Recherche de structure dans un graphe aléatoire : modèles à espace latent

Antoine Channarond

Université Paris-Sud

Faculté des Sciences d'Orsay

École doctorale de Mathématiques
de la région Paris-Sud

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI

Spécialité : Mathématiques

par

Antoine Channarond

Sujet :

Recherche de structure dans un graphe aléatoire : modèles à espace latent

Soutenue le 10 décembre 2013 devant la Commission d'examen composée de :

M.	Gérard Biau	Université Pierre et Marie Curie	(Examinateur)
M.	Étienne Birmelé	Université Paris Descartes	(Rapporteur)
M.	Jean-Jacques Daudin	AgroParisTech	(Directeur de thèse)
Mme	Anne Estrade	Université Paris Descartes	(Examinateur)
M.	Christophe Giraud	Université Paris Sud	(Examinateur)
M.	Stéphane Robin	AgroParisTech/INRA	(Co-directeur de thèse)

Après avis des rapporteurs : M. Étienne Birmelé Université Paris Descartes
M. Eric Kolaczyk Boston University

Le commencement de toutes les sciences, c'est l'étonnement de ce que les choses sont ce qu'elles sont. Et vice et versa.

librement inspiré d'Aristote.

Remerciements

Nombreuses sont les personnes qui selon moi ont contribué à cette thèse par leur soutien, professionnel ou amical. Malgré tous mes efforts, je ne saurais en faire la liste exhaustive, et je commence donc par m'excuser par avance auprès de tous ceux que j'aurais omis de citer dans ce texte déjà long !

Mes premiers remerciements sont destinés à mes directeurs de thèse, Jean-Jacques et Stéphane. Cette simple phrase me semble bien maigre en regard de ce que vous m'avez apporté tout au long de cette thèse. Prenez-la comme une poupée russe, dans laquelle se cacherait aussi mon admiration et ma reconnaissance envers vous. Entre le vouvoiement en stage de master et les corrections finales de mon manuscrit, il s'est passé trois années pendant lesquelles vous m'avez fait découvrir le monde de la recherche de la plus pédagogique des façons. Votre disponibilité et votre patience est ce dont rêve tout doctorant, et bien que l'affirmant, je ne me rends probablement pas compte à quel point cela est vrai. Je vous remercie pour avoir partagé avec moi votre immense connaissance des statistiques et votre enthousiasme pour ces intrigants objets que sont les graphes. Merci encore, pour avoir relu mes élucubrations, qui se sont transformées en articles à la lumière de vos quelques conseils magiques. Stéphane, toi que je vois un peu plus souvent en tête à tête depuis de départ à la retraite de Jean-Jacques, je te témoigne à nouveau mon admiration, pour ces moments où tu comprenais le problème avec lequel j'arrivais dans ton bureau, et ce avant même que j'aie pu t'en expliquer le quart de la moitié. Je soupçonne cependant que tu aies posé un mouchard dans mon bureau.

Je remercie mes rapporteurs, Étienne Birmelé et Eric Kolaczyk, d'avoir relu ce manuscrit. Merci pour l'appui que vous m'avez donné, ainsi que pour m'avoir fait part de vos observations sur ces travaux. Je remercie aussi Gérard Biau, Étienne Birmelé, Anne Estrade et Christophe Giraud, d'avoir accepté de faire partie de mon jury de thèse. Gérard, vous avez été un des auteurs, avec Benoît Cadre et Bruno Pelletier, de l'article qui a été le point de départ d'une grande partie de ma thèse. C'est donc un honneur pour moi que vous soyiez présent aujourd'hui. Anne, merci de t'être intéressée et d'avoir suivi mon travail avec ton regard de probabiliste.

Je remercie tous les chercheurs que j'ai pu rencontrer sur le chemin qui a mené à cette thèse. Je pense d'abord à Élisabeth Gassiat grâce à qui j'ai rencontré Stéphane, puis Jean-Jacques et enfin les graphes, alors que je tâtonnais dans la pénombre de la recherche d'un sujet de thèse. Je pense à toutes les personnes avec qui j'ai pu discuter de mon sujet et qui m'ont apporté leur regard dessus, dont Catherine Aaron, Gilles Celeux, Cécile Durot, Ulrike von Luxburg, Pierre Pudlo, Karl Rohe et Nicolas Verzelen. Gilles, merci également de m'avoir aussi initié aux modèles de mélange ainsi qu'à la formulation variationnelle de l'EM, et ce, dès l'examen de votre cours. J'ai naïvement cru à l'époque qu'il s'agissait peut-être d'un hasard. Cécile, merci d'avoir rendu captivant ton cours de statistiques asymptotiques de master par ton dynamisme. Pierre, merci pour cette discussion à Fréjus qui a ouvert la porte à la seconde partie de ce manuscrit de thèse. Merci à tous les deux également pour avoir accepté de faire partie de mon comité de

thèse. Merci aux organisateurs des séminaires que je fréquentais très régulièrement et qui m'ont beaucoup appris, notamment Tristan à l'Agro, Sophie et Mahendra à SSB.

Je remercie tous les enseignants ou enseignants-chercheurs qui ont suscité mon intérêt pour les mathématiques, puis plus tard pour les probabilités et statistiques. Je pense en premier à M.Fargeas et M.Boutemy, mes professeurs respectivement de Terminale et de 5/2, qui avaient et ont toujours mon immense admiration. M.Fargeas, votre goût communicatif des mathématiques, aussi bien que votre personnalité, y sont pour beaucoup dans ma conviction de m'orienter vers les mathématiques après le bac. M. Boutémy, merci pour votre dynamisme et votre vivacité. Vous nous avez transportés rapidement et simplement, ma classe et moi, vers les raisonnements les plus élaborés : votre cours est encore une référence pour moi. Merci à Gérard Sussel, professeur mais surtout ami de ma famille, et qui à l'occasion de quelques randonnées et casse-croûtes copieux, m'a communiqué sa passion pour les mathématiques. Merci à M.Guédès de nous avoir fait entrer dans la taupinière, ma classe de MPSI3 et moi, et à mademoiselle Chanet, en particulier pour nous avoir appris la typographie grecque correcte. Je remercie l'ENS de Cachan et tous ses professeurs, en particulier Frédéric Pascal, qui s'est occupé de notre promo avec bienveillance tout au long de notre parcours. Merci à Erwan Le Pennec, que j'ai eu comme chargé de TD de statistiques à Jussieu, à Wendelin Werner pour son excellent cours de percolation.

Je remercie toute l'unité 518 de l'Agro, qui a créé une atmosphère de travail agréable et épanouissante qui à mon avis a été essentielle à l'élaboration de cette thèse. Merci à tous les permanents, Artemio, Céline, Christophe, Denis, Émilie, Éric, Gabriel, Julie, Julien, Liliane, Marie, Marie-Laure, Maud, Michel, Pierre Baaarbillon, Sarah et Tristan. Sarah, ton aide dans ma lourde charge d'enseignements m'a été très précieuse. Merci Marie pour ta bonne humeur généreuse et ton caractère bien trempé que j'admire, Pierre B, pour la facette noire coûteuse de ton humour et également pour toutes tes réponses à mes questions et suspicions sur le modèle linéaire, et Tristan pour ta coolitude en toute circonstance (notamment en cours de master !). Merci vous trois de m'avoir supporté comme squatteur régulier de votre bureau. J'adresse mes remerciements à tous les doctorants et post-doctorants, anciens : Caroline pour ta jovialité, Florence, Guillem, Natalie, Ophélie et Stevenn, et actuels : Aurélien, Fred et Jean- Baptiste (les compères du BDDDB). Je m'arrête brusquement dans cette énumération, pour remercier Fred de m'avoir soulagé de mes enseignements aux moments cruciaux, et de son humour si fromager, et pour remercier le BDDDB d'avoir initié les visites historiques de la rue Mouffetard du vendredi soir. Merci à Anna, Guillaume, Loïc (un de mes compagnons de route dans l'aventure grignonnaise), Pierre Colin, et Trung. Merci à Xiao, et à Eleanna pour nos conversations *in english most of the time*, mais peut-être bientôt tout en français ? Merci à Francine et Sophie pour votre efficacité et votre disponibilité, mais aussi à Benjamin, Carole, Damien, Dominique, Hamid, Laurent et Marc, qui créez une atmosphère chaleureuse au labo. Je ne crois avoir oublié personne. Ah si, et pas des moindres : je rends hommage aux membres du respectable et respecté BDDDH, Alice, Aurore, Jean-Benoist, et nouvellement Marie et Souhil. Merci JBen pour tous tes tips informatiques et en tous genres, et surtout en tous genres !

Je remercie le département de Mathématiques d'Orsay, où j'ai été initié à l'enseignement via le monitorat. Je remercie Pierre Pansu qui a chapeauté la répartition des enseignements, Élisabeth Gassiat et Cécile Durot avec qui j'ai collaboré pour mes tous premiers enseignements à l'université, puis Nathalie Castelle, Yan Pautrat et Frédéric Haglund avec qui j'ai continué par la suite à travailler à l'épineuse tâche de la transmission de savoirs mathématiques à des non-spécialistes. Je remercie David Harari et Nessim Siboni à l'école doctorale pour leur aide appréciable pendant la thèse. Un grand merci aussi à Valérie Lavigne, pour sa disponibilité et son aide rassurante dans toutes les démarches anxiogènes de la préparation de la soutenance. Merci à tous les doctorants d'Orsay, notamment Patrick, Raphaël, Rémy, Vincent, qui ont eu pitié de moi et m'ont intégré à leur sympathique bureau du 440, qui avait bien plus fière allure que mon antre d'origine des abysses du 430. Merci également à Clément et Laure avec qui j'ai enseigné en L2.

Je n'ai pas encore fini ces remerciements, mais je me permets un bref interlude, avant d'entamer des attentions plus personnelles. Petit, j'avais la tête pleine de rêves tous plus vraisemblables les uns que les autres, comme devenir réalisateur de films français qui rencontreraient un succès international, devenir champion de badminton, pilote de formule 1, cascadeur ou catcheur. J'ai donc sûrement dû rêver aussi un jour d'être écrivain. Je me permets donc de savourer cet instant, pendant lequel ma prose non strictement mathématique sera peut-être lue par plus d'une personne.

Je voudrais à présent remercier l'ensemble de mes amis, sans la présence et le soutien desquels cette thèse n'aurait simplement pas été.

Merci à tout mon clan de Cachan : Émilien, Kron et Popoff, mes compagnons de toutes les situations les plus abracadabrantes. Je ne puis me retenir de témoigner ma profonde amitié pour Popoff, ce personnage improbable, cet ovni dans le paysage des mathématiciens, dont la rencontre ne laisse jamais indifférent. D'ailleurs mon humérus ne l'est pas resté. Merci à El Colonel et El Communisto d'être ce que vous êtes, de vrais bonhommes ! Merci à Arthur Mozzarelo de m'avoir rappelé à chacun de nos affrontements que je ne serai jamais champion de badminton. Par ailleurs nos discussions sur les probabilités et la physique porteront je l'espère encore d'autres fruits. Merci au passage à tes collègues du LPTMS, notamment Paul et Pierre-Élie, qui ont toujours accueilli à bras ouverts le vagabond d'Orsay que j'étais. Merci à Cécilou pour avoir partagé avec moi, le statut d'auditeur libre, puis le concours 3A, des oignons et de la crème au deuxième étage du bâtiment J, et surtout, tous les potins les plus incroyables. Merci à toi et Arthur de m'avoir fait confiance pour assumer la difficile mission de coach de Jeanne. Merci à Gaëlle pour ta gentillesse et ta générosité : je me devais de sauver une telle perle de l'évanouissement ! Et surtout, merci pour ta patience avec l'insupportable touriste photophile que je suis. Vous tous, ainsi qu'Agathe, Baptiste, Irène, JSK, Julie, Quentin, Romain, Sandra ni couette, et bien sûr Jimmy, mais aussi Bobby, Cindy, Stephen et... Stephen, j'ai adoré nos nombreuses virées cachanaises, corses, rétaises, avignonnaises, alpines, qui resteront dans les annales.

Merci aux MPSI3 : Aurore, Clémentine, Émilie, Fannie, Guillaume, Ito, Jean Q, Léo, Mathieu, Matthieu, Romuald, Samuel, Sophie, Thomas, Vincent, Yannick. Avec vous, la

prépa est devenue (presque) une véritable partie de plaisir, et une aventure aussi bien étudiante qu'humaine. Tout au long de nos cours en E 308, nos khôlles, nos soirées taupe, puis surtout nos randonnées et nos voyages dont je garde une foule de souvenirs impérissables, j'ai découvert ce qu'était un groupe d'amis solidaires. Merci à Ito, Jean Q, Matchô et Audha, avec qui j'ai sillonné le monde entier : Europe, Asie, Amérique ! Enfin, j'ai une pensée particulière pour Clémentine : merci pour le bout de chemin que nous avons parcouru ensemble.

Merci à toute la tribu rochelaise. Les « man » : Ben Brouss, Geoffrey, Julien et Médéric. Merci à Julien de m'avoir appris à rester debout même quand un hurluberlu essaye absolument de vous montrer sa nouvelle prise de jujitsu. Merci à Cyrielle de l'avoir un peu calmé. Merci Médéric pour ton humour et tes imitations... inimitables. Je pense à tous mes autres camarades de lycée (et de collège, voire primaire !) que j'ai encore souvent l'occasion de croiser : Adrien le conquérant, Alexis, Arthur, BeFa, Corentin, Peter, Thierry, et à tous les autres rochelais : Aurélie, Julien Gwenn, Laetitia, Lilian, Romain Rocheteau, Romain Jacques, Sarah, Stéphanie. Pour vous tous, j'ai toujours un immense plaisir à revenir dans ma ville natale. Pour ceux que je vois de temps en temps à Paris, je regrette de ne pas vous voir plus souvent. Je remercie enfin Anna et Clara, les deux filles les plus folles que la Terre ait jamais connues : je suis persuadé que malgré tous vos voyages autour du monde, vous n'avez jamais croisé vos alter egos.

Merci à toute la Drôme team, qui s'est formée de manière évidente en à peine quelques jours cet été, et qui ne se lâche plus depuis : Alice, Coraline, Élise, Émilie Petit, Émilien, Kron, Lucie, Marie, Mikhaël, Tigrane, Valérie. Alice, merci pour ton humour ; j'espère que tu as trouvé la référence de dialogue de film cachée dans ces remerciements. Lucie, merci pour ton rire immortel tel un phénix qui renaît éternellement de ses cendres. Merci à toi et Émilien de vos nombreuses invitations dans votre QG du 18^{e}, pour une soirée, une après-midi de luge, ou pour vaigrer avec Yann et Charlotte. Merci Valérie de libérer une place de Vélib chaque jour près de l'Agro, dans le grand chassé-croisé des thésards.

Je dois beaucoup à la personne avec qui je partage ma vie en quelque sorte depuis trois années maintenant : ma colocataire. Elsa, tu m'as soutenu pendant une période difficile, et je t'en suis infiniment redevable. Ma dette envers toi est presque aussi grande que notre désamour commun pour la vaisselle, et c'est peu dire... Merci aussi à Valentin, d'avoir fait la vaisselle. Ta gentillesse n'a d'égal que ta rhétorique. Heureusement d'ailleurs, car une telle rhétorique pourrait être dangeureuse entre de mauvaises mains !

Je remercie enfin toutes les personnes avec qui j'ai partagé des moments inoubliables : Émilie D, Juliette C et Agathe B, Justine à qui je dois ma connaissance de la Basse Normandie, Tosaki pour ta science du mahjong, quelques Cachanais : Isabelle, Jonas, Pierre-Antoine le bleauzer, Kévin MM, Léo G, Fabrice V, Mélanie C, Anaëlle P, Benjamin M, et Kalliopi. Je remercie Bertrand, Erwan et Jean-Baptiste et tout le club de Badminton de Montrouge où j'ai trouvé une foule de partenaires avec qui partager mon attrait pour ce sport pendant ces trois dernières années. J'ai ainsi pu parfois passer mes nerfs sur un volant plutôt que dans des destructions matérielles regrettables. Pour mes nerfs, merci à Matlab, au RER B, ou au type à capuche rouge qui s'enfuit devant moi sur le dernier Vélib en état de marche tous les matins. Je remercie les cartésiens avec qui

j'ai passé une formidable année de 5/2 : Amaury, Benjamin, Émilien, Morgan, Nicolas, Pierre, Xavier. Merci à tout le reste de la famille Sussel, Marie-Odile, Anne et Olivier.

Émilie, je devine ton effroi en te voyant à peine citée précédemment, mais tu connais déjà mon esprit moqueur... Et je crois qu'en fait tu avais déjà deviné la fine supercherie. Je peux maintenant mieux te remercier de ton soutien sans faille pendant mon dernier mois de rédaction, qui a duré un mois et demi... Merci pour ta patience, ta générosité, et pour ton grand rire clair.

Ce final est réservé à ma famille, qui m'a toujours soutenu dans mes choix, même les plus fantaisistes, et qui m'a encouragé dans mon attrait pour les sciences depuis tout petit, même si ce n'était pas leur domaine de prédilection. Merci à ma petite soeur Cécile, pour le lien indessoudable qui existe entre nous. Merci à mon père et à ma mère de s'être toujours souciés de notre avenir à tous les deux. Pardon si mon côté distrait vous a plus d'une fois donné des sueurs froides... Je remercie aussi ma tante et marraine, de s'être toujours intéressée à mes traits qui relient des points ! Merci pour ta bonne humeur et ton coeur sur la main. Merci à mon cousin Philippe, qui a toujours été un modèle pour moi depuis nos vacances rétaises d'enfance. Enfin, c'est non sans émotion que mes pensées vont à mes grands-parents, à ma mamie à qui je dois beaucoup dans mon parcours, et qui j'en suis persuadé, aurait aimé être là aujourd'hui. À vous et à toute ma famille, cette thèse vous est dédiée en témoignage de ma reconnaissance.

Table des matières

Première partie

Introduction

Chapitre 1

Préambule

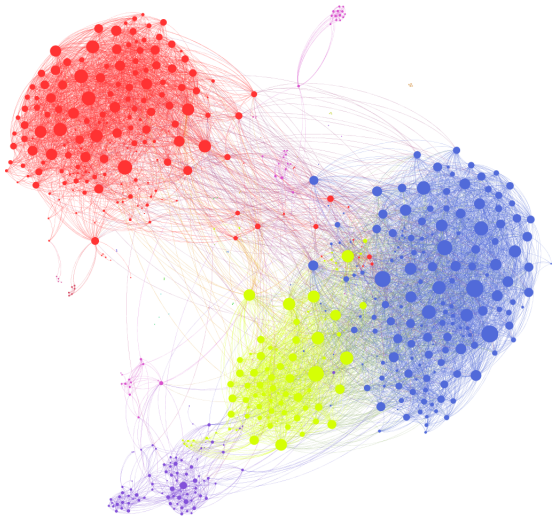
La compréhension d'un phénomène de groupe, qu'il soit sociologique, économique, biologique ou informatique, passe par l'observation et l'analyse individuelle des membres du groupe, mais aussi par celle de leurs interactions. La donnée de l'ensemble des interactions entre les membres forme ce qu'on appelle un réseau d'interactions.

Les réseaux sociaux constituent un exemple typique de réseau d'interactions. Un tel réseau est défini comme l'ensemble des liens sociaux établis entre les individus d'un groupe : liens d'amitié, politiques ou professionnels par exemple. Dès la fin du XIX^{e} siècle, Émile Durkheim ouvrait la voie à l'étude des réseaux en sociologie, en montrant que les facteurs individuels ne pouvaient à eux seuls expliquer les phénomènes sociaux (Durkheim, 1893) : par conséquent un groupe social ne se résume pas à la somme des individus, mais doit aussi sa structure aux interactions reliant les individus. L'idée était cependant loin d'être nouvelle, puisqu'au IV^{e} siècle avant Jésus-Christ, Aristote parlait de la constitution de la matière, mais aussi de l'organisation des cités grecques, en ces termes : « *La totalité est plus que la somme des parties* ». L'avènement de la sociométrie dans les années 1930 a conduit à formuler les premières questions de méthodologie pour l'analyse statistique de ce nouveau type de données.

