, where the likelihood – and thus BIC – is usually intractable. Daudin et al. (2008) adapted a variational ICL in the context of SBM. Under MAR condition, this criterion requires a slight adaptation stated in the following proposition.

Proposition 37 (Model selection)**.** *For an SBM with* Q *blocks and for* $\hat{\boldsymbol{\theta}} = \arg\max \log p_{\boldsymbol{\theta}}(\mathbf{Y}^o,\mathbf{Z})$, *the ICL criterion is given by*

$$\text{ICL}(Q) = -2\mathbb{E}_{\tilde{p}_{\boldsymbol{\tau}}}\left[\log p_{\hat{\boldsymbol{\theta}}}(\mathbf{Y}^o,\mathbf{Z};Q,\boldsymbol{X})\right] + pen_Q,$$

and

$$pen_Q = \begin{cases} \left(\frac{Q(Q+1)}{2}\right)\log\text{card}\,(\mathcal{D}^o) + N(Q-1)\log\text{card}\,(\mathcal{N}^o) & \textit{(in Model 1)}, \\ \left(\frac{Q(Q+1)}{2}+m\right)\log\text{card}\,(\mathcal{D}^o) + (Q-1)\log\text{card}\,(\mathcal{N}^o) & \textit{(in Model 2)}, \\ \left(\frac{Q(Q+1)}{2}+m\right)\log\text{card}\,(\mathcal{D}^o) + N(Q-1)\log\text{card}\,(\mathcal{N}^o) & \textit{(in Model 3)}, \end{cases} \tag{3.15}$$

Note that a dyad is only counted once since we work with symmetric networks. The number of blocks chosen is the one associated to the lowest ICL.

4 Illustrations

We consider in this section two cases where the NMAR modeling appears like a good alternative when some covariates are not available.

4.1 Example 1

In this Section we illustrate Proposition 32 on a simple example available in Figure 4.2. In this example, we consider an SBM with covariates with $n = 100$ nodes following Model 1. The network topology is affiliation, which means that diagonal parameters of the connectivity matrix $\boldsymbol{\pi}$ are greater than extra-diagonal parameters. Indeed we choose

$$\boldsymbol{\pi} = \begin{pmatrix} .2 & .05 & .05 \\ .05 & .2 & .05 \\ .05 & .05 & .2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\alpha} = (1/4, 1/2, 1/4). \tag{4.1}$$

Furthermore, the network is sampled according to a block-node sampling with sampling parameters $p = (.8, .2, .5)$ corresponding to probabilities to observe a node conditionally to his covariate. As a consequence, with these parameters, the sampling rate is varying in the interval $(.46, .828]$. Finally the parameter δ is varying and takes three different values : 1, 2/3 and 1/3. The greater δ the less $\mathbf{X}$ depends on $\mathbf{Z}$.

In Figure 4.2, boxplots correspond to estimations of the connectivity matrix on the left and to the adjusted Rand index (in short ARI Rand, 1971) between the true clustering and the predicted one on the right. We run the VEM algorithms for SBM with block-node sampling (NMAR) and with node sampling (MAR) described in (Tabouy et al., 2019a) and finally VEM for SBM with covariates corresponding to Model 1 (see Section 3.1).

Figure 4.2 shows that in this example, making the MAR assumption is not the best way to deal with missing data when covariates are missing. Furthermore, we see that the estimation of the clustering and of $\boldsymbol{\pi}$ is made with the same accuracy for VEM_NMAR and VEMCov1, moreover results do not depend on the value of the parameter δ. Finally, note that the estimations of parameter ρ (results not reported here), linked to parameter p as described in Proposition 32, appears to be good in terms of Frobenius norm when δ is equal to 1 and 2/3 but is degrading when δ gets low as when it is equal to 1/3.

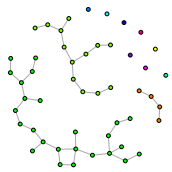
4.2 Example 2

Considering an SBM with covariates on dyads with missing entries, we explore various network topologies (namely affiliation, star and bipartite). The connectivity matrix represented in SBM with covariates context by the matrix $g^{-1}(\gamma)$ of which are given in Figure 4.3. This parameter is classically considered (see Mariadassou et al., 2010) as monitoring heterogeneity in the SBM and then the topology of the network.

In the following, we will consider that each dyad has one covariate and that they are generated conditionally to latent variables $\mathbf{Z}$ as following :

$$\mathbf{X}_{ij} = \mathbb{1}_{\{z_i = z_j\}}(1 - B_{ij}) + (1 - \mathbb{1}_{\{z_i = z_j\}})B_{ij}, \tag{4.2}$$

where B_{ij} are independent Bernoulli variables with the tuning parameter p. According to these covariates, the SBM is as in model 3 with the specificity that we give here the conditional distribution of $\mathbf{X}|\mathbf{Z}$ and not the inverse as suggested in Figure 4.1. A simple application of Bayes formula allow us to find the conditional distribution of $\mathbf{Z}|\mathbf{X}$ if wanted. Coming back to the definition of $\mathbf{X}_{ij}$, note that the smaller the parameter p, the more $\mathbf{X}$ depend on $\mathbf{Z}$. The value $p = 0$ corresponds to the case where $\mathbf{X}$ encode exactly $\mathbf{Z}$. This particular case when $\mathbf{X}$ is observed is called "Oracle" in order to show the response of the VEM described in Section 3.2 when covariates are available. Cases corresponding to $p = .2$ and $p = .5$ are explored in Figure 4.5. In parallel of the Oracle we run VEM for SBM without



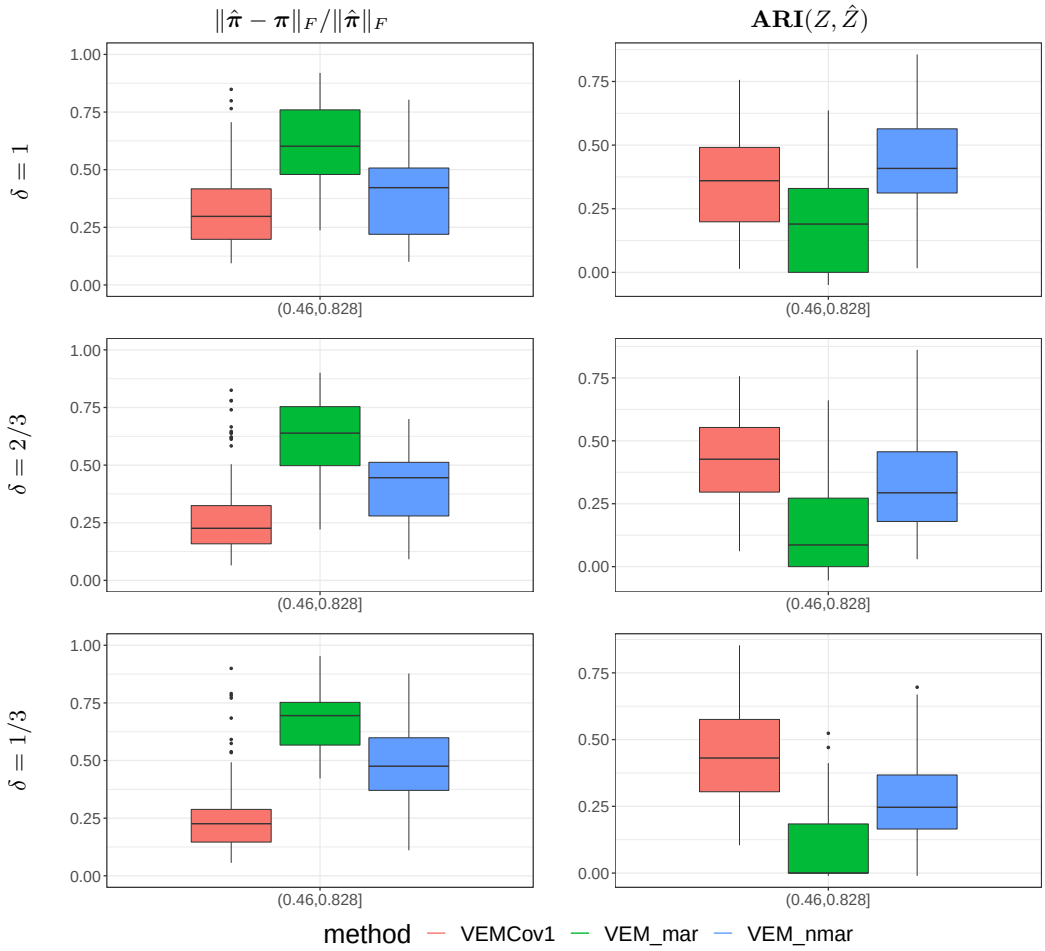


Figure 4.2 – Estimation error of $\boldsymbol{\pi}$ and adjusted Rand index averaged over 100 simulations for affiliation topology in an SBM with covariates (Model 1) and comparison with models whithout covariates (MAR and NMAR).

$$\begin{pmatrix} .3 & .05 & .05 \\ .05 & .3 & .05 \\ .05 & .05 & .3 \end{pmatrix}$$

(a) affiliation

$$\begin{pmatrix} .05 & .3 & .05 & .05 \\ .3 & .05 & .05 & .05 \\ .05 & .05 & .05 & .3 \\ .05 & .05 & .3 & .05 \end{pmatrix}$$

(b) bipartite

$$\begin{pmatrix} .3 & .3 & .03 & .03 \\ .3 & .03 & .07 & .03 \\ .03 & .07 & .3 & .3 \\ .03 & .03 & .3 & .03 \end{pmatrix}$$

(c) star

Figure 4.3 – Matrix $g^{-1}(\gamma)$ in different topologies with inter/intra block probabilities.

covariates and missing data from a double-standard sampling. This design states that every edge is sampled with probability ρ_1 and every non-edge is sampled with probability ρ_0. It implies that missing data are NMAR. We also run random-dyad sampling where every dyad is sampled with probability ρ. It implies that missing data are MAR. More details are available in Tabouy et al. (2019a) about these sampling designs and estimation methods. They are respectively called "NMAR" and "MAR".

Simulated networks have $n = 100$ nodes. Marginal probabilities α are chosen specifically for affiliation, star and bipartite topologies, respectively $(1/3, 1/3, 1/3)$, $(.15, .35, .15, .35)$ and $(1/4, 1/4, 1/4, 1/4)$. The sampling parameters δ is fixed at 1 and κ takes iteratively values $(-1.4, -.8, -.25, .1)$ in order to have a panel of sampling rates. Algorithms are initialized with spectral clustering.

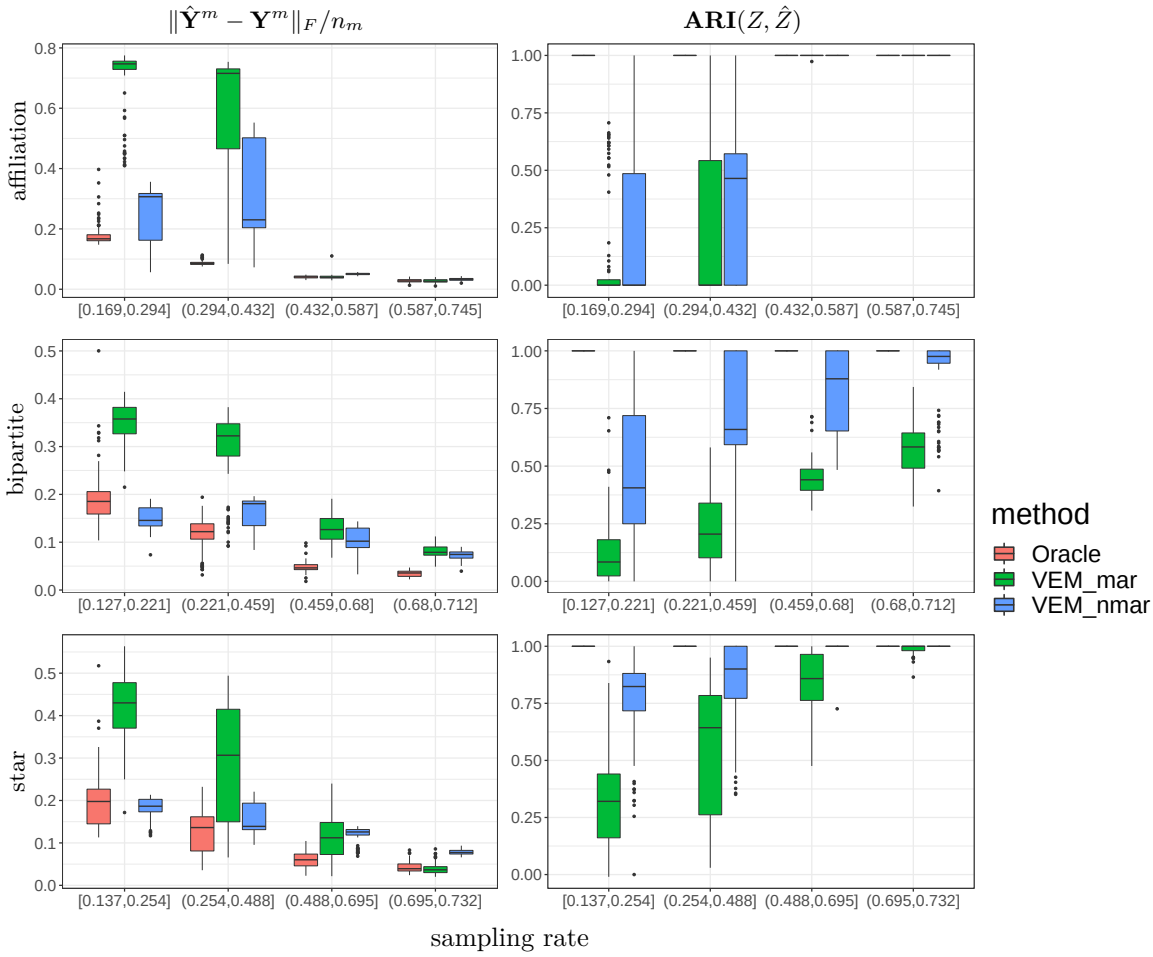


Figure 4.4 – On the left side we estimate the prediction error of missing dyads where n_m is the number of missing dyads and on the right side the ARI between the predicted clustering and the true one. In this case $p = 0$.

Simulations on Figure 4.4 show that "NMAR" algorithm is a good alternative when covariates are unknown whereas "MAR" algorithm does not estimate well parameters and does not predict correctly the clustering. However, the more the sampling rate increases, the more similar are the performance of algorithms "NMAR" and "MAR". On Figure 4.5, where parameter p denoting how much $\mathbf{X}$ encode $\mathbf{Z}$, the difference between "NMAR" and "MAR" decrease when p increase. Indeed, "NMAR" algorithm accuracy is linked to parameter p in the way that it is greater as much as p is close to 0, which corresponds to the case where $\mathbf{X}$

depends the most on $\mathbf{Z}$. We have not yet fully understood the reason for this phenomenon, it would be worth studying.

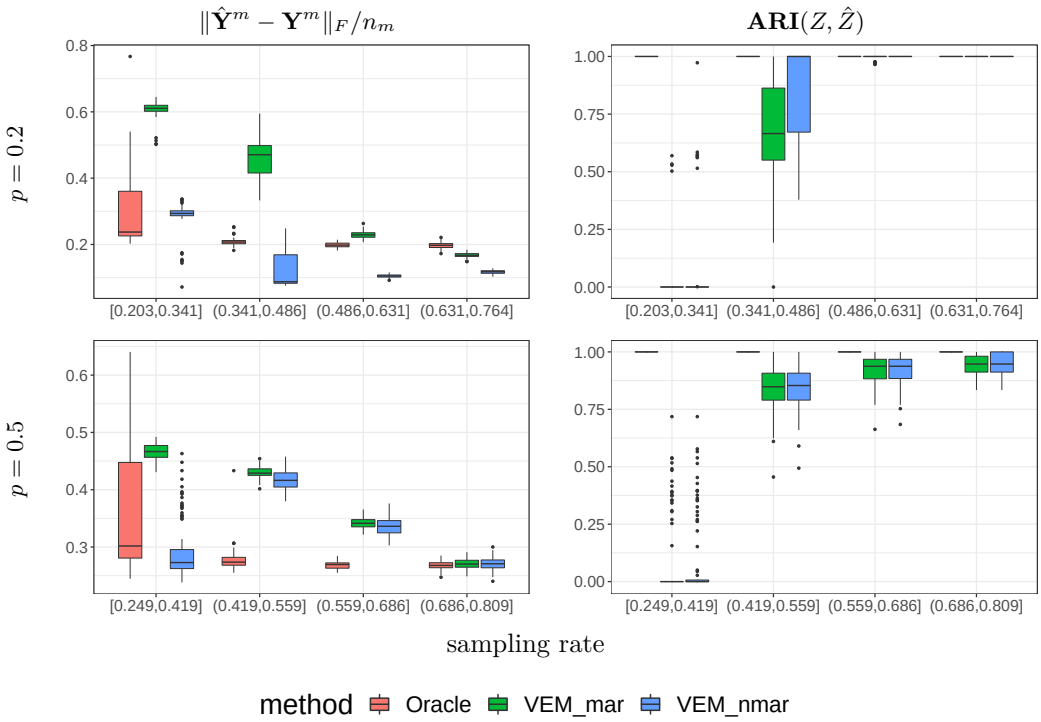


Figure 4.5 – On the left side we estimate the prediction error of missing dyads where n_m is the number of missing dyads and on the right side the ARI between the predicted clustering and the true one. In these cases $p = .2$ and $p = .5$.

5 Conclusion

In this paper we described several combining sampling of a network generated under an SBM with covariates. Algorithms have been proposed to estimate parameters of the models and recover the clustering. Furthermore, we investigate cases where some covariates are not available and show through two examples where NMAR modeling can be a better alternative than MAR modeling. During our exploration, we encounter many simulation settings leading to similar performances for MAR and NMAR models (simulations not reported here). These considerations would deserve further investigations to understand the relatively good robustness of the MAR modeling. Note that however, we never found MAR approximation to be better that NMAR approximation.

Chapter 5

missSBM: An R Package for Handling Missing Values in the Stochastic Block Model

This chapter is available as a preprint on arXiv (Tabouy et al., 2019b).

Contents

The Stochastic Block Model (SBM) is a popular probabilistic model for random graph. It is commonly used to perform clustering on network data by aggregating nodes that share similar connectivity patterns into blocks. When fitting an SBM to a network which is partially observed, it is important to account for the underlying process that originates the missing values, otherwise the inference may be biased. This paper introduces **missSBM**, an R-package fitting the SBM when the network is partially observed, i.e. the adjacency matrix contains not only 1 or 0 encoding presence or absence of edges but also `NA` encoding missing information between pairs of nodes. It implements a series of algorithms for the binary SBM, with the possibility of accounting for covariates if needed, by performing variational inference for several sampling mechanisms, the methodology of which is detailed in Tabouy et al. (2019a). Our implementation automatically explores different block numbers to select the most relevant according to the Integrated Classification Likelihood (ICL) criterion. The ICL criterion can also help to determine which sampling mechanism fits the best the data. Finally, **missSBM** can be used to perform imputation of missing entries in the adjacency matrix. We illustrate the package on a network data set consisting in interactions between blogs sampled during the French presidential election in 2007.

1 Introduction

In many fields of science, networks are a natural way to represent interaction data. To cite a few examples, a network may represent social interactions such as friendship or collaboration between people in a social network, regulation between genes and their products in a gene regulatory network, or predation between animals in a food web. Notice that we only consider here networks which can be represented by graphs composed by simple edges connecting pairs of nodes (also referred to as *dyads* in the following).

At this day, there exist many softwares performing network-related analyzes. Unsurprisingly, the R community is extremely active in this area. Indeed, the R programming language is especially well-designed for performing data manipulation and visualization, and thus appropriate for handling network data. Among the many available R packages related to networks, we suggest a classification into three groups:

i) Packages performing representation, manipulations, visualization tasks, and/or packages computing descriptive statistics on networks. This group clearly occupies the first place in terms of number of packages. We may cite non exhaustively the following top representatives: **igraph** (Csardi and Nepusz, 2006), **network** (Butts, 2008a), **statnet** (Handcock et al., 2008), or **sna** (Butts, 2008b).

ii) Packages fitting (probabilistic) models on network data. Most of the existing packages in this group are dedictated to the estimation of a specific network model: important examples include **ergm** (Hunter et al., 2008), fitting the family of exponential random graph models introduced in Hunter and Handcock (2006b); **mixer** (Ambroise et al., 2015) and **blockmodels** (Leger, 2016), fitting the family of Stochastic Block Models (Holland et al., 1983 and Nowicki and Snijders, 2001); or **latentnet** (Krivitsky and Handcock, 2008), implementing the latent space approach of Hoff et al. (2002).

iii) Packages learning the structure of a network from an external source of data, such as **huge** (Zhao et al., 2012a), or **bnstruct** Franzin et al. (2017). These packages generally rely on an specific graphical modeling of the data (e.g., Gaussian graphical models (Lauritzen, 1996) in **huge**, or Bayesian networks (Pearl, 2011) in **bnstruct**).

In addition to this brief typology, the interested reader may consult the CRAN task view on the related topic of graphical modeling (Hojsgaard, 2019).

The package **missSBM** that we introduce here belongs to the second category, that is, softwares that fit a probabilistic model to network data. **missSBM** is dedicated to the estimation of the stochastic block model (SBM). The SBM is a mixture of Erdős-Rényi random graphs (Erdős and Renyi, 1959) that allows a high degree of heterogeneity in connectivity profiles. As a consequences it generally fits well real-world network data, while retaining the advantage of being a generative model (contrary to mechanistic approaches such as the Barabási-Albert model (Albert and Barabási, 2002), defined by a preferential attachment algorithm). The main outcome of the SBM inference on a network is a clustering of its nodes – or "blocks" sharing the same connectivity properties.

Even though there already exist efficient R packages for SBM inference such as **Blockmodels** (Leger, 2015) and **mixer** (Ambroise et al., 2015), they can only be applied to complete data sets. The main feature introduced in **missSBM** is to deal with cases where the network data is only partially observed. More precisely, we consider situations where the adjacency matrix encoding the network data contains contains not only 1 or 0 encoding presence or absence of edges but also `NA` encoding missing information between dyads. In the presence of missing data, it is important to account for the underlying process that originates the missing values in the estimation of a probabilistic model, otherwise estimation of the model parameters may be biased. In particular, one has to take the type of missing data mechanisms into account (Missing at Random or Not, see Rubin, 1976). This issue has been

studied in the context of network data in Handcock and Gile (2010) for exponential random graph models and in our previous paper Tabouy et al. (2019a) for stochastic block models. The package **missSBM** is an implementation of the methodology developed therein.

Specifically, **missSBM** implements variational algorithms in the vein of Daudin et al. (2008) for estimating the SBM for binary network data, with or without covariates, under various missing data mechanisms. This includes cases of incomplete data where the inference can be made only on the observed part of the data (Missing at Random), or cases where it is necessary to take the sampling design in the inference into account (Not Missing at Random). Although version `0.2.0` of **missSBM** only deals with binary networks (either directed or not), we deploy a structure that let the possibility to easily include other variants in the future by adopting an oriented-object programming spirit thanks to R6-classes and the **R6** package of Chang (2017). In particular, extending **missSBM** to weighted SBM with exponential distributions of the edges (Mariadassou et al., 2010) should be straightforward.

The paper is organized as follows: Section 2 briefly introduces the statistical framework of the binary SBM, with or without covariates, and summarizes the key points for estimating the SBM under missing data condition, further detailed in Tabouy et al. (2019a). Section 3 presents the package structure; Section 4 provides basic user guidelines. We finally detail in Section 5 a case study which analyzes a network data set describing the French blogosphere during French presidential election of 2007, illustrating the most striking features of the package.

2 Statistical Framework

2.1 Binary Stochastic Block Model (SBM)

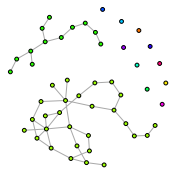
In an SBM, nodes from a set $\mathcal{N} \triangleq [\![1, n]\!]$ are distributed among a set $\mathcal{Q} \triangleq [\![1, Q]\!]$ of hidden blocks that model the latent structure of the graph. The group memberships are described by categorical variables $(\mathbf{Z}_i, i \in \mathcal{N})$ with multinomial distribution $\mathcal{M}(1, \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_Q))$. The probability of an edge between any pair of nodes (or *dyad*) in $\mathcal{D} \triangleq \mathcal{N} \times \mathcal{N}$ only depends on the blocks the two nodes belong to. Hence, the presence of an edge between i and j, indicated by the binary variable Y_{ij}, is independent on the other edges conditionally on the latent blocks:

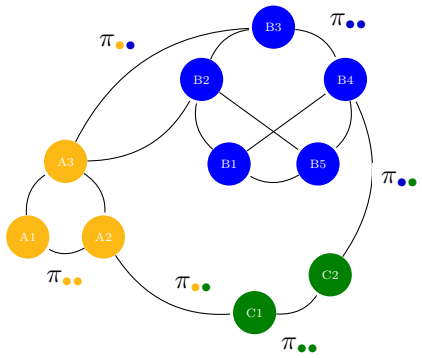
$$Y_{ij} | \mathbf{Z}_i, \mathbf{Z}_j \sim^{ind} \mathcal{B}(\pi_{\mathbf{Z}_i \mathbf{Z}_j}), \ \forall (i, j) \in \mathcal{N}^2, \tag{2.1}$$

where $\mathcal{B}$ stands for the Bernoulli distribution. In the following, we denote by $\boldsymbol{\pi} = (\pi_{q\ell})_{(q,\ell) \in \mathcal{Q}^2}$ the connectivity matrix, $\boldsymbol{\alpha}$ the mixture parameters, $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$ the $n \times Q$ membership matrix and $\mathbf{Y} = (Y_{ij})_{(i,j) \in \mathcal{D}}$ the $n \times n$ adjacency matrix. The vector encompassing all the unknown model parameters is $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$. A schematic representation of the binary SBM in the undirected case is given in Figure 5.1, where we highlight the latent clustering.