En plus de la sociologie, d'autres disciplines, bien que de nature très différentes, peuvent en fait se retrouver soeurs dans le cadre formel des réseaux.

- Une épidémie dans une population peut être formalisée de la même façon que la propagation d'une rumeur dans un réseau social, ou que celle d'un virus informatique (Draief and Massoulié, 2010). La question commune à ces problèmes est de pouvoir prédire si le phénomène touchera la totalité de la population ou si elle sera contenue dans une fraction négligeable, par l'étude des propriétés du réseau.
- L'étude de la vulnérabilité de la structure du réseau de machines informatiques permet de prévenir des pannes aléatoires ou des attaques ciblées (Bol-

FIGURE 1.1 – Exemple de réseau d'amitiés. Les noeuds représentent l'ensemble des amis d'un membre de Facebook. Les arêtes symbolisent les liens d'amitié entre ces personnes. On distingue en couleurs plusieurs cercles d'amis fréquentés par le membre. Source : Griff's Graphs, griffsgraphs.com

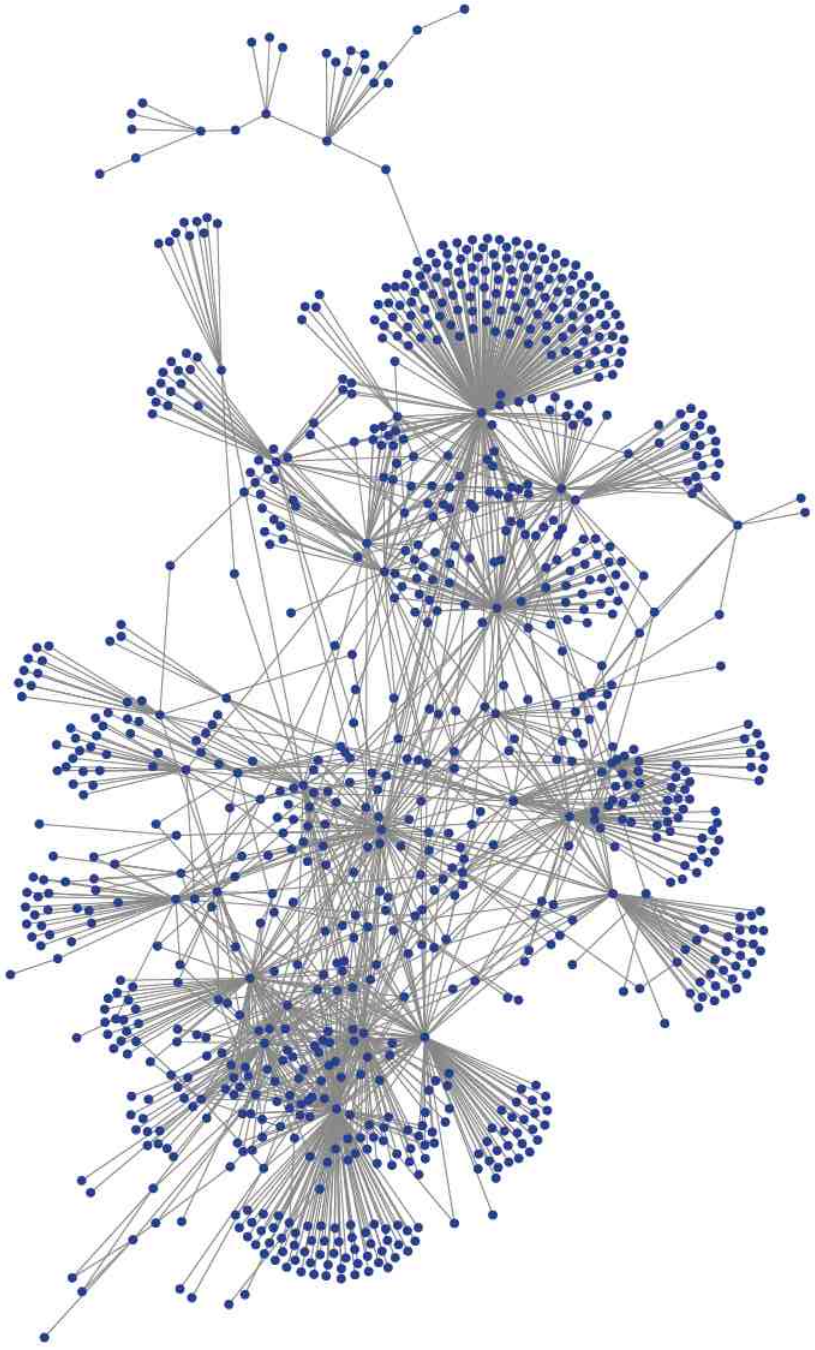


lobás and Riordan, 2004b). De même, la structure d'un d'un réseau de cultivateurs doit assurer au mieux la perpuétuation d'une variété de céréales, en cas de pertes de semences dues à des intempéries.

– En sociologie, la formalisation de la notion d'équivalence de rôle social par Lorrain and White (1971) a permis de schématiser la structure sociale d'un groupe. Elle est réinterprétée en biologie pour étudier les réseaux de régulation de protéines. Les protéines sont les briques élémentaires d'un organisme vivant. Son fonctionnement est contrôlé par un système d'auto-régulation des protéines entre elles. Le biologiste détermine expérimentalement le réseau en testant l'existence d'interactions biochimiques entre chaque paire de protéines. Les protéines peuvent être rangées par équivalence de rôles dans la régulation, qui ressemblent en fait aux rôles sociaux en sociologie. En visualisant le réseau à l'échelle des classes d'équivalence plutôt que des noeuds, le biologiste peut comprendre le système de régulation, et les fonctions biologiques complexes qui en découlent (Picard et al., 2009).

De manière générale, le nombre et la taille des données de réseau sont allés croissants depuis les débuts de l'analyse des réseaux avec quelques centaines d'éléments, jusqu'à plusieurs millions, voire milliards aujourd'hui. Une première explosion du volume des données a eu lieu dans les années 1990 avec leur numérisation massive, et une deuxième plus grande encore dans les années 2000 avec la démocratisation

FIGURE 1.2 – Exemple de réseau d'interactions protéine-protéine. Source : Edward Rietman, cancer-systems-biology.org



d'Internet, qui recèle aujourd'hui de nombreux réseaux sociaux : blogs, réseaux professionnels, réseaux de partages de données pairs-à-pairs, etc. Le World Wide Web est lui-même un réseau de pages, reliées entre elles par des liens hypertextes. En biologie, le séquençage du génome a fourni aussi de très grands réseaux d'interaction génétique. Un des enjeux majeurs pour le développement d'outils d'analyse statistique de réseaux est celui de pouvoir traiter ces très grands jeux de données. Dans cette thèse, on s'intéresse notamment à donner des algorithmes peu complexes, pour traiter de grands réseaux.

Un réseau est formalisé en mathématiques par un graphe, les membres du réseau sont appelés noeuds, et les interactions sont représentées par des paires de noeuds, appelées arêtes. En 1735, Euler initie la théorie des graphes en présentant à l'Académie de Saint-Pétersbourg une solution au problème des septs ponts de Königsberg, publiée plus tard (Euler, 1741). Les graphes sont longtemps vus comme des objets aux vocations uniquement algébriques. L'introduction de probabilités dans les graphes ne voit le jour qu'à la fin des années 1950 avec les travaux des mathématiciens hongrois Paul Erdős et Alfréd Rényi, qui étudient en 1959 un modèle qui porte leur nom depuis (Erdős and Rényi, 1959). La simplicité de ce modèle permet son étude exhaustive, mais empêche aussi un bon ajustement à des données réelles. En particulier, tous les noeuds du réseau agissent indépendamment les uns des autres et de la même façon dans le groupe, ce qui semble peu plausible. C'est un enjeu majeur dans la formulation de modèles de réseaux d'interaction que de trouver un compromis entre la richesse de leurs propriétés et la simplicité de leur étude.

Un modèle statistique de graphes aléatoires doit satisfaire un compromis supplémentaire : son inférence doit en plus, c'est-à-dire la détermination des caractéristiques de la population selon le modèle. Dans les années 1980, la sociologie produit plusieurs modèles hétérogènes permettant de prendre en compte les caractéristiques individuelles, par exemple des modèles de type régression logistique (Holland and Leinhardt, 1981), mais où les liens sont encore indépendants. L'inférence peut être faite classiquement par la méthode du maximum de vraisemblance. Le cadre de la famille exponentielle, une fois adaptée aux graphes, permet d'introduire de manière simple des dépendances entre les arêtes, par exemple entre celles qui partagent un noeud dans les modèles de Markov (Frank and Strauss, 1986). Le moindre ajout de dépendances complique très vite l'étude d'un réseau, et l'étude des techniques d'inférence dans ces modèles est d'ailleurs toujours d'actualité.

Les travaux originaux de cette thèse concernent plus précisément les modèles de réseaux hétérogènes où les caractéristiques individuelles ne sont pas observées, dits modèles à variables latentes. Un enjeu crucial dans ce contexte est de décrire la structure latente de la population, uniquement à partir du réseau observé. Cela inclut :

- Tester si cette population est effectivement hétérogène.
- Si elle l'est, faire du clustering, c'est-à-dire classer [1] les noeuds dans des sous-groupes plus homogènes.
- Estimer les paramètres du modèle.

Après ce préambule, les trois chapitres de la première partie visent à recréer le paysage dans lequel s'insèrent les travaux originaux de cette thèse.

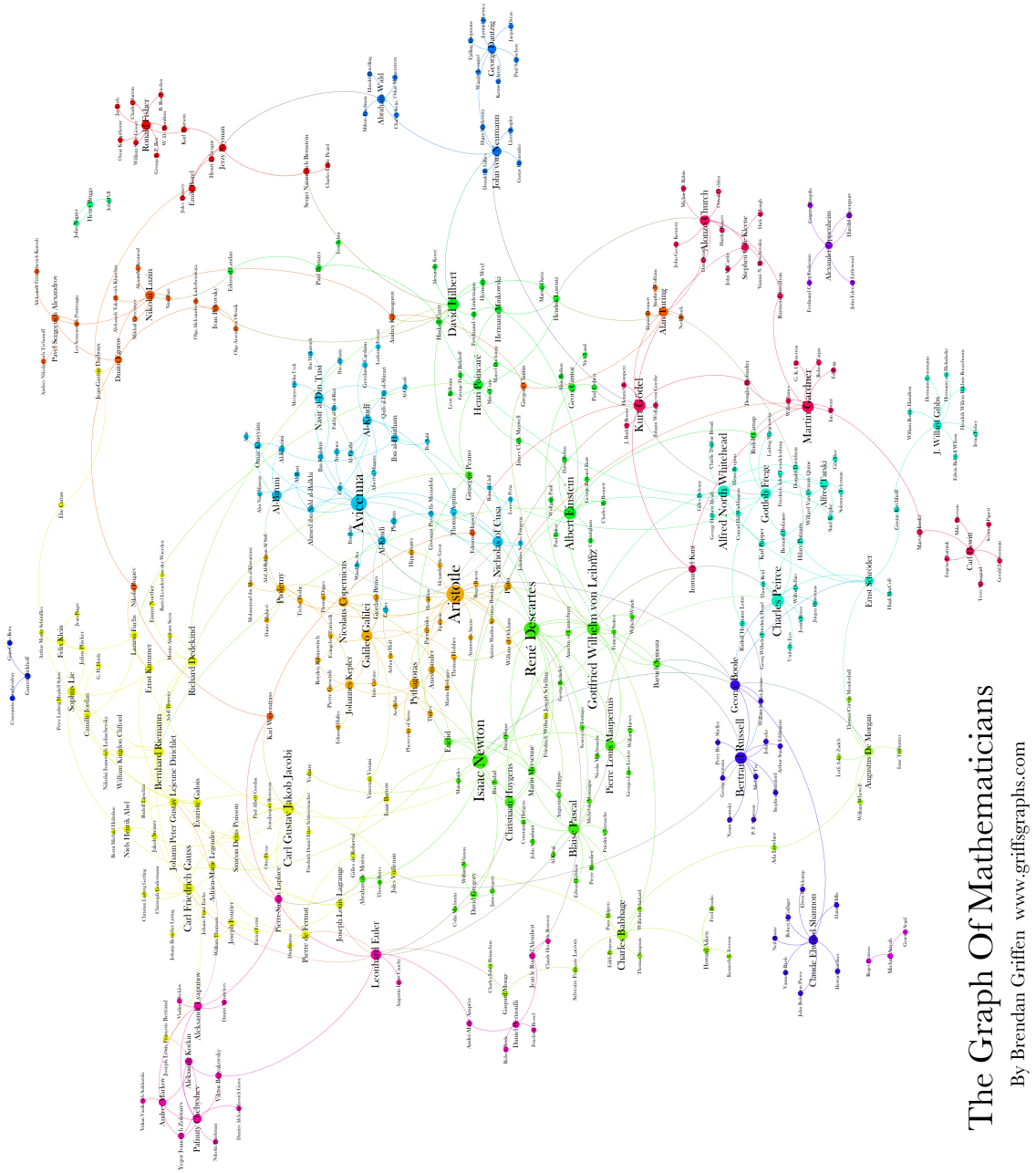
- Le Chapitre 2 commence par présenter le modèle d'Erdős-Rényi. L'étude de sa connexité est un peu plus développée, car elle sera utilisée dans le modèle KerNet au Chapitre 7, où l'on montrera que les graphes tirés dans ce modèle ressemblent localement à un graphe d'Erdős-Rényi. Quelques modèles de graphes aléatoires hétérogènes issus principalement de la physique sont aussi abordés. On montre comment ils s'y prennent pour réunir plusieurs propriétés empiriques dans un même modèle de sorte à s'approcher des réseaux réels, et pourquoi en l'état ils ne sont pas bien adaptés aux statistiques.
- Le Chapitre 3 a vocation à présenter des modèles de graphes conçus pour l'étude statistique de données de réseau. La première grande famille est celle des graphes aléatoires exponentiels, qui permettent notamment de contrôler paramétriquement les dépendances des arêtes et les tendances du graphe à produire certains motifs topologiques. Dans ces modèles, les variables individuelles sont supposées accessibles. Dans cette thèse, on s'intéresse à des modèles à variables non observées, dites latentes. La suite de ce chapitre présente ces modèles de manière générale, et comment l'inférence s'y fait classiquement. Il est conclu par la présentation du cadre commun aux modèles de graphes à variables latentes étudiés dans les travaux originaux de la thèse.
- Le Chapitre 4 s'intéresse au partitionnement de graphes aléatoires, et adopte deux points de vue. Le premier correspond à celui de la classification non supervisée et a pour objectif de retrouver les classes non observées auxquelles appartiennent les noeuds d'un graphe, dont l'hétérogénéité est décrite par un modèle de mélange, comme c'est le cas dans le Chapitre 6. On verra que le cas particulier des réseaux complique grandement l'inférence à cause de leur structure de dépendance particulièrement intriquée. Le second point de vue est purement algorithmique, et cherche des groupes homogènes selon un critère *ad hoc*. Nous choisirons finalement un compromis dans le Chapitre 7 en utilisant un algorithme associé à une définition du clustering basée sur un modèle, mais qui est peu contraint.
- Le Chapitre 5 a un statut particulier : il rappelle et montre quelques résultats qui seront utiles dans les Chapitres 6 et 7.

1. Dans cette thèse, on ne considère pas d'échantillon d'entraînement, ce qui nous distingue de la classification supervisée.

La seconde partie, composée des Chapitres 6 et 7, présentent les travaux originaux de cette thèse. Ils ont pour point commun de proposer des algorithmes très faiblement complexes (l'un est de complexité pseudo-linéaire en nombre de noeuds, l'autre linéaire en nombre d'arêtes), tout en ayant de bonnes propriétés asymptotiques dans les modèles auxquels ils sont appliqués.

- Le Chapitre 6 propose une méthode simple et consistente sous certaines hypothèses dans le Stochastic Blockmodel. Il s'agit d'un modèle où les noeuds ont une couleur non observée, et où la probabilité de connexion entre deux noeuds ne dépend que de leurs couleurs. La méthode est fondée sur l'étude fine de la distribution des degrés, qui permet toute l'inférence du modèle bien que ce ne soit qu'une marginale de la loi complète du graphe.
- Le Chapitre 7 propose une méthode tout aussi simple, et dont nous prouvons partiellement la consistence. Une étude de simulation vient étayer la partie théorique et donner des pistes pour des travaux futurs, visant à améliorer la robustesse de l'algorithme.

FIGURE 1.3 – Réseau des influences entre mathématiciens célèbres. Source : Griff's graphs, griffsgraphs.com



Chapitre 2

Modèles de graphes aléatoires

2.1 Quelques définitions et notations

2.1.1 Notations générales

Pour tout entier $n \geq 1$, on note $[n]$ l'ensemble des entiers entre 1 et n.

Le cardinal d'un ensemble fini V est noté $|V|$. Une paire d'un ensemble V est un sous-ensemble de V à deux éléments distincts. On note $\mathcal{P}(V)$ l'ensemble des paires d'éléments de V, et $\mathcal{P}_n$ l'ensemble des paires de $[n]$.

Soit $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites réelles. On dit que $(u_n)_{n\in\mathbb{N}}$ domine $(v_n)_{n\in\mathbb{N}}$ et on note $v_n = \mathcal{O}(u_n)$ s'il existe un entier N et un réel positif C tel que pour tout $n \geq N$, $|v_n| \leq C|u_n|$. On dit que $(v_n)_{n\in\mathbb{N}}$ est négligeable devant $(u_n)_{n\in\mathbb{N}}$ et on note $v_n = o(u_n)$ si pour tout $\varepsilon > 0$, il existe $N \in \mathbb{N}^*$ tel que pour tout $n \geq N$ $|u_n| \leq \varepsilon|v_n|$.

$\mathcal{B}(p)$ désigne la loi de Bernoulli de paramètre $p \in [0,1]$, $\mathcal{B}(n,p)$ la loi binomiale de paramètres $n \in \mathbb{N}^*$ et $p \in [0,1]$. $\mathcal{M}(k,\alpha)$ désigne la loi multinomiale de paramètres $k \in \mathbb{N}^*$ et $\alpha = (\alpha_1, \ldots, \alpha_Q) \in [0,1]^Q$ tel que $\sum_{i=1}^{Q} = 1$ (cas où l'on lance indépendamment k objets dans Q tiroirs au hasard selon les probabilités $\alpha_1, \ldots, \alpha_Q$). Enfin $\mathcal{N}_d(\mu, \Sigma)$ désigne la loi normale d-dimensionnelle de moyenne μ et de matrice de variance-covariance Σ de taille $d \times d$.

2.1.2 Quelques définitions dans les graphes

Un graphe simple non-orienté est une paire (V, E) où V est un ensemble et E un ensemble de paires de V. Les éléments de V sont appelés noeuds et ceux de E arêtes. Le graphe est dit fini si V est fini. Si $i, j \in V$ avec $i \neq j$, et $\{i, j\} \in E$, on dit que les noeuds i et j sont connectés, ou sont voisins.

La taille d'un graphe simple fini est égale au cardinal de V. Pour simplifier, dans toute cette thèse, sauf mention explicite, on supposera que $V = [n]$ lorsqu'il

est question d'un graphe de taille n.

On dit qu'un graphe (V', E') est inclus dans un autre graphe, ou est un sous-graphe de (V, E), si $V' \subset V$ et $E' \subset E$. Le sous-graphe de (V, E) induit par un ensemble $I \subset V$ est le graphe $(I, E \cap \mathcal{P}(I))$.

Un graphe plein est un graphe où chaque noeud est connecté à tous les autres noeuds. Une clique est un sous-graphe plein. Un triangle est une clique de trois noeuds.