2.2 Accounting for External Covariates

On top of information about connections between nodes, it is common for network data to be accompanied with additional information on nodes or dyads, that we call *covariates*: for instance in social network, nodes may belong to different categories (gender, occupation, nationality). Covariates on dyads usually represent distances between nodes covariates: for example in a context of spatial data where nodes corresponds to entities that have explicit geographic locations, dyad covariates may be the distances between the nodes. Depending on the analysis, we may want to detect a connectivity pattern beyond the covariate effect. To do so, we present here a variant of the SBM implemented in **missSBM** that allows the user to include covariates in the model. Let $\mathbf{X}_{ij}$ denote the vector of length m of covariates for dyad (i, j). If the covariates correspond to the nodes, i.e. $\mathbf{X}_i \in \mathbb{R}^N$ is associated with node i for all $i \in \mathcal{N}$, they are transferred on the dyad level through a symmetric "similarity" function





- $\mathcal{Q} = \{\bullet, \bullet, \bullet\}$ blocks
- $\alpha_\bullet = \mathbb{P}(i \in \bullet)$, $\bullet \in \mathcal{Q}, i = 1, \ldots, n$
- $\pi_{\bullet\bullet} = \mathbb{P}(i \leftrightarrow j | i \in \bullet, j \in \bullet)$

Figure 5.1 – Schematic representation of an undirected network following the stochastic block model with 3 blocks. Colors are blocks in which nodes are dispatched with probabilities $\boldsymbol{\alpha}$ and dyads distribution between nodes depends on colors of nodes with probabilities $\boldsymbol{\pi}$.

$\phi(\cdot,\cdot) : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^m$: $\mathbf{X}_{ij} \triangleq \phi(\mathbf{X}_i, \mathbf{X}_j)$. In the following, $\mathbf{X} \triangleq [\mathbf{X}_{ij}]_{i,j\in\mathcal{N}} \in (\mathbb{R}^m)^{n\times n}$ denotes the covariates. An SBM accounting for covariates is

$$\begin{aligned} \mathbf{Z}_i &\sim^{\text{iid}} \mathcal{M}(1, \boldsymbol{\alpha}), \quad \forall i \in \mathcal{N}, \\ Y_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, \mathbf{X} &\sim^{\text{ind}} \mathcal{B}(g(\gamma_{z_i z_j} + \boldsymbol{\beta}^\top \mathbf{X}_{ij}), \quad \forall (i,j) \in \mathcal{N}^2, \end{aligned} \tag{2.2}$$

where $\gamma_{q\ell} \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^m$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_Q)$, $g(x) = (1 + e^{-x})^{-1}$. In this case the vector of unknown parameters is defined by $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$. Note the connection between this model and logistic regression. In this framework, heterogeneity is modeled by parameters $(\gamma_{q\ell})_{q\ell}$ corresponding to inter and intra block probabilities (or connectivity matrix) in the binary SBM. It accounts for heterogeneity that is not explained by the regression term $\beta^T \mathbf{X}_{ij}$.

2.3 Missing Data and SBM

The main feature of **missSBM** is to deal with missing values to perform unbiased estimation of the parameters of the model underlying an incompletely observed network. The sampled data can be encoded in an adjacency matrix $\mathbf{Y}$ where missing information – here dyads not sampled – is encoded by `NA`'s. We also define the $n \times n$ sampling matrix $\mathbf{R}$ such as $R_{ij} = 1$ if the dyad Y_{ij} is sampled and $R_{ij} = 0$ otherwise. For convenience we define $\mathbf{Y}^o = \{Y_{ij} : R_{ij} = 1\}$ and $\mathbf{Y}^m = \{Y_{ij} : R_{ij} = 0\}$ the respective sets of observed and non-observed dyads. At this stage it is important to notice that the number of nodes n is assumed to be known.

In our framework, a sampling design is a stochastic process that generates $\mathbf{R}$. We then rely on the standard missing data theory of Rubin (1976) to classify those designs either into Missing Completely At Random (MCAR), Missing At Random (MAR) or Not Missing At Random (NMAR) cases. We summarize the analyzes conducted in Tabouy et al. (2019a) on missing data for network as follows:

$$\text{Sampling design for SBM is } \begin{cases} \text{MCAR} & \text{if } \mathbf{R} \perp\!\!\!\perp (\mathbf{Y}, \mathbf{Z}), \\ \text{MAR} & \text{if } \mathbf{R} \perp\!\!\!\perp (\mathbf{Y}^{\text{m}}, \mathbf{Z}) \mid \mathbf{Y}^{\text{o}}, \\ \text{NMAR} & \text{otherwise.} \end{cases} \tag{2.3}$$

Note that MCAR missingness is a particular case of MAR missingness. Denoting by $\boldsymbol{\psi}$ the set of parameters associated with the sampling design distribution that generates $\mathbf{R}$, we assume that $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are living in a product space, so that we can derive the following proposition:

Proposition 38. *From (2.3), if the sampling is MAR or MCAR then maximizing* $p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{R})$ *or* $p_{\boldsymbol{\theta}}(\mathbf{Y}^{\mathrm{o}})$ *in* $\boldsymbol{\theta}$ *is equivalent.*

In words, the inference can be conducted on the observed part of the network data when the sampling is MAR or NMAR without incurring any biased. In these cases, adaptation of existing algorithm for SBM inference is straightforward. NMAR sampling designs require, however, more refined inference strategies.

2.4 Examples of Sampling Designs for Networks

This section reviews a series of stochastic processes – or sampling designs – available in **missSBM** either for sampling an existing network or to be accounted for in the inference of an SBM model. Note that the sampling design may depend either on *i*) the values of the dyads in the network; *ii*) the latent clustering of the nodes; or *iii*) some covariates. We specify for each case when the sampling is MAR or NMAR. The sampling design examples detailed in the following assume the independence between dyad observation for dyad-centered sampling designs and between node observation for node-centered sampling design conditionally on $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{X}$.

Dyad-Centered Sampling Designs

- **Dyad sampling** (MAR): each dyad $(i, j) \in \mathcal{D}$ has the same probability $\mathbb{P}(R_{ij} = 1) \triangleq \psi$ to be observed.
- **Double standard sampling** (NMAR): let $\psi \triangleq (\rho_1, \rho_0) \in [0, 1]^2$. Double standard sampling consists in observing dyads with probabilities

$$\mathbb{P}(R_{ij} = 1|Y_{ij} = 1) = \rho_1, \qquad \mathbb{P}(R_{ij} = 1|Y_{ij} = 0) = \rho_0.$$

  The probabiliy for sampling a dyad thus intrinsically depend on the presence/absence of the corresponding edge.
- **Block-dyad sampling** (NMAR): this sampling consists in observing all dyads with probabilities $\psi \triangleq (\psi_{q\ell})_{(q,\ell)\in\mathcal{Q}^2}$ such that

$$\psi_{q\ell} = \mathbb{P}(R_{ij} = 1 \mid Z_{iq} = 1, Z_{j\ell} = 1).$$

  Thus sampling thus depends on the underlying clustering of the network.
- **Covar-dyad sampling** (MAR):let us define $\psi \triangleq (\alpha, \kappa) \in \mathbb{R} \times \mathbb{R}^m$. In this sampling, the probability to observe a dyad is driven by the effect of a given covariate:

$$\mathbb{P}(R_{ij} = 1|\mathbf{X}_{ij}) = g(\alpha + \kappa^t \mathbf{X}_{ij}).$$

Node-Centered Sampling Designs

A node-centered sampling consists in drawing some nodes to observe with probabilities given by the sampling design. Observing a node means observing all the dyads implying this node. For all $i \in \mathcal{N}$, we denote by $V_i = 1$ if the node i is observed, $V_i = 0$ otherwise. Then, if $V_i = 1$ we have $R_{ij} = 1$ for all $j \in \mathcal{N}$.

- **Node sampling** (MAR): the probabilities for observing nodes are uniform: $\mathbb{P}(V_i = 1) = \psi$ for all $i \in \mathcal{N}$.
- **Degree sampling** (NMAR): for all node $i \in \mathcal{N}$, $\mathbb{P}(V_i = 1) = \rho_i$ where $(\rho_1, \ldots, \rho_n) \in [0, 1]^n$ are such that $\rho_i = g(a + bD_i)$ for all $i \in \mathcal{N}$ where $\psi \triangleq (a, b) \in \mathbb{R}^2$ and $D_i = \sum_j Y_{ij}$.

- **Block-node sampling** (NMAR): this sampling consists in observing all dyads corresponding to nodes selected with probabilities $\psi \triangleq (\psi_1, \ldots, \psi_Q) \in [0,1]^Q$ such that $\psi_q = \mathbb{P}(V_i = 1 \mid Z_{iq} = 1)$ for all $(i,q) \in \mathcal{N} \times \mathcal{Q}$.

- **Covar-node sampling** (MAR): let $\psi \triangleq (\nu, \eta) \in \mathbb{R} \times \mathbb{R}^N$. The probability to observe a node is
$$\mathbb{P}(V_i = 1 | \mathbf{X}_i) = g(\nu + \eta^t \mathbf{X}_i).$$

2.5 Estimation Procedure

Optimization: a Variational EM

The SBM is a latent state space model which can be seen as a mixture model for random graphs. As such the EM algorithm (Dempster et al., 1977) is the natural choice for the inference. It is based on the evaluation of the expectation of the complete loglikelihood of the model, with respect to the conditional distribution of the latent variables given the data. However, this expectation is intractable in the SBM due to the structure of dependency between the latent variables $\mathbf{Z}$ and the network $\mathbf{Y}$. In fact, it would require to sum over all possible clustering for all pairs of nodes, which is out of reach even for a moderate number of nodes or clusters. To address this shortcoming in the complete data situation, Daudin et al. (2008) introduced a *variational* EM, based on the variational principles of Jordan et al. (1998). The idea is to maximize a lower bound of the loglikelihood based on an approximation of the true conditional distribution of the latent variable $\mathbf{Z}$.

In the case of an SBM with missing data, the level of difficulty is higher since the set of latent variables encompasses both $\mathbf{Z}$ (the latent clustering) and $\mathbf{Y}^{\mathrm{m}}$ (the missing dyads). We propose here a variational distribution of the conditional distribution $p_\theta(\mathbf{Z}, \mathbf{Y}^{\mathrm{m}} | \mathbf{Y}^{\mathrm{o}})$ where complete independence is forced on $\mathbf{Z}$ and $\mathbf{Y}^{\mathrm{m}}$, using a multinomial $m(\cdot)$, respectively a Bernoulli $b(\cdot)$ distribution for $\mathbf{Z}$ and $\mathbf{Y}^{\mathrm{m}}$:

$$\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}(\mathbf{Z}, \mathbf{Y}^{\mathrm{m}}) = \tilde{p}_{\boldsymbol{\tau}}(\mathbf{Z})\ \tilde{p}_{\boldsymbol{\nu}}(\mathbf{Y}^{\mathrm{m}}) = \prod_{i \in \mathcal{N}} m(\mathbf{Z}_i; \tau_i) \prod_{(i,j) \in \mathcal{D}^{\mathrm{m}}} b(Y_{ij}; \nu_{ij}),$$

where $\boldsymbol{\tau} = \{\tau_i, i \in \mathcal{N}\}$ and $\boldsymbol{\nu} = \{\nu_{ij}, (i,j) \in \mathcal{D}^{\mathrm{m}}\}$ are two sets of variational parameters respectively associated with $\mathbf{Z}$ and $\mathbf{Y}^{\mathrm{m}}$. Interestingly, $\boldsymbol{\tau}$'s are proxies for the posterior probabilities of memberships for all nodes, and $\boldsymbol{\nu}$'s correspond to the imputed values of the missing dyads in the network data. This approximation leads to the following lower bound $\mathcal{J}$ of the loglikelihood, where KL is the Kullback-Leibler divergence between the true and approximated conditional distribution:

$$\begin{aligned}
&\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{R}) \geq \\
&\quad J_{\boldsymbol{\tau},\boldsymbol{\nu},\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{R}) \triangleq \log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{R}) - \mathrm{KL}(\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\psi}}(\mathbf{Z}, \mathbf{Y}^{\mathrm{m}}) \| p_{\boldsymbol{\theta}}(\mathbf{Z}, \mathbf{Y}^{\mathrm{m}} \| \mathbf{Y}^{\mathrm{o}})), \\
&\qquad = \mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}} \left[\log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{Y}^{\mathrm{o}}, \mathbf{R}, \mathbf{Y}^{\mathrm{m}}, \mathbf{Z})\right] - \mathbb{E}_{\tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}} \left[\log \tilde{p}_{\boldsymbol{\tau},\boldsymbol{\nu}}(\mathbf{R}, \mathbf{Y}^{\mathrm{m}})\right].
\end{aligned}$$

Note that when we chose $\tilde{p}$ to be the true conditional distribution of the latent variables $\mathbf{Z}, \mathbf{Y}^{\mathrm{m}}$, we meet again the quantity maximized in the standard EM algorithm.

Based on this approximation, the variational EM algorithm consists in alternatively updates of the variational parameters $\{\boldsymbol{\tau}, \boldsymbol{\nu}\}$ (the VE-step) and updates of $\boldsymbol{\theta}, \boldsymbol{\psi}$ maximizing $\mathcal{J}_{\boldsymbol{\tau},\boldsymbol{\theta}}$ (the M-step). Steps VE and M are iterated until convergence like in a standard EM. The algorithm converges to a local maximum of the lower bound of the loglikelihood. This variational principle is translated into a series of algorithms for handling missing data with all sampling designs introduced in Section 2.4. We underline that some time consuming parts of the VEM algorithms are coded in `C++` using the **Rcpp** package (Eddelbuettel and François, 2011) to interface `C++` with `R`.

Initialization

It is well known that algorithms based on EM algorithm have a great sensitivity to initialization. This step therefore requires a special attention. In **missSBM** we implemented several methods for initializing the variational EM: the Absolute Eigenvalues Spectral Clustering (detailed in Rohe et al., 2010); a Hierarchical Ascending Classification (HAC) based on Manhattan norm between nodes; and the K-means algorithm. Experiences showed us that the HAC algorithm is sometimes more robust in cases where the network contains a large proportion of missing dyads. However, in general, we recommend the use of the Absolute Eigenvalues Spectral Clustering, which is used by default in the package.

Selection of the Number of Blocks

A main difficulty met when conducting SBM inference lies in the estimation of the number of blocks, generally unknown to the user. To remedy this problem we use the Integrated Classification Likelihood (ICL) criterion defined in Biernacki et al. (2000) and routinely used in the framework of mixture models. More precisely, we implement in **missSBM** an exploration procedure designed to produce a smooth and robust ICL curve, by avoiding to get stuck in local minima. This "smoother" is divided into two steps, forward and backward: The forward step creates new initializations for each number Q of block considered, by splitting blocks obtained from estimations with $Q-1$ blocks. On the other hand, the backward step tries new initializations for each Q by merging groups of the model with $Q+1$ groups. The best model in terms of ICL is always retain. The procedure can be iterated until an satisfying smoothing is reached.

3 Structure of the Package

The R package **missSBM** is coded using object-oriented programming with **R6** package (Chang, 2017). The hierarchical structure and inheritance relations are represented in Figure 5.2.

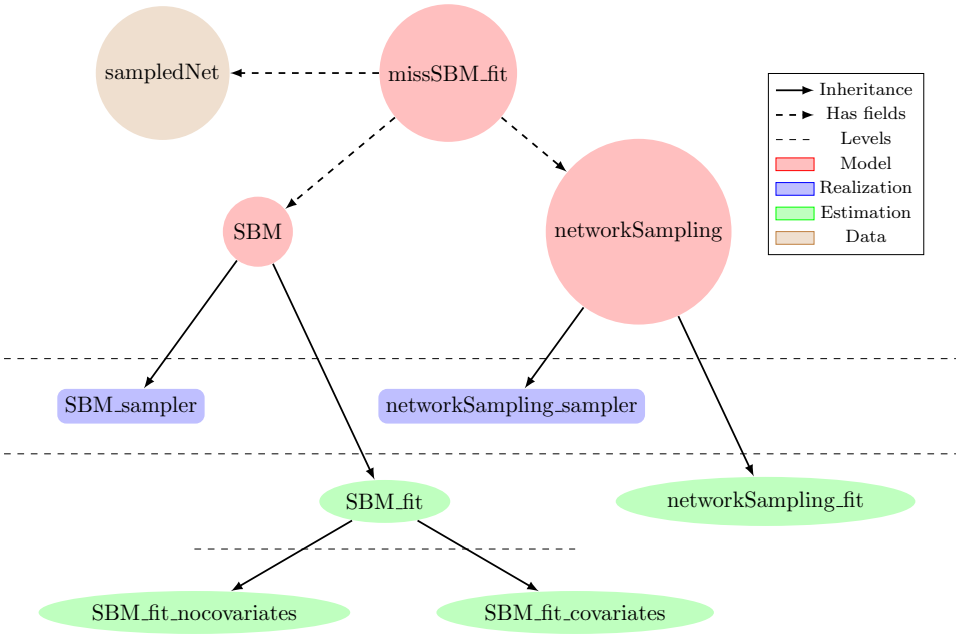


Figure 5.2 – Diagram of relation between classes of objects in **missSBM**, with three levels of inheritance.

The package contains two general **R6** classes: `SBM` and `networkSampling`. These two classes instantiate objects that collect all parameters needed to respectively define an SBM and a sampling design, as depicted in Sections 2.1 and 2.2 for SBMs and in Section 2.4 for the sampling design. These classes both give birth to two child classes, a `sampler` class and a `fit` class, which inheritate of all methods and fields of their respective mother classes. The `sampler` class allows the user to obtain a realization of network data sampled under an SBM (*i.e.* an adjacency matrix $\mathbf{Y}$), and a realization of the sampling of a network (*i.e.* a sampling matrix $\mathbf{R}$). The `fit` classes contains methods for estimating the SBM and the sampling design parameters. Finally, `SBM_fit` gives birth to two child classes: `SBM_fit_nocovariates` and `SBM_fit_covariates`, respectively dedicated to the estimation of binary SBM with or without covariates. Note that the two classes `SBM` and `network_Sampling` can be used totally independently. In particular, it is possible to fit SBM without missing data. In order to perform SBM inference when the network is partially observed, two others classes of object are introduced: `sampledNetwork` and `missSBM_fit`. The former instantiates an object which collects any information about presence or absence of dyads and nodes in a sampled network. The latter makes the connection between objects of classes `SBM` and `network_sampling`, in order to make it possible to infer SBM from partially-observed network data, under several sampling designs.

4 Guidelines for Users

The package **missSBM** allows the user to do three differents actions: *simulate* a network under the SBM, *sample* a network according to a panel of sampling designs, and *estimate* an SBM from partially observed network data. To do so, three standard R functions are exported to the user, which use internally the classes of object detailed in Section 3. We describe in the following the usage of these functions plus some additional auxiliary functions, as well as their connections with models defined in Section 2.

4.1 Simulate a network

The function `simulate` draws a realization of an SBM, with the following usage:

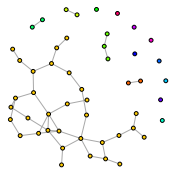
```
missSBM::simulate(nNodes, mixtureParam, connectParam, directed = FALSE,
                  covariates = NULL, covarParam = NULL)
```

The arguments of `simulate` and their mathematical counterpart from (2.1) and (2.2) are described in Table 5.1.

Argument	Description	Correspondence	Lives in
`n`	number of nodes	n	$\mathbb{N}$
`alpha`	mixture parameters	$\boldsymbol{\alpha}$	$[0,1]^Q$
`pi`	`matrix` of inter and intra clusters probabilities	$\boldsymbol{\pi}$	$\mathcal{M}_Q([0,1])$
`directed`	whether the network is directed or not		$\{\mathrm{T}, \mathrm{F}\}$
`covariates`	`list` of covariates	$(\mathbf{X}_i)_{i\in\mathcal{N}}$, $(\mathbf{X}_{ij})_{ij}$	$(\star)$
`covarParam`	regression parameters of the covariates	$\boldsymbol{\beta}$	$\mathbb{R}^m$

Table 5.1 – Arguments of `simulate` and their counterparts in the SBM. $(\star)$ is a `list` with N entries (the N covariates corresponding to nodes) or a list of m $n \times n$matrices (covariates corresponding to dyads).

Note that the network simulated are by default undirected and without covariates. In the case of an SBM with covariate(s), `covarParam` *must not* include intercepts since this role is played by parameters $(\gamma_{q\ell})_{q,\ell}$ in model (2.2).



The output of `simulate` is an **R6** object belonging to the class `SBM_sampler` (see Figure 5.2). An non-exhaustive list of the most useful fields, accessible via $ are presented in Table 5.2. A `plot` method is also available which allows the user to either draw the network by reordering rows and columns of the adjacency matrix $\mathbf{Y}$ according to the nodes memberships, or draw the predicted connectivity matrix $\mathbf{ZYZ}^\top$, i.e. the probability of connection for each nodes once reordered block-wise. This drawing is a simple yet powerful way to visualize the structure of large networks.

Field	Description	Correspondence
`adjacencyMatrix`	adjacency matrix of the simulated network	$\mathbf{Y}$
`blocks`	matrix of blocks memberships	$\mathbf{Z}$
`memberships`	vector of blocks memberships	$(\texttt{which.max}(\boldsymbol{\tau}_i))_{i\in\mathcal{N}}$
Method	**Description**	
`plot(object, type)`	draw $\mathbf{Y}$ re-ordered by blocks if `type`="network" and $\mathbf{ZYZ}^\top$ if `type`="connectivity"	

Table 5.2 – Selection of important fields and methods in class `SBM_sampler` with their mathematical counterparts

4.2 Sampling a network

The function `sample` draws a realization of a sampling matrix $\mathbf{R}$ from a distribution chosen from the series of sampling designs defined in Section 2.4. The usage is the following:

```
missSBM::sample(adjacencyMatrix, sampling, parameters, clusters = NULL,
                covariates = NULL, similarity = l1_similarity, intercept = 0)
```

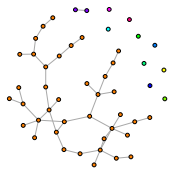
Table 5.3 describes the arguments of `sample` and their mathematical counterparts.

Argument	Description	Correspondence	Lives in
`adjacencyMatrix`	adjacency matrix the network	$\mathbf{Y}$	$\mathcal{M}_n(\{0,1\})$
`sampling`	name of the sampling design		$(\star\star)$
`parameters`	sampling parameter(s)	$\boldsymbol{\psi}$	Ψ
`clusters`	`vector` of blocks memberships	$(\texttt{which.max}(\boldsymbol{\tau}_i))_{i\in\mathcal{N}}$	$\mathcal{Q}^n$
`covariates`	`list` of covariates	$(\mathbf{X}_i)_{i\in\mathcal{N}}$, $(\mathbf{X}_{ij})_{ij}$	$(\star)$
`similarity`	similarity function for covariates on dyads	ϕ	$\mathbb{R}^N \times \mathbb{R}^N \mapsto \mathbb{R}^m$

Table 5.3 – Arguments of `sample` and their counterparts in the SBM. $(\star\star)$ possible values for `sampling` are characters string in "dyad", "double-standard", "block-dyad", "node", "block-node", "degree", "covar-node" and "covar-dyad" for SBM without covariate, and "dyad", "node", "covar-node" and "covar-dyad" SBM with covariates.

Note that the dimension (or `length`) of `parameters` depends on the sampling design selected, as described in Section 2.4. Argument `clusters` only needs to be specified for "block-dyad" and "block-node" sampling designs. Arguments `covarMatrix` is by-default `NULL`, `covarSimilarity` is set to the `l1_similarity` function defined as follows: $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto -|x - y|_{\ell^1} \in \mathbb{R}^d$ and finally `intercept` is set to 0. These last three arguments only needs to be specified in a context of SBM with covariate(s). Note that the intercept is not included in `parameters` and must be specified independently.

The output of `sample` is an **R6** object belonging to the class `sampledNetwork` (see Figure 5.2). An non-exhaustive list of the most useful fields is given in Table 5.4, where $\odot$ stands for the Hadamard (or element-wise) product. A `plot` method is defined for `sampledNetwork` objects, drawing the adjacency matrix $\mathbf{Y}$ of the network with missing entries induced by the sampling matrix $\mathbf{R}$.