Si (V, E) est un graphe fini non-orienté de taille n, sa matrice d'adjacence est la matrice (symétrique) $X \in \{0,1\}^{n \times n}$ telle que pour tout $i, j \in [n]$, $X_{ij} = \mathbb{1}_{\{i,j\} \in E}$, i.e. $X_{ij} = 0$ si et seulement si i et j ne sont pas connectés dans (V, E), 1 sinon. On identifiera toujours un graphe et sa matrice d'adjacence.

Si X est un graphe de taille n, on appelle degré d'un noeud $i \in [n]$ le nombre de voisins de i dans X. Il est noté D_i^X ou D_i si le contexte est clair, et s'écrit :

$$D_i^X = \sum_{j=1}^{n} X_{ij}$$

Soit $i, j \in [n]$. On dit qu'il existe un chemin dans le graphe X entre i et j s'il existe $i_0, i_1, \ldots, i_r \in [n]$ tels que $i_0 = i$, $i_r = j$ et pour tout $k \in [r]$, i_{k-1} et i_k sont connectés dans X. Sa longueur est le nombre r. Ceci définit une relation binaire sur les noeuds, qui est une relation d'équivalence. On appelle composantes connexes de X les classes d'équivalence de cette relation d'équivalence.

Les chemins permettent de définir une notion de distance dans le graphe : la distance entre deux noeuds est égale à la longueur du plus court chemin entre ces noeuds. Un cycle est un chemin de longueur non nulle qui part et arrive au même noeud. On appelle arbre un graphe connexe sans cycle.

On appelle densité d'un graphe X la fraction d'arêtes présentes sur le nombre d'arêtes possibles. Elle est notée $\zeta(X)$ et s'écrit :

$$\zeta(X) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n}^{n} X_{ij}$$

2.2 Le modèle d'Erdős-Rényi comme outil et référence

2.2.1 Définitions et premières propriétés du modèle

Le modèle d'Erdős-Rényi est parfois appelé « le » graphe aléatoire au sens où, dans ce modèle, la topologie du réseau est tirée uniformément au hasard. Historiquement deux modèles sont en fait associés au nom d'Erdős-Rényi, le modèle uniforme à nombre d'arêtes M fixé, noté $\mathcal{G}(n, M)$, et le modèle binomial, où la probabilité p de connexion entre deux noeuds quelconques est fixée, noté $\mathcal{G}(n, p)$. C'est

à ce dernier modèle qu'on donne communément aujourd'hui le nom d'Erdős-Rényi. Les deux modèles ne sont pas formellement équivalents, mais leurs propriétés sont liées dans une certaine asymptotique. Ils ont été étudiés sous toutes leurs coutures, plus particulièrement le second. Le lecteur peut trouver de nombreux résultats dans Bollobás (2001), Janson et al. (2000) ou dans Durrett (2007).

Modèle uniforme Dans ce modèle, le graphe aléatoire X est tiré uniformément dans l'ensemble des graphes à n noeuds et M arêtes. Le choix d'un tel graphe correspond au choix de ses paires de noeuds connectées, or on dénombre $N = \binom{n}{2}$ arêtes possibles dans un graphe à n noeuds. Le nombre possible de graphes à n noeuds et M arêtes est donc $\binom{N}{M}$. Ainsi pour tout graphe x à n noeuds et M arêtes,

$$P_{\mathcal{G}(n,M)}(X = x) = \binom{N}{M}^{-1}$$

La probabilité qu'une arête soit présente est identique pour toute paire de noeuds et vaut M/N. Ce modèle a été introduit par Paul Erdős et Alfréd Rényi en 1959 et a fait l'objet de plusieurs articles, dont l'article fondateur Erdős and Rényi (1959), ou encore Erdos and Rényi (1961); Erdős and Renyi (1961). Ce modèle théorique a été très étudié du point de vue des Probabilités, mais il reste peu pratique à utiliser puisque généralement dans les applications le nombre d'arêtes n'est pas déterministe.

Modèle binomial Dans ce modèle, on tire pour tous $i, j \in [n]$ avec $i < j$ une variable de Bernoulli X_{ij} de même paramètre $p \in]0, 1[$, les variables étant mutuellement indépendantes. Le graphe aléatoire X est représenté par la matrice d'adjacence symétrique $(X_{ij})_{i,j\in[n]}$. X a un nombre d'arêtes aléatoire, qui suit une loi binomiale de paramètres (N, p). Tout graphe à M arêtes a pour probabilité d'apparition $p^M(1-p)^{N-M}$ quelque soit le graphe choisi dans l'ensemble des graphes à n noeuds et M arêtes ; la loi conditionnelle de X sachant que X a M arêtes est donc $\mathcal{G}(n, M)$. Enfin, la loi des degrés est binomiale. Pour tout $i \in [n]$,

$$D_i^X \sim \mathcal{B}(n-1, p)$$

Le modèle $\mathcal{G}(n, p)$ déjà étudié par Erdős à la fin des années quarante, a aussi été introduit par Gilbert en 1959 dans l'article Gilbert (1959).

Lien entre les deux modèles Le lien entre ces modèles est notamment établi dans Janson et al. (2000). Ces modèles sont liés dans le cadre asymptotique particulier où le nombre d'arêtes M pour l'un, la probabilité de connexion p pour l'autre, dépendent de la taille n du graphe. En effet le nombre d'arêtes de $\mathcal{G}(n, p)$ suit une

loi binomiale qui par l'inégalité de Hoeffding (voir 5.1) se concentre très vite autour de son espérance Np. Ainsi, intuitivement si on pose $p_n = M/N$, $\mathcal{G}(n, p_n)$ se comporte de la même façon que $\mathcal{G}(n, M)$ quand n tend vers l'infini. Janson et al. (2000) éclaircit ceci : la plupart du temps, une propriété de graphe est asymptotiquement vérifiée dans $\mathcal{G}(n, p_n)$ si et seulement si elle l'est dans $\mathcal{G}(n, M)$.

2.2.2 Exemple d'application du modèle d'Erdős-Rényi

Nous donnons un exemple d'application du modèle d'Erdős-Rényi en épidémiologie, pour illustrer ses hypothèses de modélisation, et motiver l'étude de ses propriétés présentées dans les sections suivantes. Nous présentons le modèle de Reed-Frost, puis expliciterons l'équivalence avec le modèle d'Erdős-Rényi.

Le modèle de Reed-Frost[1] appartient à la classe des modèles d'épidémie de type SIR (pour Susceptible, Infecté, Retiré). Dans ces modèles, chaque individu de la population a à chaque instant un des trois états S, I ou R, selon qu'il n'a jamais été infecté par la maladie et *susceptible* de l'être, en cours d'*infection*, ou *retiré* (mort ou immunisé). Le modèle de Reed-Frost correspond à une version à temps discret de ces modèles. Un seul individu est infecté à l'origine. La dynamique est alors la suivante :

1. La population est fermée : la transmission de la maladie ne peut se faire qu'entre ses n individus.
2. La maladie ne se transmet que par un seul type de contact, et la probabilité p d'avoir ce contact dans une unité de temps est la même entre tous les individus et tout au long du temps.
3. Les contacts entre individus sont indépendants.
4. Un individu infecté le reste pendant une unité de temps puis est immunisé définitivement.

On doit à Barbour and Mollison (1990) la construction mathématique du modèle de Reed-Frost à partir du modèle d'Erdős-Rényi. On suppose que X est un graphe tiré dans le modèle $\mathcal{G}(n, p)$. Soit $i \in [n]$ l'individu infecté à l'origine.

On note δ_X la distance dans le graphe X. L'état $e_t(j)$ de l'individu $j \in [n]$ au temps $t \in \mathbb{N}$ est défini par :

$$e_t(j) = \begin{cases} S \text{ si } \delta_X(i,j) > t \\ I \text{ si } \delta_X(i,j) = t \\ R \text{ si } \delta_X(i,j) < t \end{cases}$$

Ainsi les contacts ayant donné lieu à une infection (si aucun des deux noeuds n'était pas déjà retiré) sont matérialisés par les arêtes du graphe. L'indépendance

1. Proposé par Lowell Reed et Wade Hampton Frost dans les années 1920.

des contacts correspond à l'indépendance des variables de Bernoulli dans le modèle $\mathcal{G}(n,p)$ et l'égalité de la probabilité d'avoir un contact infectieux entre tous les individus, à l'égalité du paramètre p de ces variables.

L'épidémie se propage depuis le noeud i vers ses voisins directs, puis itérativement vers la couche de voisins directement supérieure à chaque nouvelle unité de temps. La composante connexe de i dans X représente alors l'ensemble des noeuds qui sont retirés à la fin, c'est-à-dire ceux qui ont été un jour infectés. L'étude des composantes connexes du modèle $\mathcal{G}(n,p)$ et de leur taille peut donc être interprétée en termes épidémiques et répond à la question de la proportion de la population touchée par la maladie. Les grandes lignes de cette étude sont présentées dans les sections suivantes.

On peut s'apercevoir que les hypothèses faites sur la dynamique du modèle sont très simplificatrices et peu plausibles. Bien que le modèle d'Erdős-Rényi n'ait en fait pas les propriétés des réseaux réels, il fournit un cadre mathématique très agréable, qui permet une analyse approfondie. Il est une étape indispensable avant l'élaboration et la compréhension de modèles plus complexes.

2.2.3 Composante géante dans le modèle $\mathcal{G}(n,p)$

La plupart des propriétés du modèle d'Erdős-Rényi possèdent une transition de phase par rapport au paramètre p, que ce soit le fait d'avoir une composante géante ou d'être connexe, d'avoir un certain diamètre, de contenir au moins un motif donné, etc.

On aborde ici la question de la composante géante. On se place dans une asymptotique poissonienne, c'est-à-dire que l'on considère la suite de modèles $\mathcal{G}(n,p_n)$, où le degré moyen[2] $\lambda = np_n$ est un paramètre fixe. Le nom de cette asymptotique vient en particulier du fait que les degrés convergent alors en loi vers une loi de Poisson de paramètre λ. On note $\mathcal{C}_{(1)}, \mathcal{C}_{(2)} \dots$ les composantes connexes rangées dans l'ordre décroissant de taille. Le principal résultat est la transition de phase que subit la taille de la plus grande composante $\mathcal{C}_{(1)}$ par rapport au paramètre λ. Draief and Massoulié (2010) donnent les théorèmes suivants (voir aussi les références mentionnées au début de cette section) :

Théorème 2.1.

Régime sous-critique $\lambda < 1$. *Il existe une constante* $A_1 \geq 0$ *dépendant de* λ *telle que :*

$$\lim_{n\to\infty} P_{\mathcal{G}(n,p_n)} \left(|\mathcal{C}_{(1)}| \leq A_1 \log(n) \right) = 1$$

2. Le degré moyen est rigoureusement $(n-1)p_n$, mais les deux sont asymptotiquement équivalents quand $n \to \infty$.

Régime sur-critique $\lambda > 1$. *Soit* $p_{GW}(\lambda)$ *l'unique solution de* $x = e^{-\lambda(1-x)}$ *dans* $]0,1[$. *Il existe une constante* $A_2 > 0$ *dépendant de* λ *telle que pour tout* $\varepsilon > 0$ *:*

$$\lim_{n\to\infty} P_{\mathcal{G}(n,p_n)}\left(\left|\frac{|\mathcal{C}_{(1)}|}{n} - (1 - p_{GW}(\lambda))\right| \leq \varepsilon,\ |\mathcal{C}_{(2)}| \leq A_2 \log(n)\right) = 1$$

Régime critique $\lambda = 1$. *Il existe une constante* $A_3 > 0$ *telle que pour tout* $b > 0$ *:*

$$P_{\mathcal{G}(n,p_n)}\left(|\mathcal{C}_{(1)}| \geq bn^{2/3}\right) \leq \frac{A_3}{b^2}$$

Dans le régime sous-critique, les composantes sont asymptotiquement toutes de taille au plus logarithmique en n. Dans le régime sur-critique, elles sont aussi asymptotiquement toutes de taille au plus logarithmique, sauf la plus grande, qui englobe une proportion non négligeable de noeuds : il y a donc une composante géante asymptotiquement presque sûrement. Cette proportion est $1 - p_{GW}(\lambda)$, où $p_{GW}(\lambda)$ est en fait la probabilité d'extinction d'un processus de Galton-Watson dont la loi de descendance est la loi de Poisson de paramètre λ. On pouvait noter que ce processus apparaissait déjà dans la formulation du modèle de Reed-Frost, qui construit en fait un arbre de descendance dans le graphe. Le régime critique est plus compliqué que les deux autres ; nous ne donnons que cette borne supérieure simple qui permet d'avoir une idée de l'ordre de grandeur $n^{2/3}$ de la plus grande composante. On peut trouver des résultats plus détaillés dans Janson et al. (1993); Bollobás (2001).

2.2.4 Connexité et noeuds isolés dans le modèle $\mathcal{G}(n,p)$

La question de la connexité est un peu plus développée, car elle sera utilisée dans le chapitre 7. On y prouvera que certains sous-graphes d'un modèle plus complexe sont connexes avec grande probabilité, en montrant qu'ils contiennent un graphe d'Erdős-Rényi pour un certain paramètre p. Nous utiliserons alors notamment la borne de la Proposition 2.2 ci-dessous pour conclure.

Un des premiers résultats portant sur la connexité dans le modèle $\mathcal{G}(n,p)$ date de l'article Gilbert (1959). Gilbert y donne une formule exacte de la probabilité qu'un graphe aléatoire soit connexe sous le modèle $\mathcal{G}(n,p)$, notée dans ce paragraphe $\pi_{n,p}$. Il utilise une méthode courante en combinatoire, basée sur les séries génératrices : il écrit d'abord $\pi_{n,p}$ en fonction de $C_{n,l}$, défini comme le nombre de graphes connexes à n noeuds et l arêtes. On l'a vu plus tôt, si $N = \binom{n}{2}$, tout graphe à n noeuds et l arêtes a pour probabilité d'apparition $p^l q^{N-l}$ dans ce modèle, où $q = 1 - p$. D'où :

$$\pi_{n,p} = \sum_{l=n-1}^{N} C_{n,l} p^l q^{N-l}$$

On note qu'un graphe à n noeuds connexe à n noeuds a au moins $n-1$ arêtes (s'il en a exactement $n-1$, c'est même un arbre), et donc $C_{n,l} = 0$ pour tout $l < n-1$. Gilbert écrit une série génératrice pour la suite double $(C_{n,l})_{n\in\mathbb{N}^*,l\in\mathbb{N}}$ et la convertit en série génératrice pour $\pi_{n,p}$. Par des développements en série formelle et en notant $\mathcal{R}_n$ l'ensemble des partitions[3] de l'entier n, Gilbert obtient :

$$\pi_{n,p} = n! \sum_{(r_1,\dots,r_n)\in\mathcal{R}_n} \frac{(-1)^s(s-1)!q^{(n^2-1^2r_1-\cdots-n^2r_n)/2}}{r_1!\dots r_n!(1!)^{r_1}\dots(n!)^{r_n}}$$

où dans la somme, $s = r_1 + \cdots + r_n$. $\pi_{n,p}$ s'écrit donc comme une somme dont le nombre de termes est le cardinal de $\mathcal{P}_n$. Or ce nombre croît très vite avec n et la formule est donc rapidement incalculable...

Encadrement de la probabilité de non-connexité

Pour majorer la probabilité que le graphe soit non connexe, il utilise plutôt la relation de récurrence suivante :

Proposition 2.1. *(Gilbert, 1959)*

$$1 - \pi_{n,p} = \sum_{k=1}^{n-1} \binom{n-1}{k-1} \pi_{k,p} q^{k(n-k)}$$

Cette formule n'est pas prouvée dans Gilbert (1959), et peut être obtenue comme suit.

Démonstration. On considère le noeud n. Le graphe n'est pas connexe si et seulement si sa composante connexe notée $\mathcal{C}_n$ est de taille $k \in [n-1]$. De plus un ensemble J de noeuds forme une composante connexe de X si et seulement si le sous-graphe de X induit par J est connexe et s'il n'y a aucune arête venant des noeuds de J, d'où :

$$\begin{aligned} 1 - \pi_{n,p} &= \sum_{k=1}^{n-1} P(\mathrm{Card}(\mathcal{C}_n) = k) \\ &= \sum_{k=1}^{n-1} \sum_{I\in\mathcal{P}_{k-1}([n-1])} P\left((X_{ij})_{i,j\in I\cup\{n\}} \text{ connexe}, \bigcap_{\substack{i\in I\cup\{n\} \\ j\in[n-1]\setminus I}} \{X_{ij} = 0\} \right) \end{aligned}$$

3. Une partition de l'entier n est un n-uplet d'entiers naturels $(r_1,\dots,r_n)$ tel que $r_1 + 2r_2 + \cdots + nr_n = n$

où $J = I \cup \{n\}$, et $\mathcal{P}_{k-1}([n-1])$ l'ensemble des parties à $k-1$ éléments de l'ensemble $[n-1]$. Pour $I \in \mathcal{P}_{k-1}([n-1])$, le sous-graphe induit par $I \cup \{n\}$ a pour loi $\mathcal{G}(k,p)$, et sa probabilité de connexité est $\pi_{k,p}$. D'autre part la probabilité qu'aucun des $n-k$ autres noeuds ne soit connecté à un noeud de $\mathcal{C}_1$ est $q^{k(n-k)}$, puisqu'il y a $k(n-k)$ liens possibles d'un ensemble de k noeuds à son complémentaire. Par indépendance des arêtes du sous-graphe induit par $I \cup \{n\}$ et des arêtes pointant vers un noeud de $I \cup \{n\}$, le terme sommé est donc égal à $\pi_{k,p} q^{k(n-k)}$. De plus le cardinal de $\mathcal{P}_{k-1}([n-1])$ est $\binom{n-1}{k-1}$, d'où le résultat.

□

Un des résultats principaux de Gilbert (1959) est l'encadrement de la probabilité qu'un graphe aléatoire sous le modèle $\mathcal{G}(n,p)$ soit non connexe, dont nous obtenons une simplification ci-dessous, qui est donnée à la Proposition 2.2. La conséquence en est que pour tout $p > 0$ fixé (ne dépendant pas de n) le graphe est asymptotiquement presque sûrement connexe.

Proposition 2.2.

$$nq^{n-1}\left(1 - \frac{n-1}{2}q^{n-1}\right) \leq 1 - \pi_{n,p} \leq nq^{n-1}\left(1 + (n-1)q^{(n-2)/2}\exp\left((n-1)q^{(n-2)/2}\right)\right)$$

La borne supérieure est en particulier obtenu en majorant simplement par 1 les $\pi_{k,p}$ dans la relation de récurrence précédente, puis en utilisant une inégalité de convexité pour la fonction $x \mapsto q^{x(n-x)}$. La borne inférieure est obtenue en minorant la probabilité d'être non connexe par celle de l'existence d'un noeud isolé. En effet s'il existe un noeud isolé, il forme une composante connexe à lui tout seul et le graphe n'est donc pas connexe. La probabilité que le graphe ait un noeud isolé est ensuite elle-même minorée à l'aide d'une inégalité de Bonferroni à deux termes.