Field	Description	Correspondence
`adjacencyMatrix`	sampled adjacency matrix with `NA` at missing entries	$\mathbf{Y} \odot \mathbf{R}$
`covarMatrix`	the matrix of covariates (if applicable)	$\mathbf{X}$
`covarArray`	the array of covariates (if applicable)	$(\mathbf{X}_{ij})_{ij}$
`missingDyads`	the set of dyades not sampled in the network	$\mathbf{Y}^m$
`observedDyads`	the set of dyades sampled in the network	$\mathbf{Y}^o$
`observedNodes`	vector of sampled nodes	$\mathbf{V}$
`samplingRate`	rate of observed dyades	$\texttt{mean}(\mathbf{R})$
`samplingMatrix`	sampling matrix	$\mathbf{R}$
Method	**Description**	
`plot(object)`	draw the adjacency matrix $\mathbf{Y}$ and the sampling matrix $\mathbf{R}$	

Table 5.4 – Selection of important fields and methods in class `sampledNetwork`.

4.3 Parameters estimation, prediction and clustering

The purpose of the function `estimate` is to perform variational inference of a SBM from sampled adjacency matrix, with the following usage:

```
missSBM::estimate(sampledNet, vBlocks, sampling, clusterInit = "spectral",
                  useCovariates = TRUE, control = list())
```

Table 5.5 provides a basic description of the arguments of `estimate`.

Argument	Description	Class
`sampledNet`	the sampled network data	`sampledNetwork`
`vBlocks`	vector of block numbers for model exploration	`vector of integer`
`sampling`	name of the sampling design	`character`
`clusterInit`	"spectral", "hierarchical", "kmeans" or list of clusterings for each Q in `vBlocks`	`character`
`useCovariates`	specify if covariates are included in SBM or not	`boolean`
`control`	tune the optimization and the variational EM algorithm	`list`

Table 5.5 – Arguments of `sample` with short descriptions and types.

Data preparation. When the network data is not obtained from the function `sample` but from real-world network, the user has to format this data to an object with class `sampledNetwork`. We implemented a simple function for this task, taking as an input an adjacency matrix filled with $\{0, 1, \mathrm{NA}\}$:

```
missSBM::prepare_data(adjacencyMatrix, covariates = NULL,
                      similarity = missSBM:::l1_similarity)
```

Tuning the optimization. The argument `control` is a `list` to finely tune the variational EM algorithm, with the following entries:

(i) `threshold`: optimization stops when a V-EM step changes the objective function by less than `threshold`, (default $= 1.10^{-3}$);

(ii) `maxIter`: optimization stops when the number of iteration exceeds `maxIter`, default $= 50$ if `useCovariates` is `TRUE`, 100 otherwise;

(iii) `fixPointIter`: number of iterations in the fix point algorithm used to solve the Variational E step, $= 2$ if `useCovariates` is `TRUE`, 5 otherwise;

(iv) `cores`: number of thread be ran in parallel, = 1 by default;

(v) `trace`: integer for verbosity (0, 1, 2), useless when `cores` > 1.

Estimation ouputs: classes `missSBM_collection` and `missSBM_fit`. The output of `estimate` is an **R6** object with class `missSBM_collection`, fields of which are described in Table 5.6.

Field	Description
`models`	a `list` of models with class `missSBM_fit`
`ICL`	the `vector` of ICL associated to the models in the collection
`bestModel`	best model according to the ICL (a `missSBM_fit` object)
`optimizationStatus`	a `data.frame` summarizing the optimization process for all models

Table 5.6 – Structure of `missSBM_collection` (output of `estimate`).

Among the fields of `missSBM_collection`, the field `models` is a list with `length(vblocks)` elements which are **R6** objects with class `missSBM_fit` (see Figure 5.2). These `missSBM_fit` objects, fields of which are detailed in 5.7, contain all the results of the inference for a fixed number of block Q.

Field	Description	R6 object / class
`fittedSBM`	the adjusted Stochastic Block Model	`SBM_fit_(no)covariates`
`fittedSampling`	the estimated sampling process	`networkSampling_fit`
`sampledNet`	the sampled network data	`sampledNetwork`
`imputednetwork`	the adjacency matrix with imputed values	`matrix`
`monitoring`	status of the optimization process	`data.frame`
`vICL`	ICL criterion associated to Q	`double`
`vBound`	value of the variational bound $\mathcal{J}_{\boldsymbol{\tau},\boldsymbol{\theta}}$ at each step	`double`
`vExpec`	value of $\mathbb{E}_{\tilde{p}_{\boldsymbol{\tau}}}\left[\log(p_{\boldsymbol{\theta}}(\mathbf{Y},\mathbf{Z}))\right]$ at each step	`double`
`penalty`	penalty of the model with Q blocks	`double`

Table 5.7 – Selection of fields in object `missSBM_fit` with descriptions and types.

We finally give additional details on fields `fittedSBM` and `fittedSampling` in Table 5.8.

R6 object	Field	Description	Correspondence
`fittedSBM`	`blocks`	estimated probability of block belonging	$\{\boldsymbol{\tau}_i\}_{i\in\mathcal{N}}$
	`connectParam`	connectivity matrix	$\hat{\boldsymbol{\pi}}$
	`connectProb`	imputed adjacency matrix	$\mathbf{Y}^o \cup \{\nu_{ij}\}_{(i,j)\in\mathcal{D}^m}$
	`covarParam`	regression parameter	$\hat{\boldsymbol{\beta}}$
	`memberships`	vector of blocks memberships	$(\texttt{which.max}(\boldsymbol{\tau}_i))_{i\in\mathcal{N}}$
	`mixtureParam`	mixture parameters	$\hat{\boldsymbol{\alpha}}$
`fittedSampling`	`parameters`	sampling parameter	$\hat{\boldsymbol{\psi}}$

Table 5.8 – Selection of fields in `fittedSBM` and `fittedSampling` with their mathematical counterparts.

Smoothing and model exploration. At the end of the estimation process, it is common that the algorithm get stuck in some local minima for some values of Q, the number of blocks. The consequence is an "non-smooth" ICL curve, which is theoretically supposed to be convex. We thus provide to the user a post-processing function to "smooth" the ICL by exploring additional models with new initialization. Smoothing of the ICL curve is performed with function `smooth`, detailed in the following :

```
missSBM::smooth(Robject, type = c("forward", "backward", "both"),
                control = list())
```

The basic idea is to apply a split and/or merge strategy to the path of models in a collection of SBM, in order to find better initialization. This should result in a "smoothing" of the ICL, that should be close to convex. Arguments of `smooth` are described in Table 5.9.

Argument	Description	R6 object / class
`Robject`	an output from `estimate`	`missSBM_collection`
`type`	kind of ICL smoothing : "forward", "backward" or "both"	`character`
`control`	controlling the variational EM algorithm	`list`

Table 5.9 – Arguments of `smooth` function with short descriptions and their types.

This function acts directly on `Robject` (*i.e.* by reference). The argument `type` is by-default set to "forward". Finally, `control` is a `list` composed of 3 elements : (i) `cores` the number of R processes allowed set by-default to 1, (ii) `iterates` the number of iteration by-default set to 1 and (iii) `trace` which is by default `TRUE`.

5 Illustration: the 2007 French political blogosphere

This section illustrates the features of **missSBM** by conducting an analysis of a real-world network data. It is a sub-network of the French political blogosphere, extracted from a snapshot of over 1100 blogs collected during a period preceding the 2007 French presidential election, and manually classified by the "Observatoire Présidentielle project" (see Zanghi et al., 2008), . The network is composed of 196 blogs representing nodes in the network and 1432 edges indicating that at least one of the two blogs references the other. On top of **missSBM**, our analysis relies on **magrittr** and **igraph** for standard (network) data manipulation. We also fix the seed for reproducibility:

```
set.seed(1234)
library(missSBM)
library(igraph)
library(magrittr)
```

The `frenchblog2007` data set is shipped with **missSBM**[1]. We provide the data as an igraph object, which is the most convenient way of manipulating network data in R, in our opinion. We extract the adjacency matrix corresponding to the blog network and the vertex attribute party, providing the political party of each blog, also giving a natural classification of the nodes that could be used as a reference in our analyzes.

```
data("frenchblog2007", package = "missSBM")
class(frenchblog2007)

[1] "igraph"

adjacencyMatrix <- frenchblog2007 %>% as_adj(sparse = FALSE)
party <- vertex.attributes(frenchblog2007)$party
```

Once reordered row-wise and column-wise according to party, the adjacency matrix shows us that a part of the underlying structure is indeed supported by the political party (see Figure 5.3).

[1] earlier versions of this data set were available in packages **mixer** and **sand**

```
missSBM::sampledNetwork$new(adjacencyMatrix) %>%
  plot(clustering = as.factor(party), main = "")
```

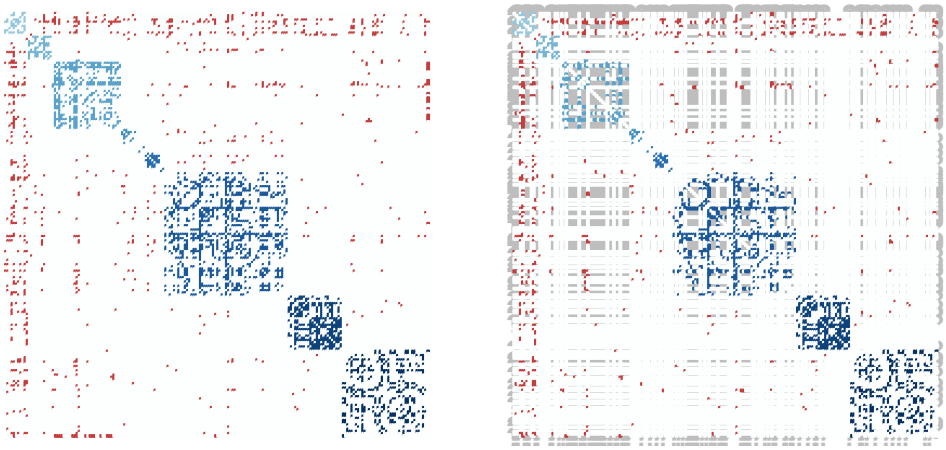


Figure 5.3 – Adjacency matrix of the French blogosphere reordered according to political party; left: fully observed; right: sampled according to a block-node sampling design. Blue color range corresponds to edges between nodes from the same cluster; red color range between node from different clusters; gray color correspond to missing dyads.

Standard SBM estimation. At this stage, the data set has no missing entries: every dyads and nodes are observed. The adjacency matrix $\mathbf{Y}$ of the fully-observed network is `adjacencyMatrix`. We can first perform a standard SBM estimation on the fully observed network. To do so we set first the algorithm parameters that will be common to all analyzes:

```
## Algorithm parameters : ##
vBlocks <- 1:15
control <- list(cores = 10, trace = 0, iterates = 2)
smoothing_type <- "both"
```

We then proceed to the estimation of the fully-observed network, including smoothing of the ICL.

```
sbm_full <-
  prepare_data(adjacencyMatrix) %>%
  estimate(vBlocks, "node", "hierarchical", control = control)
missSBM::smooth(sbm_full, smoothing_type, control)
```

Printing the object `missSBM_fit` results in a summary of the most important accessible fields:

```
sbm_full$bestModel

## missSBM-fit
## ====================================================================
## Structure for storing a SBM fitted under missing data condition
## ====================================================================
```

```
## * Useful fields (most are special object themselves)
##   $fittedSBM (the adjusted stochastic block model)
##   $fittedSampling (the estimated sampling process)
##   $sampledNetwork (the sampled network data)
##   $imputedNetwork (the adjacency matrix with imputed values)
##   $monitoring, $vICL, $vBound, $vExpec, $penalty
```

Sampling the network data. For illustrative purpose, we sample the blog network to mimic missing data and create a new adjacency matrix with missing entries. Since the data are interactions between blogs (who references who), it is natural to sample the network with a node-centered sampling design: the following code generates a realization of a 196×196 sampling matrix according to the *block-node sampling*, where blocks corresponds to the clustering estimated by SBM on the full data set and where sampling rate is either low (0.2) or high (0.8) in each block.

```
cl0 <- sbm_full$bestModel$fittedSBM$memberships
samplingParameters <- base::sample(
  x       = c(0.2, 0.8),
  size    = sbm_full$bestModel$fittedSBM$nBlocks,
  replace = TRUE)
sampledNet <- sample(adjacencyMatrix, "block-node", samplingParameters, cl0)
```

The global sampling rate obtained for this of the network data is

```
sampledNet$samplingRate

## [1] 0.7950811
```

The `sampleNet` object owns a `plot` method representing the network data in the missing data context (see right-panel in Figure 5.3):

```
plot(sampledNet, clustering = factor(party), main = "")
```

Estimation of a partially observed network. We are now in position of performing inference of SBM under missing data condition. We fit two types of model: first the SBM under the NMAR *block-node sampling* design, i.e. under the design that truly generated the missing entries; second the SBM under the *star sampling*, a.k.a MCAR *node sampling* in the package. The estimation is ran on both models with the same setting as for the fully observed data (i.e., by varying the number of blocks in $[\![1, 15]\!]$ and using the same tuning parameters for the initialization step and the optimization procedure). The ICL curve is smoothed with a forward-backward smoothing.