Par ailleurs, cet encadrement de $1 - \pi_{n,p}$ permet d'en donner le développement asymptotique ci-dessous. Ce développement pourrait être poussé plus loin en prenant plus de termes dans l'inégalité de Bonferroni du minorant, et en développant aux ordres supérieurs l'exponentielle dans le majorant.

$$1 - \pi_{n,p} = nq^{n-1} + \mathcal{O}\left(n^2 q^{3n/2}\right)$$

La minoration de la Proposition 5.1 par la probabilité qu'il existe un noeud isolé peut sembler brutale, mais en fait cette probabilité est asymptotiquement équivalente à la probabilité que le graphe soit non connexe, ce que nous déduisons ci-dessous des résultats de l'article de Gilbert. Si on note I_n l'événement « il existe un noeud isolé » :

$$nq^{n-1}\left(1 - \frac{n-1}{2}q^{n-1}\right) \leq P_{\mathcal{G}(n,p)}(I_n) \leq 1 - \pi_{n,p} \leq nq^{n-1}(1 + o(1))$$

Or le minorant et le majorant sont tous deux équivalents à nq^{n-1}, d'où :

Proposition 2.3.

$$P_{\mathcal{G}(n,p)}(I_n) \sim_{n\to+\infty} 1 - \pi_{n,p} \sim_{n\to+\infty} nq^{n-1}$$

En résumé, la non-connexité du graphe vient asymptotiquement de noeuds isolés, et pas d'une non-connexion entre deux grandes composantes. On peut interpréter ce phénomène par une observation simple. La probabilité qu'un ensemble donné de k noeuds soit non connecté à son complémentaire est $q^{k(n-k)}$, où $k(n-k)$ est le nombre d'arêtes possibles entre l'ensemble et son complémentaire (voir aussi la preuve de la Proposition 2.1). Or la fonction $x \mapsto x(n-x)$ croît très vite en 1 pour atteindre son maximum en $n/2$, de sorte qu'il y a beaucoup plus de liens à casser pour séparer une composante de taille $k > 1$ de son complémentaire qu'il n'y en a pour séparer un seul noeud. Le scénario d'un noeud isolé est ainsi beaucoup plus probable. Ce phénomène assez général n'est pas propre qu'à Erdős-Rényi et se produit dans des modèles de nature très différente, comme les graphes aléatoires géométriques (Gupta and Kumar, 1998), ou dans le modèle KerNet du Chapitre 7.

Vitesse critique du degré moyen pour la connexité

Les inégalités de Gilbert permettent de conclure à la connexité presque sûre asymptotique dans le régime à densité fixe $p > 0$. Pour trouver le régime critique pour la connexité entre ce régime et le régime $p = 0$ où le graphe est vide, on envisage que la densité tende vers zéro quand la taille du graphe croît : $p = p_n \to 0$. Le cadre asymptotique s'en trouve changé, et la précision des inégalités précédentes altérée. En particulier le minorant ne permet jamais de conclure à la non-connexité presque sûre asymptotique. Le majorant permet quand même de donner un minorant de la vitesse seuil de p_n :

$$\text{Si } \underline{\lim} \frac{np_n}{\log(n)} > \frac{4}{3}, \text{ alors } P_{\mathcal{G}_{n,p_n}}(X \text{ connexe}) \xrightarrow[n\to\infty]{} 1$$

La question qui se pose est quelle peut être le type de la vitesse critique. La clé du problème passe à nouveau par le contrôle des noeuds isolés dont le nombre N_{isol} a pour espérance :

$$\mathbb{E}\left(N_{isol}\right) = nq_n^{n-1} = \exp\left(-(np_n - \log(n)) + o(np_n)\right)$$

La quantité déterminante est donc $c_n = np_n - \log(n)$, et on distingue deux régimes différents selon que le degré moyen np_n domine asymptotiquement $\log(n)$ ou le contraire, i.e. $c_n \to +\infty$ ou $c_n \to -\infty$. Dans le premier, le nombre de points isolés explose, tandis qu'il tend vers zéro dans l'autre.

Loi limite du nombre de noeuds isolés. On se place dans le cas où $c \in \mathbb{R}$ et $p_n = \frac{\log(n)+c}{n}$. Erdős et Rényi ont montré par la méthode des moments factoriels que dans cette asymptotique, N_{isol} converge en loi vers une variable de loi de Poisson de paramètre e^{-c}, d'où :

Proposition 2.4. *(Erdős-Rényi)*

$$P_{\mathcal{G}(n,p_n)}(N_{isol} = 0) \xrightarrow[n\to\infty]{} e^{-e^{-c}}$$

Ce résultat a une démonstration plus moderne et plus simple avec la méthode de Stein-Chen, qui n'utilise que deux moments (voir Arratia et al. (1989) pour cette méthode, Draief and Massoulié (2010) pour la preuve du résultat précédent avec cette méthode). Introduite par Stein en 1972 pour majorer la distance en variation à une loi limite normale, puis généralisée à une loi limite de Poisson par Chen en 1975 dans Chen (1975), c'est un outil souvent utilisé dans les graphes (voir Penrose (2003)) et les processus stochastiques dans un contexte de dépendances faibles. On peut consulter le livre Barbour et al. (1992) pour avoir un panorama d'applications et de références.

Disparition des composantes de taille ≥ 2. Toujours dans cette asymptotique, on peut montrer que la probabilité que le graphe aléatoire du modèle $\mathcal{G}_{n,p_n}$ ait une composante de taille $k \in \{2, n/2\}$ tend vers zéro. La conséquence en est que dans cette asymptotique aussi, la non-connexité vient des noeuds isolés, d'où le théorème suivant, appelé *double exponentielle* :

Proposition 2.5.

$$\lim_{n\to\infty} P_{\mathcal{G}(n,p_n)}(X \text{ connexe}) = \lim_{n\to\infty} P_{\mathcal{G}(n,p_n)}(N_{isol} = 0) = e^{-e^{-c}}$$

On en déduit dans le théorème suivant que $\log(n)$ est la vitesse critique du degré moyen np_n pour la connexité :

Théorème 2.2. *(Erdős-Rényi) Soit $c_n = np_n - \log(n)$. Alors :*

- *Si $c_n \to +\infty$, alors $P_{\mathcal{G}(n,p_n)}(X \text{ connexe}) \xrightarrow[n\to\infty]{} 1$.*
- *Si $c_n \to -\infty$, alors $P_{\mathcal{G}(n,p_n)}(X \text{ connexe}) \xrightarrow[n\to\infty]{} 0$.*

On peut reformuler ce théorème en définissant le paramètre fixe $\lambda = \frac{np_n}{\log(n)}$ et en disant que la connexité subit une transition de phase par rapport à ce paramètre en la valeur 1 : si $\lambda > 1$, le graphe est connexe presque sûrement asymptotiquement, tandis que si $\lambda < 1$, il est non connexe presque sûrement asymptotiquement. Mais le théorème est en fait un peu plus précis, puisqu'il précise aussi certains cas critiques $\lambda = 1$, notamment les cas où c_n tend vers l'infini moins vite que $\log(n)$.

Notons aussi que cette transition de phase a lieu à une vitesse de densité plus rapide que celle de la transition de phase pour la composante géante. C'est tout à fait naturel, puisqu'on demande à la composante géante des arêtes supplémentaires pour englober tous les noeuds au lieu d'une proportion fixe. Cet effort supplémentaire est en fait faible, puisque la vitesse demandée est à peine plus rapide, $\log(n)/n$ au lieu de $1/n$ dans l'asymptotique poissonienne.

Diamètre

On se place dans le régime $\frac{np_n}{\log(n)} \to +\infty$. Alors sous le modèle $\mathcal{G}(n, p_n)$ le graphe aléatoire est asymptotiquement presque sûrement connexe d'après ce qui précède. Au cas où le graphe est connexe, il existe pour toute paire de noeuds un chemin dans le graphe qui relie ces noeuds. On peut se demander quelle est la longueur maximale de ces chemins, i.e. le diamètre du graphe. Le théorème suivant adapté de Draief and Massoulié (2010) dit que dans ce régime le diamètre du graphe dans le modèle $\mathcal{G}(n, p_n)$ croît strictement moins vite que logarithmiquement en n :

Théorème 2.3. *On suppose que* $\frac{np_n}{\log(n)} \to +\infty$. *Il existe une constante* $B \geq 0$ *telle que :*

$$\lim_{n\to\infty} P_{\mathcal{G}(n,p_n)}\left(\operatorname{diam}(X) \leq B\frac{\log(n)}{\log\log(n)}\right) = 1$$

2.2.5 Retour à la modélisation

Le modèle de Reed-Frost repose sur deux hypothèses simplistes qui correspondent à celles du modèle d'Erdős-Rényi : l'indépendance des contacts, et l'égalité de la probabilité de contact infectieux entre tous les individus. On peut interpréter de la même manière le modèle d'Erdős-Rényi dans les réseaux sociaux. Les arêtes représentent alors des interactions sociales au lieu de contacts infectieux. Plus précisément :

Indépendance des interactions Les individus se connectent à un noeud indépendamment d'avec qui ce noeud est connecté, et même d'avec qui lui-même est connecté. Cependant c'est une hypothèse peu plausible dans un réseau social : en général un ami nous présente ses autres amis.

Homogénéité des interactions La probabilité p de connexion entre tout couple d'individus est la même, ce qui ne semble pas plus plausible. En particulier, cela implique que tout individu a une probabilité non nulle de se connecter à n'importe quel autre, malgré la contrainte géographique qui limite les connexions lointaines par exemple. Le milieu socio-professionnel détermine aussi en partie les relations les plus probables. De plus dans ce modèle chaque individu a en espérance le même nombre de voisins (degré). Pourtant dans une

population hétérogène chaque individu a une sociabilité différente, et donc a plus ou moins de voisins selon cette caractéristique qui lui est propre.

De manière globale, le graphe d'Erdős-Rényi n'a en moyenne pas de structure particulière : à nombre d'arêtes fixé, toutes les topologies ont la même probabilité d'exister. Cette hypothèse n'est pas cohérente avec le point de vue des réseaux sociaux où il est clair que les individus jouent des rôles sociaux différents qui façonnent la topologie du réseau. Dans un autre exemple, celui des réseaux d'interaction protéine-protéine, chaque protéine joue un rôle particulier dans le processus global de régulation, qui dépend de la structure de chaque protéine. Ainsi ce modèle ne tient pas compte des caractéristiques de chaque individu susceptibles d'influencer ses interactions, ni de la structure de dépendance sous-jacente à la population.

En résumé, ce modèle est une référence de par la connaissance assez exhaustive que l'on en a à l'heure actuelle. Il peut s'avérer un bon outil technique, dont on verra une application au Chapitre 7. De par sa propriété principale d'avoir une topologie complètement aléatoire, il est aussi utilisé comme modèle d'hypothèse nulle pour réaliser des tests statistiques. L'idée est de quantifier si une certaine propriété de graphe peut être due à une topologie aléatoire ou si elle s'en distingue significativement. La connaissance du modèle devient alors cruciale, puisqu'on a besoin de connaître le comportement de la statistique de test sous l'hypothèse nulle pour construire la région de rejet du test. On verra une application simple de ce principe dans le Chapitre 6, on peut aussi en trouver une autre dans Arias-Castro and Verzelen (2013) pour la détection de communautés. En revanche les hypothèses simplistes du modèle se heurtent à des observations élémentaires de modélisation et s'avère peu exploitable pour modéliser directement un réseau d'interaction.

2.3 Quelques modèles hétérogènes classiques

Après avoir bousculé le modèle d'Erdős-Rényi dans ses indépendances et son homogénéité, on peut se demander quelles sont les propriétés que l'on attend d'un modèle de graphe aléatoire, en décrivant empiriquement les réseaux d'interaction réels. Beaucoup de modélisateurs, en particulier physiciens, se sont attelés à comprendre les mécanismes d'interaction pour les inclure dans les modèles et ainsi apporter dépendances et hétérogénéité.

2.3.1 Propriétés empiriques des réseaux réels

Les articles de revue Albert and Barabási (2002) et Newman (2003) établissent une liste de propriétés empiriques de différents réseaux d'interaction, notamment suivant le domaine d'applications. Malgré leur diversité, ils peuvent avoir en commun certaines propriétés parmi les suivantes.

Connexité Les réseaux réels possèdent généralement un grand groupe d'individus en interaction, et éventuellement quelques petits groupes séparés. Ils sont donc généralement connexes ou ont une composante géante.

Petit monde Propriété rendue célèbre par l'expérience du sociologue Stanley Milgram Milgram (1967) appelée « six degrés de séparation ». Elle montra d'une part que dans un réseau d'amitiés, chaque personne était à moins de six poignées de mains d'une autre. Dit en termes de graphe, le plus court chemin reliant deux personnes comptait moins de six noeuds. On dit qu'un modèle vérifie la propriété de petit monde quand la distance moyenne dans le graphe entre deux noeuds est dominée par $\log(n)$. Le modèle d'Erdős-Rényi vérifie cette propriété, mais en quelque sorte par hasard, puisque sa topologie est aléatoire contrairement aux réseaux de personnes qui ont un fondement spatial. D'autre part l'expérience de Milgram montra aussi que pour trouver le chemin optimal dans le réseau entre deux personnes, il n'est pas besoin de le connaître dans sa globalité : la connaissance géographique approximative de la destination et la vision locale du réseau que possède chaque noeud suffit. Cette possibilité qu'offre le réseau s'appelle navigabilité.

Densité Généralement, la densité des réseaux réels tend vers zéro quand leur taille n tend vers l'infini, traduisant une capacité de connexion limitée des individus. Autrement dit, le nombre de relations effectives croît strictement plus lentement que le nombre de relations possibles. On dit que le modèle de graphe est creux ou de faible densité quand $\zeta(X) = o\left(\frac{1}{n}\right)$ presque sûrement ou en espérance.

Loi des degrés Cette loi focalise beaucoup l'attention, à tel point qu'existent même des modèles dont la vocation est d'avoir une loi des degrés prescrite : ce sont les modèles dits *à configuration* (Bender and Canfield, 1978; Molloy and Reed, 1995). Bien que les degrés aient fait couler beaucoup d'encre dans l'espoir de leur trouver une loi universelle, il semble que leur loi dépende crucialement du type de réseaux. La discussion porte essentiellement sur la queue de la distribution, lourde ou exponentielle. Les lois avancées dans le cas d'une queue lourde sont les lois dites *sans échelle* qui sont les lois de puissance. Une loi μ est dite de puissance si pour tout $k \in \mathbb{N}$:

$$\mu([k, +\infty[) = \alpha k^{-\beta}$$

où $\alpha, \beta \geq 0$ sont deux paramètres. Par extension, on appelle sans échelle les modèles ayant une loi des degrés sans échelle. Les plus anciennes traces de modèle sans échelle datent à notre connaissance d'un article de Lotka de 1926 (Lotka, 1926) sur des réseaux de collaboration de chercheurs. Leur renaissance récente et l'engouement qu'ils ont suscité sont dus à l'étude du World Wide Web, dont la distribution empirique des degrés semble être à queue lourde. Sans être universelle

comme remarqué par Albert and Barabási (2002) dans General Questions, elle est la signature de l'existence de concentrateurs dans le réseau, c'est-à-dire de noeuds de très haut degré. En réalité, un certain nombre d'exemples montrent une loi de puissance à coupure exponentielle à partir d'un certain degré, interdisant de trop forts concentrateurs. La distribution des degrés possède une queue exponentielle dans le modèle d'Erdős-Rényi. Cependant Albert and Barabási (2002) précise que dans le cas de réseaux réels ayant une queue purement exponentielle, ils restent cependant loin d'avoir une distribution des degrés de type Poisson comme dans le modèle d'Erdős-Rényi à faible densité.

Transitivité Dans le cas d'un réseau social, cette propriété est illustrée par l'expression « les amis de mes amis sont mes amis ». Elle signifie que sachant que deux noeuds i et j sont connectés au même noeud k, la probabilité d'être eux-mêmes connectés l'un à l'autre est plus grande :

$$P(X_{ij} = 1 \mid X_{ik} = 1, X_{jk} = 1) > P(X_{ij} = 1)$$

Une des caractéristiques d'un modèle transitif est le très grand nombre de triangles dans le graphe. On mesure cette tendance par les deux coefficients de regroupement définis respectivement dans Watts and Strogatz (1998) et Bollobás and Riordan (2003) de cette façon, si X est un graphe à n noeuds :

$$CL_1(X) = \frac{1}{n} \sum_{i \in [n]} \zeta(X_{\mathcal{V}_X(i)}) \text{ et } CL_2(X) = \frac{\sum_{i \in [n]} \binom{D_i^X}{2} \zeta(X_{\mathcal{V}_X(i)})}{\sum_{i \in [n]} \binom{D_i^X}{2}} \tag{2.1}$$

où pour rappel, $\zeta(X_{\mathcal{V}_X(i)})$ est la densité du graphe induit par les voisins de i dans X. $CL_2(X)$ est en fait trois fois le nombre de triangles du graphe divisé par le nombre d'arêtes ayant un noeud commun. $CL_1(X)$ et $CL_2(X)$ sont tous deux nuls si X est un arbre, et valent 1 si X est une clique, mais ne sont pas égaux en général. On dit que le premier (respectivement le second) coefficient de regroupement est élevé si :

$$\liminf_{n \to \infty} CL_1(X_n) > 0 \text{ p.s. (respectivement } \liminf_{n \to \infty} CL_2(X_n) > 0 \text{ p.s.)}$$

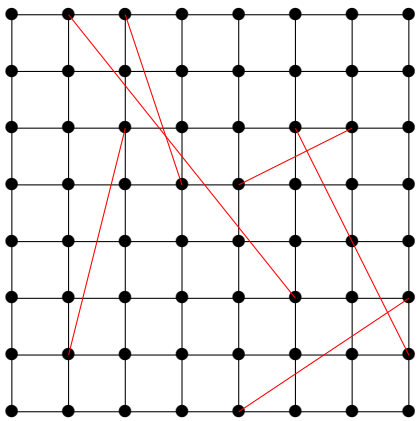
Autres propriétés Il existe beaucoup d'autres propriétés recherchées dans les modèles en fonction du domaine d'application, comme la robustesse de la connexité aux attaques supprimant des noeuds en informatique (voir Bollobás and Riordan (2004b)), la présence de certains motifs en biologie (Birmele (2012)), etc.

2.3.2 Modèles petit monde de Watts-Strogatz et de Kleinberg

Le modèle de Watts-Strogatz Watts and Strogatz (1998) a été conçu pour expliquer comment fort coefficient de regroupement et propriété de petit-monde peuvent cohabiter dans les réseaux sociaux. Cette combinaison ne peut en effet être due à une topologie aléatoire, puisque Erdős-Rényi ne possède que la propriété de petit monde. Kleinberg explique en plus la navigabilité dans Kleinberg (2000) en lui donnant une définition algorithmique rigoureuse.

Le principe commun au modèle originel et ses variantes successives (dont celle de Kleinberg) est de perturber un graphe régulier fondé sur un motif en créant aléatoirement des arêtes qui jouent le rôle de raccourcis. Dans Watts and Strogatz (1998) les raccourcis sont créés par redirection aléatoire des arêtes du graphe régulier. Dans les articles Newman and Watts (1999a,b), ils sont créés par ajout d'arêtes en plus de celles du graphe régulier. Chaque noeud crée avec probabilité $p \in]0, 1]$ un raccourci avec un noeud tiré uniformément au hasard parmi tous les autres noeuds (voir Figure 2.1). Les raccourcis font ainsi chuter le diamètre, dominé par $\log(n)$ (voir aussi Draief and Massoulié (2010)). Newman and Watts (1999a,b) étudie la transition de phase [4] entre une distance moyenne dans le graphe linéaire en n et la propriété de petit monde qui émerge quand $p > 0$.

FIGURE 2.1 – Réalisation d'un graphe dans le modèle de Watts-Strogatz sur une grille 7×7. De chaque noeud part un raccourci avec probabilité $p = 0.07$, vers un noeud choisi uniformément au hasard. En noir, les arêtes de la grille et en rouge, les raccourcis.



Cependant ces modèles n'expliquent pas la propriété de navigabilité de l'expérience de Milgram. Kleinberg (2000) montre en résumé qu'un algorithme décen-

4. Le schéma d'ajout d'arêtes au lieu de la redirection facilite techniquement cette étude.

tralisé, i.e. n'ayant connaissance que du graphe régulier et des arêtes partant des noeuds qu'il parcourt, ne peut trouver que des chemins de longueur moyenne en puissance de n, et donc bien plus grande que la distance moyenne dans le graphe qui est logarithmique en n.

Le modèle de Kleinberg est le suivant. Les noeuds sont identifiés aux points de la grille $[m] \times [m] \subset \mathbb{R}^2$, d'où $n = m^2$ noeuds. Les noeuds se trouvant à une distance L^1 inférieure à $D \in \mathbb{N}$ l'un de l'autre sont connectés de manière déterministe. En l'état, le diamètre du graphe est $\mathcal{O}(n/D)$. L'idée est à nouveau d'y ajouter des raccourcis aléatoires. De chaque noeud partent $r \in \mathbb{N}$ arêtes. Pour tous noeuds $i \neq j$ la probabilité que l'extrémité E d'une arête partant de i soit j est inversement proportionnelle à une certaine puissance de leur distance L^1 :

$$P(E = j) = \frac{\|i - j\|_1^{-\alpha}}{\sum_{k \neq i} \|i - k\|_1^{-\alpha}} \text{ où } \alpha \geq 0.$$

Pour $\alpha = 0$, les arêtes sont tirées uniformément et on retrouve un modèle à la Watts-Strogatz. Kleinberg (2000) montre que pour le paramètre critique $\alpha = 2$, et seulement pour ce paramètre, l'algorithme glouton [5] trouve des chemins de longueur moyenne dominée par $\log^2(n)$. Seul le modèle de paramètre $\alpha = 2$ est donc navigable.

Ces modèles ont une interprétation transparente en terme de réseaux sociaux. Par exemple dans le modèle de Kleinberg, la grille donne un fondement « géographique » au réseau, et le graphe régulier modélise les relations de voisinage. La pénalisation de la probabilité d'un raccourci par la distance de ses extrémités est aussi très cohérente avec l'idée selon laquelle l'éloignement géographique limite les relations. Aussi le paramètre α peut être interprété comme un niveau d'inertie des individus : plus il est grand, moins les raccourcis iront loin. Plus il est petit, plus ils pourront aller loin, jusqu'à la valeur critique $\alpha = 0$ où ils vont partout avec la même probabilité.

Les modèles petit monde affichent un premier coefficient de regroupement élevé, et la loi des degrés n'est pas à queue lourde selon Newman and Watts (1999a). Le fait que le graphe soit fondé sur un graphe régulier est à double tranchant : d'une part la structure spatiale amène de l'hétérogénéité, mais d'autre part il limite aussi fortement la variété des formes topologiques que le graphe peut prendre. Cette régularité, ainsi que le fait que la distribution des raccourcis est la même pour tous les noeuds, interdit de prendre en compte des caractéristiques sociales individuelles.

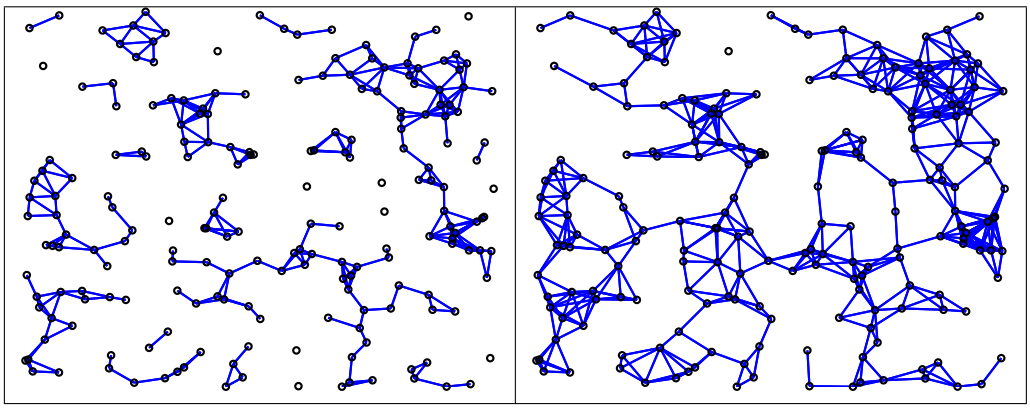
5. Le terme « glouton » désigne une stratégie générale d'optimisation itérative qui consiste à optimiser localement à chaque itération, en espérant obtenir l'optimisation globale au terme des itérations.

2.3.3 Graphes aléatoires géométriques

Dans la classe des modèles de graphes aléatoires géométriques, les noeuds sont identifiés à des positions aléatoires $(Z_i)_{i\in[n]}$ dans un espace muni d'une distance δ. Les arêtes se créent entre les noeuds à une distance inférieure à un seuil $r > 0$. Elles sont donc déterministes conditionnellement aux positions :

$$X_{ij} = \mathbb{1}_{\delta(Z_i,Z_j)\leq r}$$

FIGURE 2.2 – Réalisation de deux graphes aléatoires géométriques de taille $n = 200$ sur le carré unité de $\mathbb{R}^2$ avec une densité uniforme, $r = 0.07$ (gauche) et $r = 0.1$ (droite). Les positions $(Z_i)_{1\leq i\leq n}$ sont les mêmes dans les deux graphes, seul le seuil r change.



Cette classe de modèles modélise des réseaux de télécommunications sans fil ou des phénomènes tels que la propagation d'un incendie de forêt ou d'une maladie par exemple. Ces graphes sont étudiés de manière exhaustive dans Penrose (2003) notamment dans les cas où les positions sont tirées selon un processus de Poisson homogène sur $\mathbb{R}^d$, ou une densité uniforme sur $[0,1]^d$ par exemple. Notons le résultat de Gupta and Kumar (1998), qui montre qu'à l'instar des graphes d'Erdős-Rényi, la probabilité que le graphe aléatoire géométrique sur le disque unité de $\mathbb{R}^2$ soit connexe est asymptotiquement équivalente à la probabilité d'avoir un noeud isolé (voir 2.2.4). De plus la vitesse critique pour le degré moyen est aussi la même, malgré la nature différente de ces deux modèles. La démonstration faite en 7.3 dans un modèle proche fait un lien entre les deux types de graphes : localement, les graphes géométriques contiennent un graphe d'Erdős-Rényi.

2.3.4 Modèle de Barabási-Albert

Presque en même temps que Watts et Stogatz à la fin des années 90, Albert et Barabási proposent un modèle sans échelle dans Barabási and Albert (1999). Une

des caractéristiques de leur modèle, et de toute la classe des modèles de croissance, est qu'ils sont construits itérativement en attachant un noeud à ceux préexistant, à chaque itération. Du point de vue physique, la dynamique de croissance du graphe est ainsi décrite au niveau microscopique, éclairant les mécanismes topologiques du réseau qui conduisent à leurs propriétés macroscopiques.

En l'occurrence le principe d'*attachement préférentiel* consiste à relier le nouveau noeud de préférence aux noeuds déjà les plus populaires (ceux de plus grand degré) expliquant ainsi la formation des concentrateurs et la loi de puissance des degrés. Introduit en 1925 par Yule dans Yule (1925), ce schéma suscite un regain d'intérêt avec la modélisation de l'évolution du Web ; il s'applique aussi particulièrement bien aux réseaux de citations de chercheurs. Le modèle de Barabási and Albert (1999) est défini de manière rigoureuse et rebaptisé *Linearized Chord Diagram* (LCD) dans Bollobás et al. (2001).

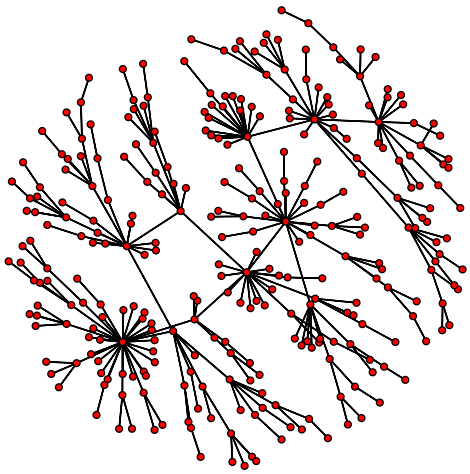
Dans la variante suivante de Draief and Massoulié (2010) qui dépend d'un paramètre p, la loi des degrés converge vers une loi de puissance d'exposant $(3-p)/(1-p)$, couvrant ainsi les exposants de 3 à $+\infty$. On part d'un graphe de base $X_0 = (V_0, E_0)$. On construit récursivement la suite de graphes aléatoires $(X_n)_{n\in\mathbb{N}}$, où pour tout $n \in \mathbb{N}$, $X_n = (V_n, E_n)$. À l'étape $n \in \mathbb{N}^*$, un noeud $i \notin V_{n-1}$ est ajouté, duquel part une arête. On a donc $V_n = V_{n-1} \cup i$, le nombre de noeuds à cette étape est donc $n+|V_0|$, et le nombre d'arêtes $n+|E_0|$. L'autre extrémité de la nouvelle arête est tirée uniformément parmi V_{n-1} avec probabilité $p \in [0, 1[$, et est tirée avec probabilité $1-p$ suivant la loi d'attachement préférentiel proportionnel, définie de la façon suivante ; pour tout $j \in V_{n-1}$, la probabilité que i se connecte à j est :

$$P(E = j) = \frac{D_j^{X_{n-1}}}{2(|E_0| + n)}$$

La loi des degrés est à queue « d'autant moins lourde » que p est proche de 0, au sens où l'exposant de la loi de puissance augmente. Par construction, ce paramètre contrôle la proportion d'attachement préférentiel qu'on introduit à chaque ajout de noeud, et donc le phénomène de création de concentrateurs qui en découle. On retrouve le modèle LCD [6] pour $p = 0$. Bollobás and Riordan (2003) établit une revue des résultats obtenus sur le modèle LCD originel. La loi des degrés est sans échelle et le modèle vérifie la propriété de petit-monde grâce aux concentrateurs par lesquels peuvent passer beaucoup de plus courts chemins. En revanche son deuxième coefficient de regroupement est faible.

6. On note que dans Bollobás and Riordan (2004a); Bollobás et al. (2001); Bollobás and Riordan (2003), le modèle LCD considère le noeud i lui-même dans la création de la nouvelle arête, créant ainsi éventuellement des boucles. Ceci permet de décrire la distribution du graphe à l'étape n dans un cadre qui la simplifie. D'autre part, ces articles envisagent aussi le cas de la création de $m > 1$ arêtes sur le nouveau noeud.

FIGURE 2.3 – Réalisation d'un graphe dans le modèle d'Albert et Bárabasi, de taille $n = 300$, construit à partir d'un graphe X_0 de taille 10, qui est la réalisation d'un graphe d'Erdős-Rényi avec $p = 0.3$



On notera que l'hétérogénéité dans le réseau est amenée par un mécanisme déséquilibré qui force automatiquement la déviation du degré de quelques noeuds. Elle ne dépend donc pas d'un paramètre déterminé à l'avance et intrinsèque aux noeuds. Au contraire même, les noeuds en arrivant dans le graphe suivent en fait tous le même schéma : la distribution de leur arête dépend seulement du graphe préexistant, mais d'aucun paramètre individuel.

Une généralisation de ce modèle est possible en considérant d'autres attachements qui sont fonctions croissantes du degré autres que linéaires, voir Dommers et al. (2010) and Buckley and Osthus (2004).

Une autre sous-classe usuelle de modèles de croissance est celle des modèles à mécanisme de copie. À chaque itération, le noeud ajouté « copie » un noeud préexistant tiré uniformément, en se connectant à certains de ses voisins. Il est introduit dans Kleinberg et al. (1999) et s'inspire originellement d'autres aspects de la création de pages Web. Il a en fait rencontré beaucoup de succès dans la modélisation de l'évolution du génôme par duplication de gènes, voir par exemple Chung et al. (2003).

2.3.5 Conclusion

Ces modèles schématisent les mécanismes de formation des structures topologiques locales, de sorte à expliquer le plus simplement possible l'émergence de propriétés globales des réseaux réels. Ils sont le fruit d'une démarche *ad hoc* qui répond avec succès aux questions de la Physique.

Cependant, bien qu'ils possèdent une certaine hétérogénéité, leur schéma de construction est commun à tous les noeuds. Ils ne tiennent donc pas compte d'éventuelles caractéristiques individuelles influençant les interactions. Tout se passe comme si l'ensemble des paramètres individuels étaient remplacés par un petit nombre de paramètres moyens sur la population. Leur variabilité s'en trouve limitée et leur topologie plutôt rigide. Comme le fait remarquer le livre Kolaczyk (2009), cela n'interdit pas l'estimation de ces paramètres, qui présentent l'intérêt d'être interprétables dans les applications. En revanche ces modèles se révèlent généralement peu adaptés à une étude statistique de variabilité dont un des buts majeurs est de séparer au mieux, puis d'estimer, les effets individuels et les effets de masse. Bien que les Statistiques aient besoin de modèles dont les propriétés s'ajustent à la réalité, ils doivent en plus leur permettre de faire le trajet inverse de celui de la Physique : revenir du réseau global aux propriétés individuelles de chaque noeud.

Chapitre 3

Quelques modèles statistiques de graphes aléatoires hétérogènes

Ce chapitre a pour but de présenter des classes de modèles qui attaquent de front la modélisation d'une population hétérogène, notamment celle des modèles à variables latentes. Le principe de ces modèles est de tenir compte de variables caractéristiques pertinentes de chaque individu de la population, dites *latentes* ou *cachées* car inaccessibles dans les données, puis de considérer que les observations sont engendrées conditionellement à ces variables.

La Section 3.2 discute de manière générale de la modélisation de l'hétérogénéité en Statistiques, avec des exemples de modèles de réseau. Elle part de modèles où les variables décrivant l'hétérogénéité sont observées puis présente les modèles à variables latentes, où certaines caractéristiques individuelles ne sont pas observées. Les modèles de mélanges et les modèles à chaîne de Markov cachée (HMM) illustrent ces modèles de manière simple. La Section 3.3 présente la méthode classique d'inférence de ces modèles que constitue l'algorithme EM. Quelques variantes sont évoquées, dont la forme variationnelle qui pallie aux problèmes de modèles à variables cachées avec dépendances. Enfin, la Section 3.4.1 introduit la classe des modèles de graphes aléatoires qui sert de fil rouge aux travaux originaux de cette thèse, présentés aux chapitres 6 et 7.

3.1 Modèles statistiques de graphes aléatoires hétérogènes

3.1.1 Modélisation de l'hétérogénéité en Statistiques

Une des propriétés principales que l'on attend d'un modèle statistique est son ajustement aux données observées. Si par exemple on veut modéliser la taille adulte des individus d'une espèce animale asexuée, une loi normale s'y prête bien : elle

TABLE 3.1 – Modèles de régressions

Régression linéaire de la taille sur l'âge	$X_i \sim \mathcal{N}(a + bz_i, \sigma^2)$
covariable explicative	z_i âge de l'individu i
paramètres	a la taille moyenne à la naissance b la croissance par unité de temps
Régression logistique de l'interaction sur les distances et la sociabilité	$X_{ij} \sim \mathcal{B}\left(\frac{e^{\eta_i+\eta_j-\beta d_{ij}}}{1+e^{\eta_i+\eta_j-\beta d_{ij}}}\right)$
covariables explicatives	η_i, η_j sociabilités téléphoniques des individus i et j d_{ij} distance de i à j
paramètres	α effet constant β poids de la distance par rapport à la sociabilité téléphonique

modélise efficacement une variable centrée sur la moyenne caractéristique de l'espèce, à laquelle se superposent des fluctuations symétriques imputables à la variabilité génétique et environnementale de chaque individu. Cependant l'ajustement se détériore si l'on utilise ce même modèle sur tous les individus de l'espèce quel que soit leur stade de développement, ou si l'espèce est sexuée et que mâle et femelle sont significativement différents en taille. La population n'est alors plus homogène et la variabilité n'est plus due aux seules fluctuations, mais à des caractéristiques individuelles qu'il convient d'ajouter au modèle. Dans ces exemples simples, âge ou sexe sont généralement des covariables accessibles, et on pense en premier à une régression linéaire de la taille des individus sur leur âge et à une ANOVA de la taille sur le sexe des individus. Le modèle linéaire permet de prendre en compte une hétérogénéité linéaire dans la moyenne, la variance étant la même pour tous les individus.

Du côté des réseaux, le modèle d'Erdős-Rényi correspond de la même façon à une population homogène. Gilbert avait introduit le modèle $\mathcal{G}(n,p)$ en le motivant par une modélisation de réseaux téléphoniques. Mais ces réseaux sont très marqués par des caractéristiques individuelles, comme ce que nous appellerons « sociabilité téléphonique » (mesurable par le nombre, la durée des appels) et par la géographie. Pour les prendre en compte, on peut tout aussi bien mettre en oeuvre une régression, le cadre le plus agréable dans le cas des variables de Bernoulli étant le cadre logistique. Soit $(\eta_i)_{i\in[n]}$ des réels mesurant la sociabilité des individus, les valeurs négatives caractérisant les personnes allergiques au téléphone, et les

valeurs positives des personnes accrochées au téléphone. Soit $(d_{ij})_{i,j\in[n]}$ la matrice des distances entre les personnes, de sorte à pénaliser la probabilité d'appel par l'éloignement géographique, comme dans le modèle de Kleinberg. Soit $\alpha \in \mathbb{R}$ un effet constant, et $\beta \geq 0$ un paramètre qui pondère l'effet de la distance par rapport à la sociabilité. On suppose les variables $X = (X_{ij})_{i,j\in[n]}$ indépendantes, et on pose le modèle de graphe suivant :

$$\text{logodd}(X_{ij} = 1) = \alpha + \eta_i + \eta_j - \beta d_{ij} \tag{3.1}$$

ou de manière équivalente :

$$X_{ij} \sim \mathcal{B}\left(\frac{e^{\alpha+\eta_i+\eta_j-\beta d_{ij}}}{1+e^{\alpha+\eta_i+\eta_j-\beta d_{ij}}}\right)$$

Ainsi à η_i fixé, plus η_j est grand, plus i et j ont de chances d'être connectés. D'autre part plus leur distance d_{ij} est petite, et moins leur probabilité de connexion est réduite. Si $\eta_1 = \eta_2 = \cdots = \eta_n$ et $d_{12} = d_{13} = \cdots = d_{n-1,n}$ la population est à nouveau homogène et tout le monde vit à la même distance ; on retrouve alors le modèle d'Erdős-Rényi $\mathcal{G}\left(n, (1+e^{-(\alpha+2\eta_1-\beta d_{12})})^{-1}\right)$. La vraisemblance du modèle du graphe aléatoire X est pour tout graphe $x = (x_{ij})_{i,j\in[n]}$:

$$P_{\alpha,\beta}(X = x) = \frac{1}{\Lambda_\beta} \exp\left(\sum_{1\leq i<j\leq n} x_{ij}(\alpha + \eta_i + \eta_j - \beta d_{ij})\right)$$

avec Λ_β la constance de normalisation. L'inférence de ce modèle (limitée à l'estimation de β) est très simple et peut être réalisée par la méthode du maximum de vraisemblance. Ce modèle peut être vu comme une version non dirigée du modèle p_1 (Holland and Leinhardt, 1981).

L'apport de ces modèles que l'on souligne est qu'ils tiennent compte d'un certain type d'hétérogénéité en établissant un lien explicite entre un caractère individuel z_i et la loi de l'observation X_i associée (voir table 3.1).

Un aspect décevant du modèle de graphe hétérogène mentionné dans ce passage est que les arêtes sont mutuellement indépendantes. Bien que les probabilités de réalisations des arêtes soient liées, par exemple dans le cas des distances via l'inégalité triangulaire, les réalisations des arêtes elles-mêmes ne sont pas dépendantes au sens probabiliste usuel.