We first run the estimation for the block-node sampling and report ICL values before and after smoothing:

```
sbm_block <- estimate(sampledNet, vBlocks, "block-node", control = control)
ICL_block_nsm <- sbm_block$ICL
missSBM::smooth(sbm_block, smoothing_type, control)
```

Figure 5.4 illustrates the dramatic change in the model selection process due to the smoothing.

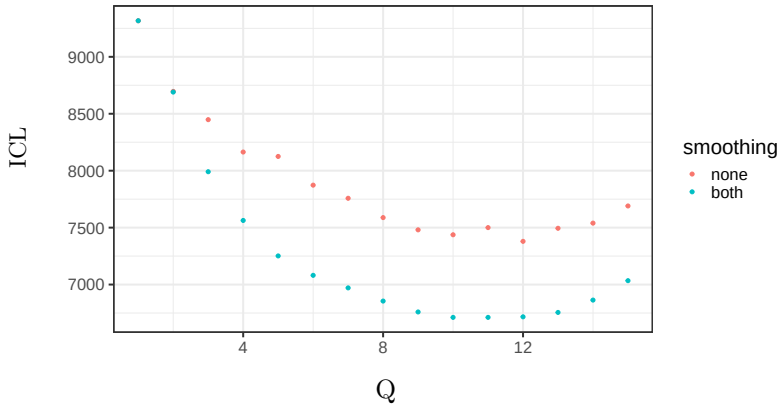


Figure 5.4 – ICL criterion for block-sampling design with/without smoothing of the ICL.

Sampling design comparison. Now, we consider the simple MCAR *node sampling* which basically performs inference only on the observed part of the network, neglecting the process that originates the missing values:

```
sbm_node <- estimate(sampledNet, vBlocks, "node", control = control)
missSBM::smooth(sbm_node, smoothing_type, control)
```

Figure 5.5 shows how ICL can be used to select which sampling design fit at best the data by comparing the smoothed ICL curves for "block-node" and "node" sampling designs. Note that the curve associated with the *block-node sampling* uniformly dominates the curve associated with the *star sampling*, showing that this sampling design is more adapted to the network data at play. We also represent the ICL curve of the SBM estimated on the fully observed network: although the values of the ICLs cannot be compared with the ones obtained for the partially observed network (indeed, data are not the same in this case), the number of block selected in the different cases remains comparable.

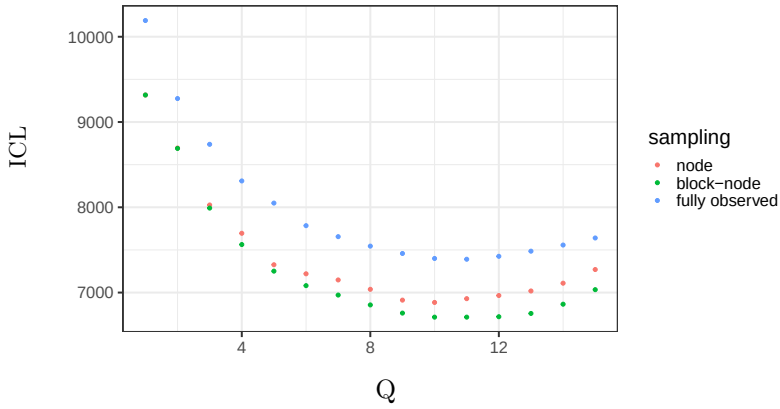


Figure 5.5 – ICL criterion curves for th block-node and node sampling designs.

The ICL criterion selects 11 blocks for an SBM adjusted on the fully observed network, while the SBM with missing entries accounted for by block-node sampling only selects 9 blocks, with a more flatter ICL curve. Indeed, due to the partial sampling, some blocks are less well represented than others; and it seems more likely to gather some blocks together considering the information available.

Regarding the clusterings obtained by the three variants (fully observed, missing entries with MCAR modeling and missing entries with NMAR block-node sampling), we compare them with the Adjusted Rand Index (ARI, Rand, 1971), computed with the **aricode** package (Chiquet and Rigaill, 2018). We use the classification of the SBM fitted on the fully observed network as a reference, since its clustering was used to sample the network with the block-node sampling design. We typically expect that an SBM relying on a better modeling of the missing values shall lead to a clustering closer to the reference. This is indeed the case when looking at the next piece of code, where it is shown that the ARI with the reference clustering is 50% higher for the SBM with block-sampling than for SBM with MCAR node sampling:

```
aricode::ARI(sbm_block$bestModel$fittedSBM$memberships, cl0)

[1] 0.6031785

aricode::ARI(sbm_node$bestModel$fittedSBM$memberships , cl0)

[1] 0.4412234
```

Extraction of the SBM with block-sampling design. The model that we finally retain it thus block-sampling with 9 blocks.

```
myModel <- sbm_block$bestModel
```

As seen from Figure 5.2, `myModel` is an object with class `missSBM_fit` with two important fields used for storing the results of the estimation of both the SBM (field `fittedSBM`) and the sampling design (`fittedSampling`). The important fields and methods are recalled to the user thanks to the `print` methods:

```
myModel$fittedSBM

Fit of a Stochastic Block Model
==================================================================
Model undirected with 196 nodes, 10 blocks and no covariate(s).
==================================================================
* Useful fields
  $nNodes, $nBlocks, $nCovariates, $nDyads
  $mixtureParam, $connectParam
 $covarParam, $covarArray
* Additional fields
  $blocks, $memberships, $adjMatrix, $connectProb, $vBound, $vICL, $penalty
* S3 methods
  plot, print
```

```
myModel$fittedSampling

block-node-model for network sampling
==================================================================
Structure for handling network sampling in missSBM.
==================================================================
* Useful fields
  $type, $parameters, $prob, $df
  $penalty, $vExpec
```

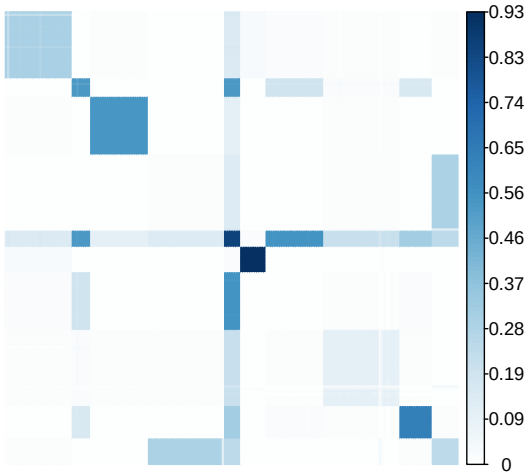


Figure 5.6 – Probability predicted by the SBM with block-node sampling

Representation and validation With `myModel`, we now have at hand a tool for analyzing the clustering of the French political blogosphere. The first output is the connectivity matrix of the network, which puts into light the community structure of the blogosphere. Indeed, it is revealed by a diagonal filled with high probabilities and off-diagonal with low probabilities. Thus, nodes (blogs) into blocks connects with high probability with other nodes of the same block and with low probability with nodes of other blocks. Such a network concentrates most of its edges between nodes of the same blocks. This can be seen by displaying the probability of connection predicted by the SBM at the whole network scale:

```
plot(myModel$fittedSBM, type = "connectivity")
```

For validation, we suggest to compare the clustering of the model with the node attribute corresponding to the political parties to which blogs belong to. First, we remark that the SBM fitted on missing entries carry the same amount of information regarding the political party than the SBM adjusted on the fully observed network:

```
aricode::ARI(party, myModel$fittedSBM$memberships)

[1] 0.4279665

aricode::ARI(party, sbm_full$bestModel$fittedSBM$memberships)

[1] 0.4244866
```

A more detailed comparison between blocks inferred by the SBM and political parties is reported in Figure 5.7 with an alluvial diagram.

Also remember that **missSBM** performs imputation of the missing dyads in the adjacency matrix. Thus, we can compare the imputed values with the value sof the dyad in the fully observed network to validate the performance of our approach. Using the R package **pROC** (Robin et al., 2011), we check the quality of the imputation. The following piece of code generates a singe ROC curve for the current sampling and imputation. Results in Figure 5.8 display ROC curves for 100 samplings of missing entries (always with block-sampling design). The sampling rate varies between ≈ 0.4 and ≈ 0.9. It shows the robustness and the good performance of the imputation method.

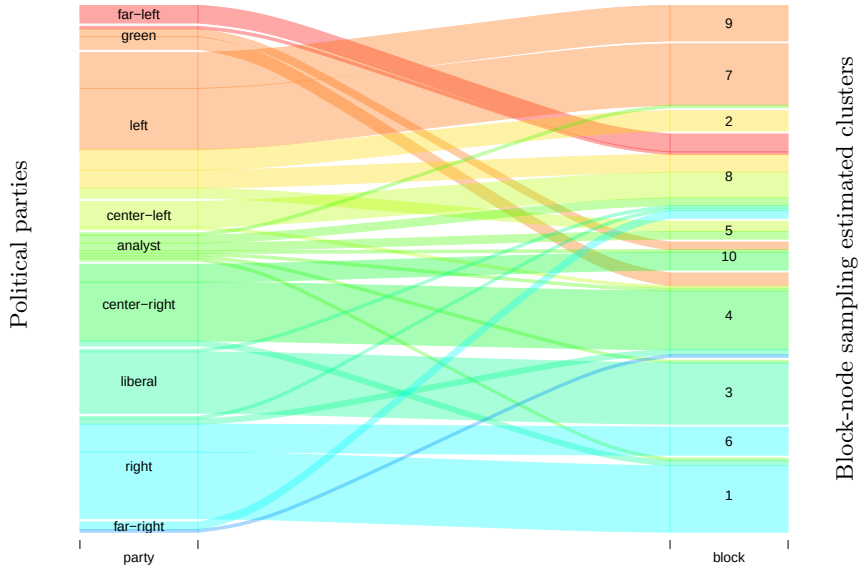


Figure 5.7 – Alluvial plot between block-node sampling clustering and political parties.

```
nu <- sbm_block$bestModel$imputedNetwork[sampledNet$NAs == TRUE]
Ym <- adjacencyMatrix[sampledNet$NAs == TRUE]

AUC <- pROC::roc(Ym,nu)
ggroc(AUC) + theme_bw(base_size = 20) + theme(axis.title = element_blank())
```

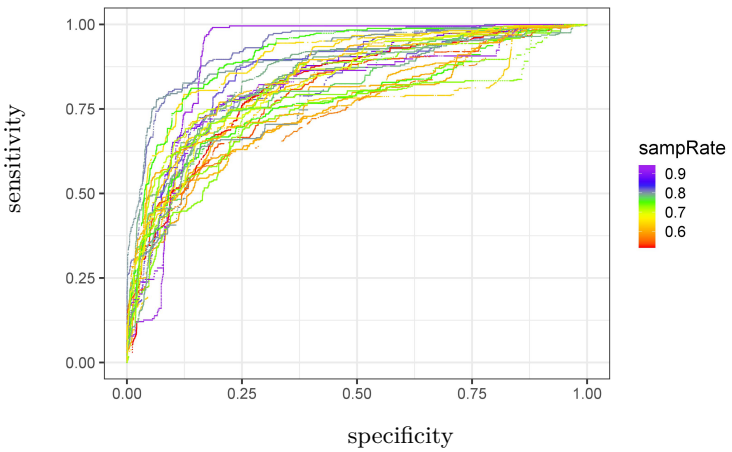


Figure 5.8 – ROC curves measuring quality of the imputation for varying sampling rate

Finally, the Area Under the Curve (AUC) when the sampling rate is varying between 0.43 and 0.9 is plotted in Figure 5.9. We can see that the more the sampling rate increases the more the AUC increases.

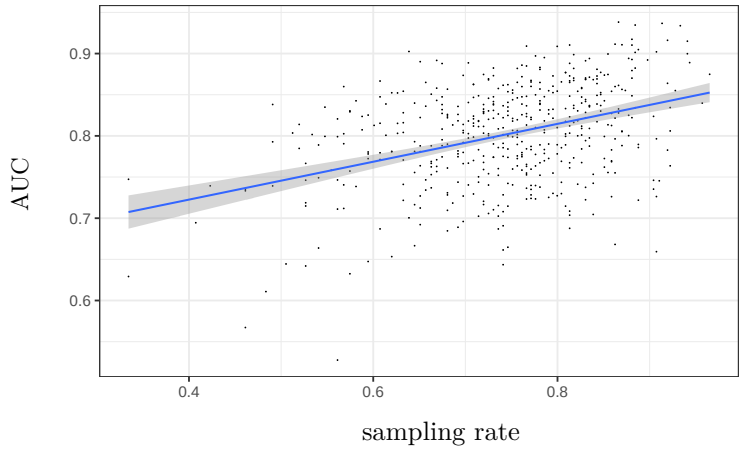


Figure 5.9 – Area Under the Curve (AUC) in function of the sampling rate.

Acknowledgments

The authors thank all members of MIRES group for fruitful discussions on network sampling designs. This work is supported by public grants overseen by the French National research Agency (ANR) as part of the "Investissement d'Avenir" program, through the "IDI 2017" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02, and second by the "EcoNet" project, ANR-18-CE02-0010.

Conclusion et perspectives

Contents

1 Conclusion

Dans le chapitre 2 nous avons étudié les différents types de données manquantes selon la théorie développée dans Rubin (1976) et nous l'avons adaptée au cas du SBM. Fort de la dichotomie MAR/NMAR nous avons donné des exemples de stratégies d'échantillonnage aussi bien centrées sur les nœuds que sur les dyades illustrant les cas de données manquantes MAR et NMAR pour le SBM. Un algorithme EM avec approximation variationnelle a été proposé pour s'adapter à tous les cas de données manquantes dont la loi du processus d'échantillonnage suit un modèle connu. Une étude de simulation poussée a permis de montrer la pertinence de la prise en compte de la nature des données manquantes dans l'inférence. De plus elle a permis de montrer qu'il était possible de choisir parmi plusieurs stratégies laquelle expliquait le mieux les données. Enfin l'identifiabilité du SBM sous trois des modèles de données manquantes définis dans ce chapitre a été établie.

Dans le chapitre 3 nous nous sommes intéressés à l'étude asymptotique des estimateurs du maximum de vraisemblance et de l'approximation variationnelle du SBM en présence de données manquantes. Le cadre considéré est celui d'un SBM dont la loi d'émission d'une dyade appartient à la famille exponentielle à un paramètre avec des données manquantes telles que chaque dyade est observée indépendamment et avec même probabilité. Ainsi, les estimateurs du maximum de vraisemblance et de l'approximation variationnelle sont consistants et asymptotiquement normaux. De plus, la variance asymptotique est explicite. Enfin, nous avons montré l'identifiabilité du SBM pour toute la famille exponentielle à un paramètre en présence de données manquantes.

Le chapitre 4 introduit différents modèles de SBM avec covariables et données manquantes. Les liens entre données manquantes et covariables sont définis ainsi que des exemples de stratégies d'échantillonnage. Nous avons montré que sous certaines conditions, un SBM sans covariable avec données manquantes NMAR est équivalent à un SBM avec covariables et données manquantes MAR. De plus, à travers des simulations nous montrons que sans la connaissance de covariables, les algorithmes NMAR développés dans le chapitre 2 sont dans certains cas une alternative intéressante pour l'estimation des paramètres du modèle et la prédiction de la classification par rapport au fait de considérer les données manquantes MAR pour un SBM sans covariables.

Finalement, le chapitre 5 décrit le package **missSBM**, développé dans la language R et actuellement sur le CRAN. Chaque fonction est précisément présentée ainsi que son utilisation. Plusieurs exemples reproductibles fondés sur des jeux de données disponibles directement dans le package sont proposés pour permettre une meilleure introduction et prise en main de **missSBM**.

2 Perspectives

Au cours de cette thèse, nous avons abordé la question des données manquantes dans le SBM. Cependant de nombreuses questions que nous n'avons pas eu le temps d'aborder restent en suspens et mériteraient de l'attention.

Une extension naturelle serait d'étudier d'autres lois pour les dyades du SBM comme par exemple la loi de Poisson pour des données de comptage et la loi normale pour des données continues. En effet, définir des stratégies d'échantillonnage et une méthode d'inférence pour les cas NMAR n'est pas évident et ne découle pas directement de notre travail. Il apparaît que l'approximation variationnelle que nous avons proposée (*i.e.* supposer les dyades manquantes indépendantes, de même lois avec leurs propres paramètres) pour le cas binaire avec des données manquantes NMAR ne produise pas des estimateurs explicites des paramètres variationnels et du SBM. Ceci résulte du fait que dans le cas binaire, la vraisemblance est linéaire en les dyades et s'intègre directement, comme pour la loi binomiale et la loi multinomiale. Cependant ce n'est le cas par exemple pour la loi de Poisson où la vraisemblance d'une dyade donnée par

$$\log p_{\boldsymbol{\theta}}(Y_{ij}|Z_{iq}=1, Z_{j\ell}=1) = -\pi_{q\ell} + Y_{ij}\log(\pi_{q\ell}) - \log(Y_{ij}!)$$

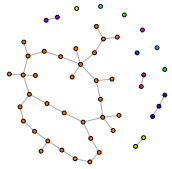
dont l'espérance quand Y_{ij} est manquant n'est pas directement calculable avec une approximation de loi de type "champs moyen". De plus, il n'est pas évident que des algorithmes d'optimisation soient efficaces compte tenu de la non convexité en général du problème. Il pourrait être intéressant aussi de se demander si des méthodes comme des algorithmes consistant à simuler conditionnellement au reste les Y_{ij} manquants permettraient de résoudre ce problème.

De nombreux problèmes théoriques persistent encore à propos de l'étude asymptotique des estimateurs du maximum de vraisemblance, ainsi que de celle des estimateurs du maximum de vraisemblance issus de l'approximation variationnelle pour des stratégies d'échantillonnage noeuds centrés MAR et toutes les stratégies NMAR. Ce problème est difficile et nécessite probablement pour chaque stratégie une preuve spécifique. De plus, il serait intéressant de montrer l'identifiabilité des modèles définis lorsqu'elle n'est pas connue. Il semble que les preuves classiques d'identifiabilité du SBM ne s'appliquent pas pour toutes les stratégies que nous avons définies. Notre intuition est que ces modèles sont identifiables, les simulations allant dans ce sens. Finalement, il nous semblerait important de mieux comprendre pourquoi dans les cas où des nœuds sont échantillonnés, faire l'hypothèse que des données manquantes sont MAR quand elles sont NMAR constitue une hypothèse robuste dans de nombreux cas que nous avons pu observer sur des simulations. Il serait important de pouvoir identifier ces cas de manière à pouvoir prévoir et prévenir quand l'hypothèse MAR est crédible. D'autres stratégies d'échantillonnage pour le cas de réseaux binaires qui paraîtraient naturelles pourraient aussi être définies. Le package **missSBM** a d'ailleurs été pensé pour pouvoir accueillir facilement de nouvelles stratégies.

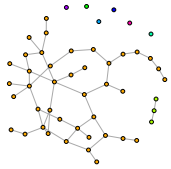
Dans toute cette thèse, la taille des réseaux étudiés est supposé connue. Cette hypothèse pose naturellement une question récurrente que nous n'avons pas explorée qui est de ne plus supposer la taille du réseau connue. C'est ce qu'il est proposé de faire dans l'article de Vincent and Thompson (2015) dans lequel les auteurs estiment conjointement la taille du réseau et les paramètres du SBM qu'ils explorent avec un échantillonnage *one-wave snowball sampling*. On pourrait se demander si leurs méthodes s'appliquent à d'autres cas comme ceux que nous étudions. Notre intuition est que cela ne devrait pas modifier la nature des données manquantes, ce qui devrait rendre simple l'inférence des cas MAR. Les cas NMAR nécessite d'estimer la taille du réseau, on pourrait imaginer par exemple que n soit aléatoire avec une loi puissance et l'estimer alternativement avec les paramètres du modèle.

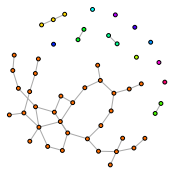
Il serait intéressant aussi de se poser la question de l'impact des "liens fallacieux" (*i.e.* observés avec du bruit) décrits dans Clauset et al. (2008) et Guimerà and Sales-Pardo (2009) sur la modélisation et la nature des données manquantes ainsi que leur prise en compte dans l'inférence.

Enfin, faisant le lien avec des problèmes venant de l'écologie, on pourrait se demander comment prendre en compte des stratégies NMAR pour le SBM qui échantillonnent uniquement des liens existant (*i.e.* des arêtes) et ceci en tenant compte éventuellement de covariables. Ceci consisterait *a priori* à appliquer une stratégie d'échantillonnage comme le



deux poids deux mesures en prenant en compte à la fois la valeur des covariables associées à une dyade et la valeur de la dyade dans la probabilité de l'échantillonner.





Bibliography

C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *J. Compl. Net.*, 3.2:221–248, 2014.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9(Sep):1981–2014, 2008.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

C. Ambroise and C. Matias. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *J. R. Stat. Soc. B-Met.*, 74(1):3–35, 2012.

C. Ambroise, G. Grasseau, M. Hoebeke, P. Latouche, V. Miele, and F. e. a. Picard. *MixeR: Random Graph Clustering*, 2015. R package version 1.8.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25(1):25, 2000.

P. Balachandran, E. D. Kolaczyk, and W. D. Viles. On the propagation of low-rate measurement error to subgraph counts in large networks. *Journal of Machine Learning Research*, 18(61):1–33, 2017.

A. Bar-Hen, P. Barbillon, and S. Donnet. Block model for multipartite networks. Applications in ecology and ethnobiology. working paper or preprint, July 2018.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.509.

P. Barbillon, S. Donnet, E. Lazega, and A. Bar-Hen. Stochastic block models for multiplex networks: an application to networks of researchers. *J. R. Stat. Soc. C-Appl.*, 2015.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974. ISSN 00359246.

S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. *Ann. Appl. Probab.*, 21(6):2146–2170, 12 2011. doi: 10.1214/10-AAP740.

P. Bickel, D. Choi, X. Chang, H. Zhang, et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Stat.*, 41(4):1922–1943, 2013.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725, July 2000. ISSN 0162-8828. doi: 10.1109/34.865189.

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 03 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx008.

B. Bollobas. *Random Graphs*. Cambridge University Press, 2001.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255.

C. Bouveyron, P. Latouche, and R. Zreik. The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, 2016. doi: 10.1007/s11222-016-9713-7.

V. Brault. *Estimation and model selection for the latent block model*. Theses, Université Paris Sud - Paris XI, Sept. 2014.

V. Brault, C. Keribin, and M. Mariadassou. Consistency and Asymptotic Normality of Latent Blocks Model Estimators. working paper or preprint, Apr. 2017.

C. Butts. network: A package for managing relational data in r. *Journal of Statistical Software, Articles*, 24(2):1–36, 2008a. ISSN 1548-7660. doi: 10.18637/jss.v024.i02.

C. Butts. Social network analysis with sna. *Journal of Statistical Software, Articles*, 24(6): 1–51, 2008b. ISSN 1548-7660. doi: 10.18637/jss.v024.i06.

G. Celeux and J. Diebolt. The sem algorithm : a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Rapport de recherche RR-1364, Inria, Institut National de Recherche en Informatique et en Automatique*, 1985.

G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics Quarterly*, vol. 2(no 1):p. 73–82, 1991.

A. Celisse, J.-J. Daudin, L. Pierre, et al. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.*, 6:1847–1899, 2012.

W. Chang. *R6: Classes with Reference Semantics*, 2017. R package version 2.2.2.

A. Channarond. *Clustering in a random graph : models with latent space*. Theses, Université Paris Sud - Paris XI, Dec. 2013.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *Ann. Statist.*, 41(5):2428–2461, 10 2013. doi: 10.1214/13-AOS1155.

J. Chiquet and G. Rigaill. *aricode: Efficient Computations of Standard Clustering Comparison Measures*, 2018. R package version 0.1.1.

D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with growing number of classes. *Biometrika*, 99 2:273–284, 2012a.

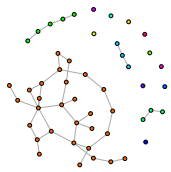
D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2):273–284, 04 2012b. ISSN 0006-3444. doi: 10.1093/biomet/asr053.

F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, Nov 2002. ISSN 0219-3094. doi: 10.1007/PL00012580.

A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008.

M. Corneli, C. Bouveyron, P. Latouche, and F. Rossi. The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, 2018. doi: 10.1007/s11222-018-9832-4.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org.



J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Stat. comp.*, 18(2):173–183, 2008.

M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B-Met.*, 39(1):1–38, 1977.

W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, Sep. 1973. ISSN 0018-8646. doi: 10.1147/rd.175.0420.

M. E J Newman, D. Watts, and S. H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 Suppl 1:2566–72, 03 2002. doi: 10.1073/pnas.012582999.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08.

P. Erdős and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

O. Frank and F. Harary. Cluster inference by using transitivity indices in empirical graphs. *J. Am. Stat. Soc.*, 77(380):835–840, 1982.

A. Franzin, F. Sambo, and B. di Camillo. bnstruct: an r package for bayesian network structure learning in the presence of missing data. *Bioinformatics*, 33(8):1250–1252, 2017. doi: 10.1093/bioinformatics/btw807.

A. Ghasemian, H. Hosseinmardi, and A. Clauset. Evaluating overfit and underfit in models of network community structure. *CoRR*, abs/1802.10582, 2018.