3.1.2 Modèles de graphes aléatoires exponentiels (ERGM) : Dépendances à la carte

Le modèle précédent 3.1 s'insère dans la classe des modèles de graphes aléatoires exponentiels (ERGM), appelés aussi modèles p^*. Il s'agit en fait des modèles de

graphes qui s'écrivent dans la famille exponentielle. Dans ce paragraphe on appelle configuration un ensemble d'arêtes, c'est-à-dire une partie C de $\mathcal{P}_n$. Les modèles ERGM sont ceux qui possèdent une loi de la forme suivante :

$$P_\theta(X = x) = \frac{1}{\Lambda_\theta} \exp\left(\sum_{C \subset \mathcal{P}_n} \theta_C g_C(x)\right) \tag{3.2}$$

avec $x = (x_{ij})_{i,j\in[n]}$ une matrice d'adjacence, identifiée au graphe (V_x, E_x), et :

- $g_C(x) = \prod_{\{i,j\}\in C\}} x_{ij}$ (ou de manière équivalente, $g_C(x) = \mathbb{1}_{C\subset E_x}$), qui dit si la configuration C est présente dans x, i.e. si les arêtes qu'elle contient sont présentes dans x.
- θ_C le coefficient associé à la configuration C. Si $\theta_C = 0$, les arêtes de la configuration sont mutuellement indépendantes conditionnellement aux arêtes $\mathcal{P}_n \backslash C$, donc la configuration en question se produira du seul fait du hasard. Au contraire, si $\theta_C > 0$ (respectivement $\theta_C < 0$) la configuration est encouragée (resp. pénalisée) et se produira plus souvent (resp. moins souvent) qu'au hasard : cela institue une dépendance entre les arêtes de la configuration. Chaque coefficient θ_C peut dépendre lui-même d'un ou plusieurs paramètres, et de covariables explicatives, comme dans l'exemple du paragraphe précédent.
- Λ_θ est la constance de normalisation de la loi.

Les modèles ERGM permettent ainsi de paramétrer directement les tendances du modèle à produire les configurations désirées. C'est un bouton de réglage au coeur de la topologie du réseau et de sa structure de dépendances : en associant un coefficient θ_C non nul aux configurations C contenant plusieurs arêtes, on crée des dépendances entre elles. On notera que toutes les spécifications des coefficients $(\theta_C)_{C\in\mathcal{P}_n}$ ne donnent pas nécessairement un modèle valide : les coefficients des configurations impliquant des arêtes communes sont liés et s'imposent des contraintes mutuelles. Par exemple, pour $k \in [n-1]$, on appelle k-étoile une configuration de k arêtes partageant un même noeud [1]. Si C et C' sont respectivement une k-étoile et une $(k+1)$-étoile telles que $C \subset C'$, on ne pourrait pas avoir $\theta_C = 0$ et $\theta_{C'} > 0$. La présence de C' dans x implique celle de C, i.e. la réalisation d' une $(k+1)$-étoile nécessite celle d'une k-étoile. Le théorème de Hammersley-Clifford (voir Besag, 1974) précise les conditions de validité d'un tel modèle.

Quand des coefficients non nuls sont associés à des configurations de tailles de plus en plus grandes, la structure de dépendance se complique de plus en plus. Pour se la représenter, on utilise souvent le *graphe de dépendance* $D = (V_D, E_D)$. Il est défini de cette façon : V_D est l'ensemble des variables aléatoires du modèle, ici $V_D = \{X_{ij};\ 1 \le i < j \le n\}$, identifié à $\mathcal{P}_n$; puis deux noeuds sont liés dans D si les

1. Une 1-étoile est une configuration singleton, i.e. ne contenant qu'une arête.

variables correspondantes sont dépendantes conditionellement à toutes les autres. Notons que pour toute configuration $C \subset \mathcal{P}_n$, θ_C est non nul si et seulement si C est une clique dans le graphe de dépendance D.

Dans le modèle du paragraphe précédent, les seules configurations spécifiées étaient les singletons $\{\{i,j\}\}$, pour tous $i \neq j$, de sorte que les arêtes étaient toutes mutuellement indépendantes. De manière équivalente, D était le graphe vide. Le coefficient θ_{ij} associé à l'arête $\{i,j\}$ était $\theta_{ij} = \alpha + \eta_i + \eta_j - \beta d_{ij}$; il incluait les covariables explicatives η_i, η_j et d_{ij} et un seul paramètre β.

Graphes aléatoires de Markov

Frank and Strauss (1986) se sont affranchis de cette hypothèse d'indépendances peu plausible en proposant une notion générale de dépendance markovienne dans les réseaux : pour tout $k \in [n-1]$, k arêtes sont interdépendantes si et seulement si elles partagent un même noeud, autrement dit, si leur ensemble forme une k-étoile. Frank and Strauss (1986) montre que les modèles vérifiant cette hypothèse constituent une sous-classe des ERGM, dont les seuls coefficients θ_C non nuls sont ceux associés aux configurations C qui sont des k-étoiles ou des triangles. Ils sont appelés modèles de graphes de Markov. Ils permettent notamment d'obtenir des modèles transitifs.

En associant des coefficients distincts aux différentes configurations, on crée de l'hétérogénéité : on donne ainsi des caractéristiques propres aux arêtes de la configuration. Cependant, il est illusoire de donner des coefficients distincts à toutes les configurations, puisqu'alors on obtiendrait un modèle surparamétré et probablement non identifiable. On peut faire une hypothèse d'homogénéité dans les classes de configurations isomorphiques, c'est-à-dire associer un paramètre à chaque motif.

Dans le cas des modèles de graphes de Markov, Frank and Strauss (1986) associe le même coefficient θ_1 à toutes les arêtes, θ_k à toutes les k-étoiles pour tout $k \in \{2, \ldots, n-1\}$, et θ_τ à tous les triangles. La loi du graphe aléatoire X est alors :

$$P_\theta(X = x) = \frac{1}{\Lambda_\theta} \exp\left(\theta_1 L(x) + \sum_{k=2}^{n-1} \theta_k S_k(x) + \theta_\tau T(x)\right)$$

avec $L(x)$ le nombre d'arêtes de x, pour tout $k \in \{2, \ldots, n-1\}$, $S_k(x)$ le nombre de k-étoiles dans x et $T(x)$ le nombre de triangles. Pour à nouveau éviter d'avoir trop de coefficients, on peut poser des contraintes sur les $(\theta_k)_{2 \leq k \leq n-1}$, en forçant les derniers à être nuls, ou afin d'obtenir des effets plus fins, en lui imposant une certaine progression dépendant d'un seul paramètre, comme le propose Snijders et al. (2006).

Ainsi, les paramètres $\theta_1, \ldots, \theta_{n-1}, \theta_\tau$ permettent de régler l'apparition des motifs tels que les triangles et les étoiles dans le réseau. Snijders et al. (2006) propose aussi d'ajouter des dépendances supplémentaires, c'est-à-dire des motifs plus

élaborés, en tenant compte des dépendances entre X_{ij} et X_{kl} où les noeuds i, j, k, l sont distincts, quand une arête joint une extrémité de l'une à une extrémité de l'autre, par exemple si l'arête $\{i, k\}$ est présente.

Caractéristiques de groupes

Bien que cela rende le modèle plus souple, attribuer un coefficient à chaque motif plutôt qu'à chaque configuration peut nuire à l'hétérogénéité du réseau, puisque le coefficient ne dépend alors que du motif et pas des individus impliqués dans configuration. On peut restaurer une certaine hétérogénéité sans aller dans l'excès du surparamétrage en considérant que la population est divisée en groupes, et que les coefficients ne dépendent que du motif et des groupes des noeuds impliqués dans la configuration. Par exemple, en reprenant le précédent exemple, dans un réseau téléphonique à une échelle internationale, un nouvel effet apparaît : celui des frontières, qui rompt la régularité géographique. Si on se limite à deux pays, on peut proposer le modèle suivant. La population totale est partitionnée entre les deux pays : $[n] = G_1 \sqcup G_2$. Soit $\gamma_1, \gamma_2 \geq 0$ l'effet additif constant favorisant les appels nationaux et $\gamma_{12} = \gamma_{21}$ l'effet soustractif constant pénalisant les appels internationaux [2]. Si $i \in G_k$ et $j \in G_l$, on note :

$$\theta_{ij}^{kl} = \begin{cases} \alpha + \gamma_1 + \eta_i + \eta_j - \beta d_{ij} \text{ si } k = l = 1 \\ \alpha + \gamma_2 + \eta_i + \eta_j - \beta d_{ij} \text{ si } k = l = 2 \\ \alpha - \gamma_{12} + \eta_i + \eta_j - \beta d_{ij} \text{ si } k \neq l \end{cases}$$

$$P_\theta(X = x) = \frac{1}{\Lambda_\theta} \exp\left(\sum_{\{i,j\} \in \mathcal{P}(G_1)} \theta_{ij}^{11} x_{ij} + \sum_{\{i,j\} \in \mathcal{P}(G_2)} \theta_{ij}^{22} x_{ij} + \sum_{(i,j) \in G_1 \times G_2} \theta_{ij}^{12} x_{ij} \right)$$

Inférence statistique dans ERGM

La famille exponentielle constitue souvent en Statistiques un cadre agréable pour l'inférence ; c'est aussi le cas par exemple dans les modèles à variables cachées avec l'algorithme EM. Dans le cas d'observations indépendantes, les estimateurs fondés sur le maximum de vraisemblance y ont une expression analytique explicite. Les ERGM indépendants sont ceux dont les seules configurations qui sont spécifiées sont des arêtes. Dans cette classe d'ERGM, le cadre est en fait celui du modèle linéaire généralisé, et l'inférence n'y pose aucun problème. En revanche, et le même problème se posera avec l'algorithme EM, l'ajout de dépendances complique beaucoup l'inférence, même dans le cas basique des modèles de Markov. La

2. Le modèle en l'état n'est pas identifiable, il faudrait ajouter des contraintes sur les effets constants pour en faire l'inférence.

log-vraisemblance d'un ERGM s'écrit, pour tout graphe x :

$$\mathcal{L}(x, \theta) = \theta^T g(x) - \psi(\theta)$$

où g est le vecteur contenant les fonctions g_H, θ le vecteur des paramètres θ_H pour toutes les configurations H possibles, et $\psi(\theta)$ est le logarithme de la constante de normalisation Λ_θ. La méthode du maximum de vraisemblance n'est pas applicable parce qu'elle requiert la minimisation de la constante de normalisation Λ_θ, qui présente une somme de $2^{\binom{n}{2}}$ termes. Le lecteur peut trouver un panorama sur les ERGM et beaucoup de références sur les aspects de leur inférence sont dans Kolaczyk (2009); Robins et al. (2007).

Deux classes de méthodes approchées d'inférence ont principalement été développées. Les méthodes les plus fondées théoriquement et les plus employées actuellement se déclinent en de multiples versions, toutes à base de MCMC. L'une de ces méthodes, détaillée dans Hunter and Handcock (2006), consiste à maximiser une approximation de $\mathcal{L}(x, \theta) - \mathcal{L}(x, \theta^{(0)})$. Il s'agit du logarithme du rapport de vraisemblance entre un paramètre θ quelconque et le paramètre fixé $\theta^{(0)}$, qui peut s'écrire de la façon suivante grâce aux propriétés algébriques de la famille exponentielle :

$$\begin{aligned}\mathcal{L}(x, \theta) - \mathcal{L}(x, \theta^{(0)}) &= (\theta - \theta^{(0)})^T g(x) - (\psi(\theta) - \psi(\theta^{(0)})) \\ &= (\theta - \theta^{(0)})^T g(x) - \log\left(\mathbb{E}_{\theta^{(0)}}\left(\left(\theta - \theta^{(0)}\right)^T g(X)\right)\right)\end{aligned}$$

L'approximation est obtenue en estimant le terme d'espérance ci-dessous par MCMC. Une autre méthode courante (Snijders, 2002) utilise l'algorithme de Robins-Monro, décrit par Kolaczyk (2009) comme une version stochastique de l'algorithme de Newton-Raphson.

L'autre méthode est fondée sur une pseudo-vraisemblance, et utilise une propriété du modèle exponentiel pour se ramener à l'inférence d'un modèle logistique. Elle a été introduite dans Besag (1975) et adaptée aux ERGM dans Strauss and Ikeda (1990). On note $X^{\setminus\{k,l\}}$ l'ensemble des variables $X = (X_{ij})_{i,j\in[n]}$ privé de X_{kl} (et X_{lk}). Alors on peut écrire :

$$\operatorname{logodd}_\theta(X_{kl} = 1 \mid X^{\setminus\{k,l\}} = x^{\setminus\{k,l\}}) = \sum_{\substack{C \subset \mathcal{P}_n \\ \{i,j\} \in C}} \theta_C \delta_C^{ij}(x) \tag{3.3}$$

avec $\delta_C^{ij}(x)$ la différence entre $g_C(x)$ quand $x_{ij} = 1$ et quand $x_{ij} = 0$, comme si les $(X_{ij})_{i,j\in[n]}$ suivaient un modèle de régression logistique sur les δ_C^{ij}, en oubliant qu'on conditionne chacune par $X^{\setminus\{i,j\}}$. Puis on réalise l'inférence comme s'il s'agissait véritablement d'un tel modèle, c'est-à-dire en oubliant les dépendances des X_{ij}, et on maximise alors la log-pseudo-vraisemblance :

$$\sum_{\{i,j\} \in \mathcal{P}_n} \log P_\theta(X_{ij} = 1 \mid X^{\setminus\{i,j\}} = x^{\setminus\{i,j\}})$$

Wasserman and Robins (2005) explique le calcul des statistiques de différence δ_C^{ij}, qui n'est pas complexe algorithmiquement. Les auteurs de Robins et al. (2007) rappellent que (3.3) n'est pas un modèle logistique, du fait que les X_{ij} ne sont pas indépendantes, et que cela conduit à d'éventuels biais et à une variance sous-estimée. Enfin, les limites de cette méthode sont théoriques, puisque le comportement de la pseudo-vraisemblance n'est encore pas complètement compris (Besag, 2001; Snijders, 2002). Une fois encore, on pourra observer un parallèle avec l'algorithme EM. L'idée de ne pas tenir compte des dépendances dans une approximation de la vraisemblance sera mise à profit dans la méthode d'inférence approchée VEM. Elle se substitue à l'algorithme EM, en particulier dans le cas des réseaux qui ont une structure de dépendance très intriquée.

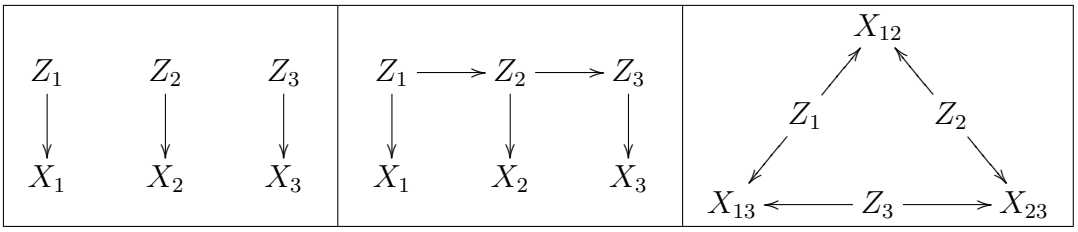
3.2 Modèles statistiques à structure latente

3.2.1 Modélisation de données cachées

La question principale à laquelle les modèles statistiques du début de ce chapitre répondent est celle de l'existence d'un lien significatif entre la variable d'étude et une ou plusieurs covariables explicatives accessibles. Par exemple dans l'exemple de la régression de la taille sur l'âge, l'âge était observé, et on pouvait tester l'influence de cette variable sur la taille. Cependant dans beaucoup de phénomènes complexes, l'utilisateur peut ne pas avoir accès à toutes les variables qu'il veut inclure dans le modèle, et qui pourraient expliquer l'hétérogénéité des observations. Ces covariables peuvent être physiquement inaccessibles, ou leur acquisition peut être trop coûteuse par exemple. On considère alors l'ajout de variables aléatoires dans le modèle, dites *cachées* ou *latentes*. On distingue dans le modèle trois statuts de variables :

1. **Variables observées** Ce sont les données d'étude ; dans cette thèse, c'est un graphe $X = (X_{ij})_{i,j\in[n]}$.
2. **Covariables** Ce sont des variables explicatives fixes auxquelles l'utilisateur a accès. Elles peuvent décrire des effets individuels des noeuds (dans le modèle (3.1), ce sont les sociabilités téléphoniques) ou des effets de paires d'individus (dans (3.1), ce sont les distances).
3. **Variables *cachées* ou *latentes*** Ce sont les variables aléatoires du modèle dont la réalisation n'est pas connue. Elles ne pourront donc pas être utilisées directement dans l'inférence. Dans cette thèse, chaque individu $i \in [n]$ sera caractérisé par une variable latente Z_i. On notera aussi $Z = (Z_i)_{i\in[n]}$). Les observations seront toujours supposées indépendantes conditionnellement aux variables latentes Z.

FIGURE 3.1 – Graphes de dépendances dans les modèles de mélanges (gauche), des HMM (centre), des modèles de réseaux à variables latentes (droite) pour trois individus.



Exemple. Citons un exemple intéressant dans les réseaux. Le réseau belge des appels téléphoniques se caractérise à l'échelle nationale par deux grandes communautés très connectées (Blondel et al., 2010), qui demeurent totalement inexplicables avec le modèle de réseau téléphonique (3.1). En effet, les covariables individuelles et d'éloignement géographique ne suffisent pas à expliquer la séparation de la Belgique en deux. Pour expliquer cette structure hétérogène, on ajoute alors dans le modèle une variable latente Z_i pour chaque noeud i, qui vaut 1 ou -1. Soit $\gamma \geq 0$ le paramètre qui pondère l'effet de la structure. Le modèle devient alors :

$$\text{logodd}(X_{ij} = 1) = \eta_i + \eta_j - \beta d_{ij} + \gamma Z_i Z_j$$

Le produit $Z_i Z_j$ avantage ou pénalise la probabilité de connexion, selon qu'il vaut 1 ou -1. *A posteriori*, on s'aperçoit que cette variable latente correspond en fait à la langue majoritaire (français ou flamand) des noeuds, constitués par les municipalités belges. La structure du réseau téléphonique reflète ainsi l'hétérogénéité linguistique belge.

Ajout de dépendances dans les modèles à variables latentes

Dans les modèles de mélanges, les variables latentes sont supposées indépendantes. En conséquence, par l'indépendance des observations conditionnellement à ces variables latentes, les observations sont elles-même indépendantes.

Cependant, en général, même en dehors des réseaux, les observations peuvent ne pas l'être. Si on observe par exemple un signal cohérent en fonction du temps, il existe une dépendance entre les observations successives. Les modèles à chaîne de Markov cachée (HMM) sont adaptés à ce type d'observations. Ils consistent à supposer que Z est une chaîne de Markov (voir aussi le graphe de dépendance en Figure 3.1). De même, si on observe une image cohérente en deux ou trois dimensions, il existe des dépendances entre pixels voisins, et on peut par exemple adopter un modèle à champ de Markov caché.

Dans un modèle de réseau à variables latentes, les observations sont des interactions entre paires d'individus. Chaque observation implique donc deux variables latentes (voir le graphe de dépendance à la Figure 3.1). Par conséquent, même si les variables latentes sont indépendantes, de la dépendance se crée entre les observations, du fait de la nature même de la structure de réseau. Dans cette thèse, les modèles de réseaux à variables latentes seront toujours supposées indépendantes. Cela revient à supposer que les interactions sont entièrement décrites par le graphe observé, et qu'il n'y a pas de graphe latent décrivant des interactions supplémentaires.

Questions statistiques.

Du fait même de l'absence d'observation des variables latentes, potentiellement explicatives de l'hétérogénéité, ces modèles ont vocation à répondre à de nouvelles questions statistiques. Elles portent sur la possibilité de pouvoir caractériser, plus ou moins précisément, les variables latentes et leur loi. À partir des observations, peut-on, de manière consistente :

1. tester la présence effective d'hétérogénéité, et donc l'existence même des variables latentes ?
2. trouver le meilleur modèle de variables latentes adapté aux données parmi une collection ?
3. estimer les variables latentes [3] et leur loi ?
4. sinon, au moins distinguer dans la population hétérogène des sous-groupes d'individus plus homogènes ?

Le cas des modèles de mélanges finis gaussiens est un cas simple typique de modèle à variables cachées, où l'on sait répondre à toutes ces questions. Au chapitre 6, on montrera qu'on sait les résoudre dans le modèle de réseau à variable latentes Stochastic Blockmodel sous certaines hypothèses. La première question, qui devrait d'abord être traitée lorsqu'on a des données à analyser, n'est pas traitée à notre connaissance dans la littérature pour le cas des réseaux hétérogènes. Elle est traitée dans une annexe ajoutée au Chapitre 6 dans le cas simple du Stochastic Blockmodel. Le modèle KerNet, introduit en 3.4.2, fournit un cadre que nous pensons adapté pour créer un test répondant à cette question dans les réseaux de manière assez générale. Le Chapitre 7, où un algorithme estime le nombre de sous-groupes homogènes dans ce modèle, constitue une première étape vers un test du type de la question 1.

3. Les variables latentes ne sont pas identifiables généralement. Leur éventuelle estimation ne pourra se faire qu'à une certaine relation d'équivalence près.

3.2.2 Modèles de mélanges : le cas indépendant

Nous développons le cas indépendant et son inférence classique, puis nous verrons comment les dépendances complexifient le problème de l'inférence de ces modèles à variables latentes.

De manière générale, les modèles de mélanges forment la classe des modèles à variables latentes indépendantes, et donc à observations indépendantes. Dans le cas de mélanges finis, on suppose que l'hétérogénéité peut être décrite par seulement $Q \in \mathbb{N}^*$ profils distincts d'individus. La population est donc séparée en Q sous-groupes aux caractéristiques différentes et inconnues. Les individus avec leurs caractéristiques sont supposés indépendants. Pour tout $q \in [Q]$, on note $\alpha_q = P(Z_i = q)$ la probabilité pour chaque individu d'appartenir au groupe étiqueté q. On note $\alpha = (\alpha_1, \ldots, \alpha_Q)$. Alors la loi des variables latentes est un n-échantillon multinomial :

$$(Z_i)_{i\in[n]} \text{ i.i.d. } \sim \mathcal{M}(1; \alpha)$$

Comme dans les modèles linéaires ou de régression, on spécifie la loi de la variable étudiée en fonction des caractéristiques de l'individu, i.e. en fonction de son groupe. Ainsi tous les individus d'un même groupe ont la même loi conditionnellement à leur groupe. On définit f_q comme la loi de X_i sachant que l'individu i appartient au groupe étiqueté q :

$$(X_i \mid Z_i = q) \sim f_q$$

Puisque la loi de chaque X_i est définie conditionnellement à son groupe Z_i, et que les Z_i sont mutuellement indépendants, les X_i sont aussi mutuellement indépendants. La loi des X_i est une combinaison convexe des lois associées à chaque groupe avec pour coefficients les $(\alpha_q)_{q\in[Q]}$:

$$X_i \sim \sum_{q=1}^{Q} \alpha_q f_q$$

Mélanges gaussiens Par souci de simplicité, les lois conditionnelles des observations appartiennent généralement à une même famille paramétrique. Dans le cas d'observations X continues, la famille gaussienne est très couramment utilisée pour le contrôle des paramètres qu'elle offre, le formulaire disponible et sa relative flexibilité permettant généralement un bon ajustement. Ainsi dans le cas des mélanges gaussiens, chaque groupe est caractérisé par les paramètres de moyenne μ_q et de matrice de variance-covariance Σ_q :

$$(X_i \mid Z_i = q) \sim \mathcal{N}(\mu_q, \Sigma_q) \text{ et } X_i \sim \sum_{q=1}^{Q} \alpha_q \mathcal{N}(\mu_q, \Sigma_q)$$

Identifiabilité Les groupes sont définis à permutation près de leur numéro. En classification, on ne cherche donc pas à retrouver le numéro de groupe (arbitraire) de chaque individu, mais les classes d'équivalence de la relation « être dans le même groupe ». Même en utilisant ces classes d'équivalence, certaines familles de lois souffrent de non-identifiabilité de leurs paramètres, par exemple les mélanges de variables de Bernoulli.

Vraisemblances du modèle

La méthode d'inférence paramétrique la plus classique est celle du maximum de vraisemblance. En présence de variables cachées, la log-vraisemblance peut avoir deux définitions :

- la log-vraisemblance de tout le modèle, en tenant compte de toutes les variables $(X_i)_{i\in[n]}$ et $(Z_i)_{i\in[n]}$, appelée vraisemblance complète et définie comme suit :
$$\mathcal{L}(x, z, \theta) = \log\left(P_\theta(X = x, Z = x)\right)$$
- la vraisemblance des seules données observées, i.e. celle qui ne tient compte que des variables $(X_i)_{i\in[n]}$; elle est appelée vraisemblance observée ou incomplète et définie ainsi :
$$\mathcal{L}(x, \theta) = \log\left(P_\theta(X = x)\right)$$

Les deux notions de vraisemblance sont liées par la relation suivante :

$$\mathcal{L}(x, z, \theta) = \mathcal{L}(x, \theta) + \log\left(P(Z = z \mid X = x)\right) \tag{3.4}$$

La vraisemblance complète dépend des variables latentes, et ne peut donc pas être utilisée directement. Pour l'inférence du modèle, on ne peut *a priori* se reposer que sur la vraisemblance observée, qui s'écrit aussi :

$$\mathcal{L}(x, \theta) = \sum_{(z_1,\dots,z_n)\in[Q]^n} P(X = x, Z = z)$$

Or son optimisation s'avère très coûteuse du point de vue combinatoire, car elle nécessite l'exploration de l'espace des configurations possibles pour les labels, qui est de taille Q^n. En pratique elle se révèle impossible à appliquer au-delà de quelques dizaines d'observations.

3.3 Inférence des modèles à variables latentes : l'algorithme EM

L'algorithme Expectation-Maximization, dit EM, a déclenché un grand engouement pour le problème des données manquantes dans les années 1980 et 1990, et

est toujours couramment utilisé pour sa mise en oeuvre simple et sa formulation explicite des estimateurs notamment dans la famille exponentielle. Il consiste à maximiser itérativement l'espérance de la log-vraisemblance complète conditionnellement aux observations, au lieu de la vraisemblance observée. Cette stratégie présentée dans l'article fondateur Dempster et al. (1977) présente l'intérêt de faire croître indirectement la vraisemblance observée à chaque itération. On note :

$$Q_{\theta,\theta'}(X) = \mathbb{E}_Z^{\theta'}(\mathcal{L}(X, Z, \theta) \mid X) \tag{3.5}$$

$$H_{\theta,\theta'}(X) = -\mathbb{E}_Z^{\theta'}(\log P(Z \mid X, \theta) \mid X) \tag{3.6}$$

où $\mathbb{E}_Z^{\theta'}$ est l'espérance contre la loi des variables Z avec pour paramètre θ'. $Q_{\theta,\theta'}(X)$ est donc l'espérance sachant X de la log-vraisemblance en θ, contre la loi des variables Z qui a pour paramètre θ'. Par ailleurs H est un terme positif appelé entropie emprunté à la Physique. En prenant l'espérance sur les variables Z pour un paramètre $\theta' \in \Theta$ de l'équation 3.4, on obtient :

$$\mathcal{L}(X, \theta) = Q_{\theta,\theta'}(X) + H_{\theta,\theta'}(X, Z)$$

Algorithme 3.1. $EM(\theta^{(0)})$

- *Initialisation du paramètre à la valeur* $\theta^{(0)}$.
- *TANT QUE l'algorithme n'a pas atteint la condition de terminaison, FAIRE* $k = k + 1$ *puis :*
  Étape E : Calcul de l'espérance de la log-vraisemblance complète sachant X *et avec le paramètre* $\theta^{(k)}$ *:* $Q_{\theta,\theta^{(k)}}(X)$
  Étape M : Maximisation par rapport à θ *de l'espérance et mise à jour du paramètre :* $\theta^{(k+1)} = \arg\max_\theta Q_{\theta,\theta^{(k)}}(X)$
- *Retourner* $\theta^{(k+1)}$

L'algorithme EM dépend d'une valeur d'initialisation en fonction de laquelle son résultat peut beaucoup varier, puisqu'il n'assure pas de croître vers un maximum global. Il convient donc de faire tourner l'algorithme plusieurs fois pour obtenir de meilleurs résultats. Quelques variantes présentées plus loin essaient de pallier ce problème en insérant une dose d'exploration aléatoire dans l'algorithme.

Notation. Dans le cas où la loi des Z est à support fini, identifié à $[Q]$ où $Q \in \mathbb{N}^*$ (cas des modèles de mélanges finis, des HMM à nombre d'états fini, du Stochastic Blockmodel par exemple), l'étape E revient au calcul des probabilités d'appartenance de chaque individu à chaque classe, dites *a posteriori* car conditionnelles aux observations. On note :

$$\tau_{iq}^{(k)} = P_{\theta^{(k)}}(Z_i = q \mid X)$$

Les conditions de terminaison sont variées : cela peut être un nombre déterminé à l'avance d'itérations, un seuil de différence entre $\theta^{(k)}$ et $\theta^{(k+1)}$: dès qu'il est assez petit, on arrête l'algorithme, ou encore un seuil de différence sur les $(t_{iq}^{(k)})_{i,q}$, etc.

3.3.1 Convergence de l'algorithme EM

On peut montrer que l'accroissement de $Q_{\theta,\theta^{(k)}}(X)$ entraîne un accroissement de la vraisemblance observée, qui en plus est strict si celui de $Q_{\theta,\theta^{(k)}}(X)$ l'est. En notant $(\theta^{(k)})_{i\in\mathbb{N}}$ la suite des paramètres produite par l'algorithme EM si on le faisait tourner à l'infini et θ un paramètre quelconque, on a pour tout $i \in \mathbb{N}$:

$$\mathcal{L}(X, \theta^{(k+1)}) - \mathcal{L}(X, \theta^{(k)}) \geq Q_{\theta,\theta^{(k)}} - Q_{\theta^{(k)},\theta^{(k)}} \tag{3.7}$$

Cette inégalité équivaut en fait à l'inégalité de Gibbs qui s'applique à l'entropie : $H_{\theta,\theta^{(k)}}(X) \geq H_{\theta^{(k)},\theta^{(k)}}(X)$ (voir Dempster et al. (1977); Little and Rubin (1987) pour une preuve). Elle-même provient de l'inégalité de Jensen qui s'applique par concavité du logarithme. En conséquence la vraisemblance observée augmente à chaque itération de l'algorithme EM. Mais en général cette croissance ne garantit rien sur la convergence de la suite $(\theta^{(k)})_{i\in\mathbb{N}}$, encore moins vers le lieu d'un maximum global ni même local de la vraisemblance.

Dempster et al. (1977) montre sous certaines conditions de régularité du modèle que si la suite $(\theta^{(k)})_{i\in\mathbb{N}}$ converge, le point limite est un point critique de la vraisemblance, qui peut être un maximum global, mais aussi un maximum local ou un point selle. La vitesse de convergence est régie par les valeurs propres de la différentielle au point limite de l'opérateur M associé à l'algorithme, c'est à dire :

$$M(\theta) = \arg\max_{\theta' \in \Theta} Q_{\theta',\theta}$$

Si elles sont toutes plus petites que 1 en valeur absolue, l'opérateur est contractant et la convergence est rapide à condition que la valeur d'initialisation $\theta^{(0)}$ soit suffisamment proche du point limite, d'où la nécessité d'essayer plusieurs initialisations. Dempster et al. (1977) écrit la différentielle en fonction de la différence de la hessienne de H et de la hessienne de la log-vraisemblance au point limite. Il interprète ces matrices comme des informations de Fisher, l'une étant l'information observée sur θ (donnée par X), l'autre étant l'information cachée (donnée par Z). Ainsi, plus l'information cachée est faible par rapport à l'information observée plus la convergence sera rapide. L'idée de certaines variantes de l'algorithme mettent à profit cette idée en essayant de traiter moins d'information cachée à chaque itération. Pour plus de développements sur la convergence et ses conditions, on pourra consulter Wu (1983) qui donne des résultats revus, corrigés et augmentés.

3.3.2 Variantes de l'algorithme EM

L'article Dempster et al. (1977) a beaucoup stimulé la recherche dans le domaine des données manquantes et de nombreuses variantes ont ainsi fait suite à l'algorithme originel.

Si la maximisation de l'étape M pose problème, par exemple si elle n'a pas de solution analytique, notamment en dehors de la famille exponentielle :

GEM pour Generalized EM. Il s'agit en fait d'un algorithme aussi proposé dans Dempster et al. (1977). À l'étape M, GEM cherche simplement un θ qui fait augmenter $Q_{\theta,\theta^{(k)}}(X)$, sans nécessairement la maximiser. En effet la maximisation n'est pas nécessaire, puisque d'après l'inégalité 3.7, un simple accroissement de $Q_{\theta,\theta^{(k)}}(X)$ fait croître la vraisemblance observée.

Les autres variantes concernent l'étape E.

CEM pour Classification EM. Introduit dans Celeux and Govaert (1992), il s'agit d'une modification de l'algorithme EM qui fournit une classification en plus d'un estimateur des paramètres. Après l'étape E, une étape C construit une classification par la règle du MAP (Maximum A Posteriori) basée sur les probabilités *a posteriori* d'appartenance aux classes calculées à l'étape E. À l'étape M, la log-vraisemblance complète prise en les Z estimés à l'étape C, soit $\mathcal{L}(X, \widehat{Z}, \theta)$, est directement maximisée par rapport au paramètre θ. Dans certains cas, la maximisation de cette quantité peut être plus simple que celle de $Q_{\theta,\theta^{(k)}}(X)$.

SEM pour Stochastic EM. Développé par exemple dans Broniatowski et al. (1983); Celeux and Diebolt (1985), cette variante tire aléatoirement une classification en utilisant les probabilités *a posteriori* $(\tau_{ij})_{q\in[Q],i\in[n]}$ d'appartenance aux classes, i.e. pour tout $i \in [n]$, on tire indépendamment le label de i :

$$\text{Pour tout } i,\ \widehat{Z}_i^{(k)} \sim \mathcal{M}\left(1, (\tau_{i1}^{(k)}, \ldots, \tau_{iQ}^{(k)})\right)$$

On note que l'étape E peut être court-circuitée s'il est difficile de déterminer la loi conditionnelle des Z sachant X, mais qu'il est possible de simuler directement sous cette loi. L'idée de SEM est de perturber la trajectoire de l'algorithme EM classique de sorte à explorer de manière plus exhaustive l'ensemble des classifications et par là, celui des paramètres. Les cas où $\theta^{(k)}$ est piégé dans un maximum local se produisent ainsi moins souvent. Un inconvénient collatéral est que la propriété fondamentale d'accroissement de la log-vraisemblance au fil des itérations n'est plus vérifiée.

SAEM pour Simulated Annealing EM, introduit dans Celeux and Diebolt (1992). L'estimateur $\theta^{(k)}$ est une combinaison convexe des estimateurs de EM et de SEM. Le coefficient de la combinaison tend vers zéro au fil des itérations, de

sorte à aller du SEM aux premières itérations vers le EM. Le coefficient joue le rôle de la température dans le recuit simulé, d'où le nom de l'algorithme.

MCEM pour Monte Carlo EM, voir Wei and Tanner (1990). Il s'avère utile dans les cas où l'espérance conditionnelle $Q_{\theta,\theta^{(k)}}(X)$ est difficile à déterminer à l'étape E, mais qu'il est possible de simuler sous la loi des Z sachant X avec le paramètre courant. $Q_{\theta,\theta^{(k)}}(X)$ est alors estimée via la méthode de Monte-Carlo par chaînes de Markov (MCMC). Si on tire un m-échantillon $(Z^{(1)}, \ldots, Z^{(m)})$ i.i.d. sous cette loi, l'espérance est estimée par :

$$\widehat{Q}_{\theta,\theta^{(k)}}(X) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(X, Z^{(k)}, \theta)$$

Le cas où $m = 1$ revient à SEM, ce pourquoi cet algorithme peut être interprété comme une généralisation de SEM.

VEM pour Variational EM. Dans les cas où la loi conditionnelle des variables cachées Z sachant les observations X est compliquée, même EM peut encore être trop coûteux algorithmiquement. VEM est une reformulation équivalente de l'algorithme EM, qui se prête à des approximations efficaces et moins coûteuses dans ces cas. Il est décrit dans le paragraphe suivant et illustré dans le cas du Stochastic Blockmodel à la Section 4.1.1.

3.3.3 Formulation variationnelle de l'algorithme EM

L'étape E de l'algorithme EM peut être vue comme la résolution d'un problème variationnel, c'est-à-dire d'un problème d'optimisation sur une classe de distributions. Le critère à optimiser naît de la décomposition de la vraisemblance observée en utilisant une distribution intermédiaire quelconque pour les variables cachées, notée R :

Proposition 3.1.

$$\mathcal{L}(x, \theta) = \mathcal{F}_{x,\theta}(R) + D_{KL}(R||P_\theta(Z|X = x))$$

où $\mathcal{F}_{x,\theta}(R) = \mathbb{E}_R\left(\log\left(\frac{P_\theta(X=x,Z)}{R}\right)\right)$ *et* D_{KL} *est la divergence de Kullback-Leibler. De plus :*

$$\mathcal{F}_{x,\theta}(R) \leq \mathcal{L}(x, \theta)$$

avec égalité si et seulement si R *est la loi conditionnelle de* Z *sachant* X*.*

Démonstration. Comme $\mathcal{L}(x, \theta)$ ne dépend pas des Z, on peut d'abord écrire :

$$\begin{aligned}
\mathcal{L}(x,\theta) &= \mathbb{E}_R\left(\mathcal{L}(x,\theta)\right) \\
&= \mathbb{E}_R\left(\log\left(\frac{P_\theta(X=x,Z)}{R}\frac{R}{P_\theta(Z \mid X=x)}\right)\right) \\
&= \mathbb{E}_R\left(\log\left(\frac{P_\theta(X=x,Z)}{R}\right)\right) + \mathbb{E}_R\left(\frac{R}{P_\theta(Z \mid X=x)}\right) \\
&= \mathcal{F}_{x,\theta}(R) + D_{KL}(R||P_\theta(Z|X=x))
\end{aligned}$$

Les dernières affirmations viennent des propriétés de la divergence de Kullback-Leibler : elle est positive d'après l'inégalité de Jensen et par stricte concavité du logarithme, elle est nulle si et seulement si les deux distributions sont égales. □

$\mathcal{F}_{x,\theta}$ est appelée énergie libre, terme aussi emprunté à la Physique. D'après la proposition, en maximisant l'énergie libre sur l'ensemble des distributions possibles pour Z, on retrouve la vraisemblance observée. Ce problème de maximisation a en fait pour unique solution la loi conditionnelle des Z sachant X. Dans l'algorithme VEM, on réalise cette maximisation sur un ensemble de distributions possibles pour Z, qu'on note $\mathcal{Q}$. L'algorithme VEM est alors le suivant :

Algorithme 3.2. *$VEM(\theta^{(0)})$*

- *Initialisation du paramètre à la valeur $\theta^{(0)}$.*
- *TANT QUE l'algorithme n'a pas atteint la condition de terminaison, FAIRE $k = k + 1$ puis :*

  Étape VE : Maximiser $\mathcal{F}_{X,\theta^{(k)}}(R)$ par rapport à R et mise à jour de la distribution courante des Z :

$$R^{(k+1)} = \arg\max_{R\in\mathcal{Q}} \mathcal{F}_{X,\theta^{(k)}}(R) \tag{3.8}$$

  Étape VM : Maximisation par rapport à θ de $\mathcal{F}_{X,\theta^{(k)}}(R^{(k+1)})$ et mise à jour du paramètre :

$$\theta^{(k+1)} = \arg\min_{\theta} \mathcal{F}_{X,\theta^{(k)}}(R^{(k+1)}) \tag{3.9}$$

- *Retourner $\theta^{(k+1)}$*

Structure de dépendance et approximation variationnelle à l'étape VE

Dans les modèles de mélanges finis, les variables cachées sont indépendantes conditionnellement aux observations, et leur loi conditionnelle sachant les observations est factorisable. Pour tout $z = (z_1, \ldots, z_n) \in [Q]^n$ elle s'écrit :

$$P(Z = z \mid X) = \prod_{i\in[n]} P(Z_i = z_i \mid X)$$

Le problème de maximisation (3.8) dans l'espace $\mathcal{Q}$ de toutes les distributions possibles des variables cachées est résolu analytiquement au moins dans la famille exponentielle. Dans les HMM, cette loi est encore factorisable de proche en proche du fait de la structure de dépendance markovienne unidimensionnelle, et s'écrit pour tout $z = (z_1, \ldots, z_n) \in [Q]^n$:

$$P(Z = z \mid X) = P(Z_1 = z_1 \mid X) \prod_{i=2}^{n} P(Z_i = z_i \mid X, Z_{i-1} = z_{i-1})$$

Moyennant une procédure supplémentaire appelée forward-backward à l'étape E (Devijver, 1985), on peut obtenir les probabilités *a posteriori* d'appartenance aux classes de manière analytique. Plus généralement, si la loi conditionnelle des Z sachant X a une structure de dépendance en arbre (non nécessairement filaire comme ici) elle pourrait être factorisée de cette façon, et un algorithme de type somme-produit (Wainwright and Jordan, 2008) permettrait encore un calcul analytique de la loi conditionnelle. Dans ces situations, aucune approximation n'est faite ; l'algorithme VEM est exact et est équivalent à l'algorithme EM.