E. N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30(4):1141–1144, 12 1959. doi: 10.1214/aoms/1177706098.

W. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995. ISBN 9780412055515.

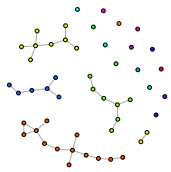
C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36 (2):463 – 473, 2003. ISSN 0031-3203. Biometrics.

R. Guimerà and M. Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0908366106.

M. Handcock, D. Hunter, C. Butts, S. Goodreau, and M. Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software, Articles*, 24(1):1–11, 2008. ISSN 1548-7660. doi: 10.18637/jss.v024.i01.



M. S. Handcock and K. J. Gile. Modeling social networks from sampled data. *The Annals of Applied Statistics*, 4(1):5–25, 2010.

M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170 (2):301–354, 2007. doi: 10.1111/j.1467-985X.2007.00471.x.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

S. Hojsgaard. Cran task view: graphical models in R, 2019. URL https://cran.r-project.org/web/views/gR.html.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

J. Hu, H. Qin, T. Yan, and Y. Zhao. On consistency of model selection for stochastic block models. https://arxiv.org/abs/1611.01238, 2017.

D. Hunter and M. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 9 2006a. ISSN 1061-8600. doi: 10.1198/106186006X133069.

D. Hunter, M. Handcock, C. Butts, S. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software, Articles*, 24(3):1–29, 2008. ISSN 1548-7660. doi: 10.18637/jss.v024.i03.

D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006b.

T. Jaakkola. *Advanced Mean Field Methods: Theory and Practice*, chapter : Tutorial on variational approximation methods. MIT Press, Cambridge, 2000.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. Springer, 1998.

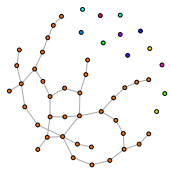
B. Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, Jan 2011.

C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015. doi: 10.1007/s11222-014-9472-2.

E. D. Kolaczyk. *Statistical analysis of network data, methods and models*. Springer, 2009.

P. Krivitsky and M. Handcock. Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software, Articles*, 24(5):1–23, 2008. ISSN 1548-7660. doi: 10.18637/jss.v024.i05.

V. Labeyrie, M. Deu, A. Barnaud, C. Calatayud, M. Buiron, P. Wambugu, S. Manel, J.-C. Glaszmann, and C. Leclerc. Influence of ethnolinguistic diversity on the sorghum genetic patterns in subsistence farming systems in eastern kenya. *PLoS One*, 9(3):e92178, 2014.



V. Labeyrie, M. Thomas, Z. K. Muthamia, and C. Leclerc. Seed exchange networks, ethnicity, and sorghum diversity. *P. Natl. Acad. Sci.*, 113(1):98–103, 2016.

P. Latouche and S. S. Robin. Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Statistics and Computing*, 26(6):1173 – 1185, 2016. doi: 10.1007/s11222-015-9607-0.

P. Latouche, É. Birmelé, and C. Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *Ann. Appl. Stat.*, pages 309–336, 2011.

P. Latouche, É. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Stat. Modelling*, 12(1):93–115, 2012.

P. Latouche, S. Robin, and S. Ouadah. Goodness of fit of logistic regression models for random graphs. *Journal of Computational and Graphical Statistics*, 27(1):98–109, 2018. doi: 10.1080/10618600.2017.1349663.

L. S. Lauritzen. *Graphical models.* Clarendon Press, 1996.

J.-B. Leger. *blockmodels: Latent and Stochastic Block Model Estimation by a 'V-EM' Algorithm*, 2015. R package version 1.1.1.

J.-B. Leger. Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates. Technical report, 2016. URL https://arxiv.org/abs/1602.07587.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data.* John Wiley & Sons, 2014.

L. Lovász. *Large Networks and Graph Limits.* American Mathematical Society colloquium publications. American Mathematical Society, 2012. ISBN 9780821890851.

L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006. ISSN 0095-8956.

M. Ludkin, I. Eckley, and P. Neal. Dynamic stochastic block models: parameter estimation and detection of changes in community structure. *Statistics and Computing*, 28:1201–1213, 2018.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

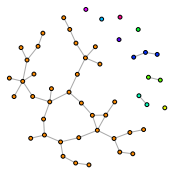
M. Mariadassou and C. Matias. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21(1):537–573, 2015.

M. Mariadassou and T. Tabouy. Consistency and asymptotic normality of stochastic block models estimators from sampled data. 2019. URL https://arxiv.org/abs/1903.12488.

M. Mariadassou, S. Robin, and C. Vacher. Uncovering latent structure in valued graphs: A variational approach. *Ann. Appl. Stat.*, 4(2):715–742, 06 2010.

C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. B-Met.*, 2016.

C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017. doi: 10.1111/rssb.12200.



C. Matias and S. Robin. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM Proc. Sur.*, 47:55–74, 2014.

C. Matias, T. Rebafka, and F. Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680, Sept. 2018. doi: 10.1093/biomet/asy016.

V. Miele, F. Picard, and S. Dray. Spatially constrained clustering of ecological networks. *Methods in Ecology and Evolution*, 5(8):771–779, 2014. doi: 10.1111/2041-210X.12208.

G. Molenberghs, C. Beunckens, C. Sotto, and G. M. Kenward. Every missing not at random model has got a missing at random counterpart with equal fit. *J. R. Stat. Soc. B-Met.*, 2008.

M. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. doi: 10.1137/S003614450342480.

M. Noroozi, R. Rimal, and M. Pensky. Estimation and clustering in popularity adjusted stochastic block model. working paper or preprint, 2019. URL https://arxiv.org/abs/1902.00431.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Soc.*, 96(455):1077–1087, September 2001.

J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence*, AAAI'82, pages 133–136. AAAI Press, 1982.

J. Pearl. Bayesian networks. 2011.

M. Penrose. *Random Geometric Graphs*. Oxford studies in probability 5. Oxford University Press, 2003.

C. E. Priebe, D. L. Sussman, M. Tang, and J. T. Vogelstein. Statistical inference on errorfully observed graphs. *Journal of Computational and Graphical Statistics*, 24(4):930–953, 2015.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Soc.*, 66(336):846–850, 1971.

X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Muller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

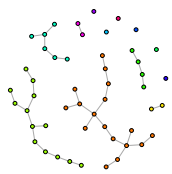
G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191, 2007. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2006.08.002. Special Section: Advances in Exponential Random Graph (p*) Models.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *Ann. Stat.*, 2010.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978. doi: 10.1214/aos/1176344136.

S. Sengupta and Y. Chen. A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80 (2):365–386, 2018. doi: 10.1111/rssb.12245.



J. Shanthikumar and U. Sumita. A central limit theorem for random sums of random variables. *Operations Research Letters*, 3(3):153 – 155, 1984. doi: http://dx.doi.org/10.1016/0167-6377(84)90008-7.

T. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 2, 2002. ISSN 1529-1227.

T. A. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37: 131–153, 2011.

T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. class.*, 14(1):75–100, 1997.

H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci. Cl. III. 4*, pages 801–804, 1956.

T. M. Sweet. Incorporating covariates into stochastic blockmodels. *Journal of Educational and Behavioral Statistics*, 40(6):635–664, 2015. doi: 10.3102/1076998615606110.

D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic. Acids Res.*, 43, 2015.

T. Tabouy, P. Barbillon, and J. Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 0(ja):1–20, 2019a. doi: 10.1080/01621459.2018.1562934.

T. Tabouy, P. Barbillon, and J. Chiquet. misssbm: An r package for handling missing values in the stochastic block model. 2019b. URL https://arxiv.org/abs/1906.12201.

T. Tabouy, P. Barbillon, and J. Chiquet. *An R package for adjusting Stochastic Block Models from networks data sampled under various missing data conditions*, 2019c. R package version 0.2.0.

C. Tallberg. A bayesian approach to modeling stochastic blockstructures with covariates. *The Journal of Mathematical Sociology*, 29(1):1–23, 2004. doi: 10.1080/00222500590889703.

S. K. Thompson and O. Frank. Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26(1):87–98, 2000.

S. K. Thompson and G. Seber. *Adaptive Sampling*. New-York : Wiley, 1996.

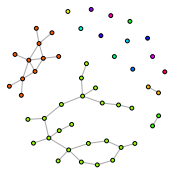
T. Valles-Catala, T. P. Peixoto, M. Sales-Pardo, and R. Guimerà. Consistencies and inconsistencies between model selection and link prediction in networks. *Physical review. E*, 97 6-1:062316, 2018.

R. K. Vinayak, S. Oymak, and B. Hassibi. Graph clustering with missing data: Convex algorithms and analysis. *Adv. Neu. In.*, 2014.

K. Vincent and S. Thompson. Estimating the size and distribution of networked populations with snowball sampling. Technical report, 2015. URL http://arxiv.org/abs/1402.4372v2.

M. J. Wainwright. Basic tail and concentration bounds. https://www.stat.berkeley.edu/ mjwain/stat210b/Chap2 2015.

Y. X. R. Wang and P. J. Bickel. Likelihood-based model selection for stochastic block models. *Ann. Statist.*, 45(2):500–528, 04 2017. doi: 10.1214/16-AOS1457.



S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. doi: 10.1017/CBO9780511815478.

G. Wei and M. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85 (411):699–704, 1 1990. ISSN 0162-1459. doi: 10.1080/01621459.1990.10474930.

W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977. doi: 10.1086/jar.33.4.3629752.

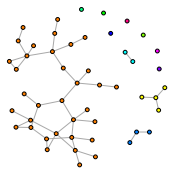
H. Zanghi, C. Ambroise, and V. Miele. Fast online graph clustering via erdős-rényi mixture. *Pattern Recognition*, 41(12):3592 – 3599, 2008. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2008.06.019.

H. Zanghi, S. Volant, and C. Ambroise. Clustering based on random graph model embedding vertex features. *Pattern Recognition Letters*, 31(9):830 – 836, 2010. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2010.01.026.

Y. Zhang, E. Levina, and J. Zhu. Community detection in networks with node features. *Electron. J. Statist.*, 10(2):3153–3178, 2016. doi: 10.1214/16-EJS1206.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in r. *J. Mach. Learn. Res.*, 13(1):1059–1062, Apr. 2012a. ISSN 1532-4435.

Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.*, 40(4):2266–2292, 08 2012b. doi: 10.1214/12-AOS1036.



université PARIS-SACLAY
École doctorale de mathématiques Hadamard (EDMH)

Titre : Impact de l'échantillonnage sur l'inférence de structures dans les réseaux : application aux réseaux d'échanges de graines et à l'écologie

Mots Clefs : Réseaux, modèle à blocs stochastiques, données manquantes

Résumé : Dans cette thèse nous nous intéressons à l'étude du modèle à bloc stochastique (SBM) en présence de données manquantes. Nous proposons une classification des données manquantes en deux catégories Missing At Random et Not Missing At Random pour les modèles à variables latentes suivant le modèle décrit par D. Rubin. De plus, nous nous sommes attachés à décrire plusieurs stratégies d'échantillonnages de réseau et leurs lois. L'inférence des modèles de SBM avec données manquantes est faite par l'intermédiaire d'une adaptation de l'algorithme EM : l'EM avec approximation variationnelle. L'identifiabilité de plusieurs des SBM avec données manquantes a pu être démontrée ainsi que la consistance et la normalité asymptotique des estimateurs du maximum de vraisemblance et des estimateurs avec approximation variationnelle dans le cas où chaque dyade (paire de nœuds) est échantillonnée indépendamment et avec même probabilité. Nous nous sommes aussi intéressés aux modèles de SBM avec covariables, à leurs inférence en présence de données manquantes et comment procéder quand les covariables ne sont pas disponibles pour conduire l'inférence. Finalement, toutes nos méthodes ont été implémentées dans un package R disponible sur le CRAN. Une documentation complète sur l'utilisation de ce package a été écrite en complément.

Title : Impact of sampling on structure inference in networks: application to seed exchange networks and to ecology

Keywords : Networks, Stochastic Block Model, missing data

Abstract : In this thesis we are interested in studying the stochastic block model (SBM) in the presence of missing data. We propose a classification of missing data into two categories Missing At Random and Not Missing At Random for latent variable models according to the model described by D. Rubin. In addition, we have focused on describing several network sampling strategies and their distributions. The inference of SBMs with missing data is made through an adaptation of the EM algorithm: the EM with variational approximation. The identifiability of several of the SBM models with missing data has been demonstrated as well as the consistency and asymptotic normality of the maximum likelihood estimators and variational approximation estimators in the case where each dyad (pair of nodes) is sampled independently and with equal probability. We also looked at SBMs with covariates, their inference in the presence of missing data and how to proceed when covariates are not available to conduct the inference. Finally, all our methods were implemented in an R package available on the CRAN. A complete documentation on the use of this package has been written in addition.