Cependant il ressort qu'une structure de dépendance plus intriquée interdirait une telle factorisation. On serait alors ramenés à la case départ, puisque la complexité algorithmique se rapprocherait de celle de la maximisation directe de la vraisemblance observée. C'est le cas par exemple des réseaux bayesiens, des champs de Markov, ou en ce qui nous concerne, des modèles de réseaux avec des dépendances.

En fait la puissance de cette formulation se révèle en jouant sur l'ensemble $\mathcal{Q}$ des distributions de Z sur lequel on maximise l'énergie libre. En maximisant non dans l'ensemble de toutes les distributions, mais dans un sous-ensemble de distributions où l'on saura résoudre le problème de maximisation, on fait une approximation, dite approximation variationnelle. Ici, il s'agit des distributions factorisables $\mathcal{Q}_{\text{fact}}$. On cherche donc R_0 tel que :

$$\mathcal{F}_{X,\theta}(R_0) = \inf_{R \in \mathcal{Q}_{\text{fact}}} \mathcal{F}_{X,\theta}(R)$$

Ainsi cet algorithme est exact sur les structures d'arbres, puisque la vraie loi conditionnelle des variables latentes sachant les observations est dans $\mathcal{Q}_{\text{fact}}$, et devient un algorithme approché sur les structures de dépendance plus complexes. De ce fait, on diminue la complexité algorithmique. La divergence de Kullback-Leibler augmente aussi, mais ce n'est pas nécessairement un problème, puisqu'on n'attend pas que $\mathcal{F}_{X,\theta}(R)$ approche bien la log-vraisemblance, mais plutôt que le lieu de son maximum en θ approche bien celui de la log-vraisemblance. L'idée de cette approximation vient de la physique, où elle est plus connue sous le nom d'approximation de champ moyen. On peut consulter Wainwright and Jordan (2008); Jaakkola (2000) pour plus de détails.

Consistence. Bien qu'elle donne souvent des résultats satisfaisants en pratique, cette méthode n'est pas encore bien comprise, et il existe peu de résultats généraux sur sa consistence à l'heure actuelle. Il existe des résultats dans des cas particuliers : Alain Celisse et al. (2012) prouvent sa consistence dans le cas du Stochastic Blockmodel. Il existe aussi des résultats de consistence concernant sa version bayesienne [4], par exemple pour les mélanges gaussiens (Wang and Titterington, 2006), et d'autres modèles appartenant à la famille exponentielle (Wang and Titterington, 2004a).

D'autre part, les estimateurs fournis par l'algorithme ne sont pas consistents dans certains modèles (Wang and Titterington, 2004b). Une classe de modèles dans laquelle elle semble toujours consistente est celle où l'approximation variationnelle est asymptotiquement exacte. C'est justement le cas dans le Stochastic Blockmodel, ce que montrent Alain Celisse et al. (2012).

3.4 Cadre général des modèles de graphes aléatoires à espace latent

3.4.1 Présentation du cadre général

Cette section introduit le cadre commun aux travaux originaux de la thèse présentés aux Chapitres 6 et 7. Il est inspiré de la famille générale de modèles de graphes aléatoires hétérogènes proposée par Bollobás et al. (2007), que nous réinterprétons ici du point de vue des modèles à variables latentes. En résumé, Bollobás et al. (2007) proposent dans leur modèle d'attribuer des caractéristiques individuelles aléatoires $(Z_i)_{i\in[n]}$ aux noeuds, mutuellement indépendantes. Puis les arêtes $(X_{ij})_{i,j\in[n]}$ du graphe sont tirées conditionnellement à Z_i et Z_j, de sorte que les caractéristiques individuelles introduisent de l'hétérogénéité dans leurs interactions. Dans toute cette thèse, les variables Z seront supposées latentes [5], tandis que le graphe X constituera notre seule donnée observée.

Définition 3.1. *On appelle espace latent de graphe aléatoire un triplet $(\mathcal{S}, \nu, \kappa)$ où $\mathcal{S}$ est un espace métrique séparable, ν une mesure de probabilité sur $\mathcal{S}$, et :*

$$\kappa : \mathcal{S} \times \mathcal{S} \to [0, 1]$$

une application symétrique (i.e. vérifiant pour tout $z, z' \in \mathcal{S}$, $\kappa(z, z') = \kappa(z', z)$) appelée fonction de connexion (ou noyau dans Bollobás et al., 2007).

4. Cette version est décrite dans Beal (2003) par exemple.

5. Nous n'empruntons en fait à cet article que le modèle. Ses auteurs l'ont conçu originellement pour étudier des transitions de phases pour la composante géante du type de celles que subit le modèle Erdős-Rényi (voir la Section 2.2) mais dans des modèles de graphes aléatoires hétérogènes suffisamment généraux.

L'espace latent de graphe correspond à l'espace des caractéristiques individuelles des noeuds, muni de leur loi ν, ainsi que de la fonction κ qui établit le lien explicite entre les caractéristiques et la loi des arêtes. Sa symétrie est due au caractère non orienté des graphes considérés. Plus précisément :

Définition 3.2. *On appelle graphe aléatoire à espace latent de taille n, tout graphe aléatoire de taille n tel qu'il existe :*

- *un espace latent de graphe aléatoire* $(\mathcal{S}, \nu, \kappa)$,
- *une suite de variables aléatoires indépendantes et identiquement distribuées* $Z = (Z_i)_{i\in[n]}$ *à valeurs dans* $\mathcal{S}$ *et de loi* ν,

et dont la loi vérifie :

- *l'indépendance conditionnelle des arêtes sachant Z :*

$$P(X = x \mid Z) = \prod_{1\leq i<j\leq n} P(X_{ij} = x_{ij} \mid Z) \text{ pour tout graphe } x \text{ à } n \text{ noeuds}$$

- *pour tout graphe x à n noeuds et pour tout* $i, j \in [n]$ *la loi conditionnelle de* X_{ij} *sachant Z vérifie :*

$$P(X_{ij} = 1 \mid Z) = \kappa(Z_i, Z_j).$$

Remarque. Contrairement aux HMM et autres modèles à champ de Markov caché, les variables Z sont indépendantes, comme dans les modèles de mélanges. Mais contrairement aux modèles de mélanges, les observations $(X_{ij})_{i,j\in[n]}$ ne sont pas indépendantes. Du fait de la nature même des observations qui sont des interactions, la dépendance se crée, non pas au niveau des variables latentes Z, mais au niveau des observations $(X_{ij})_{i,j\in[n]}$, la loi conditionnelle de chaque arête X_{ij} impliquant les caractéristiques latentes Z_i et Z_j de deux individus. Ajoutons de plus qu'elle n'implique que ces deux individus, et aucun autre. Par conséquent, la structure de dépendance est similaire à celle des modèles de Markov (voir le paragraphe 3.1.2) : des arêtes sont dépendantes si et seulement si elles partagent un même noeud.

3.4.2 Exemples

Nous présentons brièvement les deux modèles qui font l'objet de résultats originaux dans cette thèse, d'une part le Stochastic Blockmodel, qui équivaut au cas de l'espace $\mathcal{S}$ fini. Il est commenté en 4.1.1 et les travaux originaux sont exposés au Chapitre 6. Le deuxième modèle, appelé KerNet, est un cas particulier de modèle de graphes aléatoires à positions latentes, où $\mathcal{S}$ est l'espace continu $\mathbb{R}^d$. Un algorithme de type K-linkage est étudié théoriquement dans ce modèle au Chapitre 7.

Stochastic Blockmodel. Chaque noeud appartient aléatoirement à une classe ou a une couleur aléatoire parmi $Q \in \mathbb{N}^*$ possibles : $\mathcal{S} = \{1, \dots, Q\}$. La probabilité de connexion entre deux noeuds ne dépend que de leurs couleurs. Pour tout $q \in [Q]$, on note α_q la probabilité que tout noeud soit dans la classe q, soit $\nu(\{q\}) = \alpha_q = P(Z_1 = q)$. On note $\alpha = (\alpha_1, \dots, \alpha_Q)$. On note aussi π_{qr} la probabilité de connexion entre un noeud de la classe q et un noeud de la classe r, et $\pi = (\pi_{qr})_{q,r\in[Q]}$, i.e. $\kappa(z, z') = \pi_{z,z'}$ pour tous $z, z' \in \mathcal{S}$.

Variables latentes. $(Z_i)_{i\in[n]}$ i.i.d. $\sim \mathcal{M}(1, \alpha)$

Graphe observé. $(X_{ij})_{i,j\in[n]}$ indépendantes conditionellement à Z et pour tout $i, j \in [n]$, $P(X_{ij} = 1 \mid Z) = \pi_{Z_i,Z_j}$.

KerNet. Chaque noeud a une position latente dans l'espace euclidien $\mathcal{S} = \mathbb{R}^d$ selon une loi ν à densité f par rapport à la mesure de Lebesgue sur cet espace. Soit $k : \mathbb{R}^d \to [0, 1]$ un noyau à support compact isotrope, et h_n un réel strictement positif. On pose $\kappa(z, z') = k\left((z - z')/h_n\right)$ pour tous $z, z' \in \mathbb{R}^d$. Puis :

Variables latentes. $(Z_i)_{i\in[n]}$ i.i.d. f

Graphe observé. $(X_{ij})_{i,j\in[n]}$ indépendantes conditionellement à Z et pour tout $i, j \in [n]$, $P(X_{ij} = 1 \mid Z) = k\left(\frac{Z_i - Z_j}{h_n}\right)$.

Le Chapitre 7 porte sur un tel modèle ; d'autres hypothèses y seront faites par ailleurs. Le Latent Cluster Position Model (LPCM) introduit dans (Handcock et al., 2007) est une variante où f est un mélange gaussien et la fonction de connexion κ est une fonction logistique décroissante de la distance.

Les graphes aléatoires géométriques (Penrose, 2003) en sont aussi une variante, mais où $\kappa(x, y) = \mathbb{1}_{\{\|x-y\|\leq 1\}}$. Cela implique que le graphe n'est pas aléatoire conditionnellement aux positions : deux noeuds se connectent de manière déterministe quand ils sont à distance inférieure à h_n l'un de l'autre.

Asymptotique particulière. On notera qu'on se placera parfois dans une asymptotique particulière, où l'espace latent de graphe $(\mathcal{S}_n, \nu_n, \kappa_n)$ dépend lui-même de n. C'est par exemple le cas quand on veut obtenir un modèle à faible densité ou au moins dans la densité tend vers zéro quand n tend vers l'infini. On se placera dans une telle asymptotique dans le Chapitre 6 en 6.5 dans le cas du Stochastic Blockmodel, par exemple quand on suppose le nombre de classes $Q_n \to +\infty$, ou $\pi^n_{qr} \to 0$ pour avoir un modèle dont la densité de graphe tend vers zéro. Dans le modèle KerNet au Chapitre 7, κ dépend de n via le paramètre d'échelle h_n, mais $\mathcal{S}$ ne dépend pas de n.

3.4.3 Justification théorique du cadre général : théorème de représentation par un modèle de graphon

La famille de modèles vérifiant la définition 3.2 semble raisonnable d'un point de vue de modélisation, mais se voit aussi justifiée par un théorème de représentation de Lovász and Szegedy (2006). Nous donnons d'abord le modèle de graphon, qui est une brique de ce théorème, et en fait un cas particulier de la précédente famille. Soit $\mathcal{S} = [0,1]$, $\nu = \mathcal{U}([0,1])$ la loi uniforme sur $[0,1]$ et $\kappa : [0,1]^2 \to [0,1]$ une application symétrique. On note $\mathcal{H}_{n,\kappa}$ le modèle de graphe défini de la manière suivante :

Variables latentes. $(Z_i)_{i\in[n]}$ i.i.d. $\sim \mathcal{U}([0,1])$

Graphe. $(X_{ij})_{i,j\in[n]}$ indépendantes conditionellement à Z et pour tout $i,j \in [n]$, $P(X_{ij} = 1 \mid Z) = \kappa(Z_i, Z_j)$.

Alors d'après le théorème 2.7 de Lovász and Szegedy (2006), tout modèle de graphe de la précédente famille de taille n peut être réécrit comme un modèle de graphon $\mathcal{H}_{n,\kappa}$ pour une certaine fonction κ. Plus généralement :

Théorème 3.1. *(Lovász and Szegedy, 2006)*
Pour toute suite de modèles $(\mathbf{G}_n)$ de graphes aléatoires de taille n vérifiant les conditions suivantes, il existe une fonction symétrique $\kappa : [0,1]^2 \to [0,1]$ telle que le modèle a la même loi que $\mathcal{H}_{n,\kappa}$:

1. *La loi de $\mathbf{G}_n$ est invariante par permutation des numéros des noeuds.*
2. *La loi du sous-graphe induit par l'ensemble de noeuds $[n-1]$ est égale à celle de $\mathbf{G}_{n-1}$.*
3. *Pour tout $1 < i < n$, les sous-graphes de $\mathbf{G}_n$ induits par les ensembles de noeuds $\{1,\dots,i\}$ et $\{i+1,\dots,n\}$ sont indépendants.*

Par exemple pour tout SBM de taille n et de paramètres α et π, une fonction κ telle que $\mathcal{H}_{n,\kappa}$ lui corresponde peut être trouvée en prenant une fonction constante par pavés de $[0,1]^2$. On décompose $[0,1]^2$ en entier en Q^2 pavés disjoints, chaque pavé correspondant à un couple de classes. Pour le pavé correspondant aux classes q et r, les longueurs des côtés des pavés sont α_q et α_r, et la valeur de la fonction κ sur ce pavé est égale à π_{qr}.

Cependant en général ce théorème reste théorique, et ne fournit bien sûr pas la forme explicite de $\kappa : [0,1]^2 \to [0,1]$, qui peut éventuellement être très complexe et sans aucune régularité. Dans le modèle KerNet, il existe ainsi théoriquement une telle fonction quand $h_n = h$ est fixé et ne dépend pas de n. Il est en revanche plus compliqué d'exhiber cette fonction. D'autre part, le cadre asymptotique où l'espace latent de graphe dépend de n ne satisfait plus les hypothèses du théorème de représentation.

Chapitre 4

Partitionnement des noeuds de graphes aléatoires

Le partitionnement de données ou *clustering*, tout comme l'hétérogénéité, sont des notions subjectives, qui n'ont pas de définition générale autre qu'heuristique ; le lecteur peut consulter l'article de Von Luxburg and Ben-David (2005) qui essaie de tracer une feuille de route vers une théorie unifiée. De manière générale, le clustering désigne les méthodes qui ont pour but de construire des groupes d'individus tels que les éléments d'un même groupe soient les plus semblables possibles, et ceux issus de groupes distincts soient les plus différents possibles. Dans le cadre classique hors graphes, le clustering est effectué sur des observations individuelles ; les graphes forment donc un cadre tout à fait avantageux pour cette tâche, puisqu'il décrivent en plus les interactions entre individus. Les techniques de clustering utilisent d'ailleurs parfois les graphes (Von Luxburg, 2007) pour faire du clustering sur des observations individuelles.

Dans la première partie de ce chapitre, on donnera quelques exemples de modèles auxquels le phénomène de clustering est inhérent, de par leur construction impliquant un modèle de mélange paramétrique fini. Dans ce cas, on met en oeuvre une classification non supervisée. Dans la seconde partie, on évoquera des algorithmes de clustering généraux. Elles se démarquent des méthodes développées dans les mélanges au sens où elles n'essaient pas de retrouver des groupes définis par un modèle, mais des groupes qui satisfont une heuristique, par exemple en optimisant un critère *ad hoc* mesurant les qualités de la partition en groupes. On conclura sur un compromis de ces deux parties pour introduire l'algorithme du chapitre 7.

4.1 Clustering fondé sur un modèle de mélange paramétrique fini

Dans un modèle de mélange fini, chaque individu i appartient à une classe aléatoire latente. Le critère de similarité est automatiquement défini par l'appartenance à une même composante du mélange, et les méthodes, qu'on veut consistentes, cherchent à classer les noeuds selon ce critère à partir du graphe observé X.

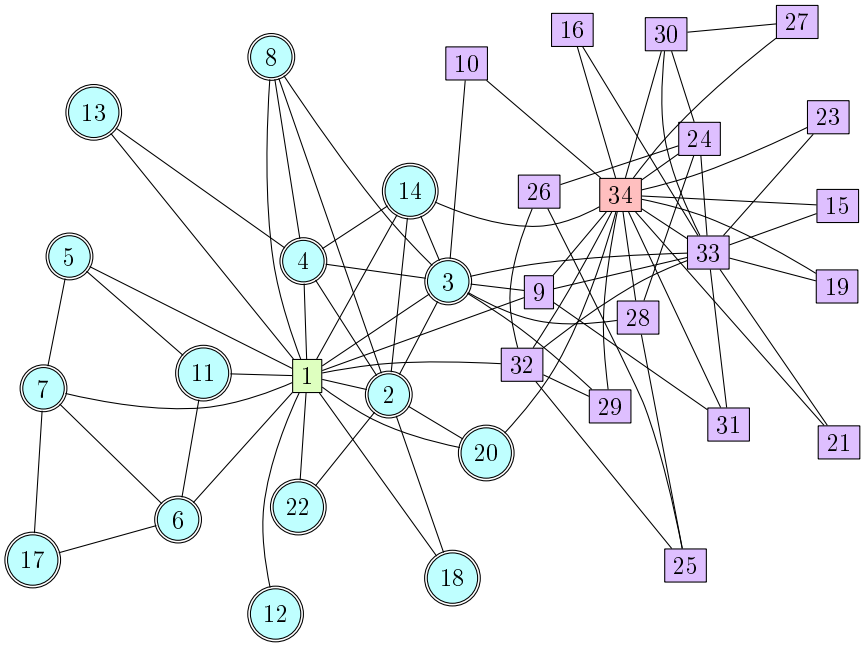
4.1.1 Stochastic Blockmodel (SBM)

Il est défini en 3.4.2. Ce modèle se fonde sur le concept d'équivalence structurelle défini par Lorrain and White (1971). Les auteurs y disent que « deux individus a et b d'une population P sont structurellement équivalents si pour tout élément x de la population, a et b interagissent tous deux avec x de la même façon ». Autrement dit, deux individus sont structurellement équivalents s'ils jouent le même rôle social dans la population. Cela conduit à formuler l'hypothèse selon laquelle la population est formée de Q blocs ou classes, dans lesquels les individus ont le même profil socio-relationnel. Par exemple, le célèbre exemple du Zachary Karaté Club (Zachary, 1977) est un réseau qui illustre bien cette idée. Il est traité dans de nombreux articles (Leger et al., 2013; Bickel and Chen, 2009; Girvan and Newman, 2002; Newman and Girvan, 2004...) et est devenu un benchmark classique dans la littérature. Il peut se décomposer en 4 classes d'équivalence structurelle (Leger et al., 2013) visibles en couleurs distinctes sur la Figure 4.1. On observe une classe composée d'un noeud vert (le président du club), une autre d'un noeud rouge (le professeur de karaté), et une pour la communauté qui gravite autour de chacune de ces deux personnes : en bleu pour celle du président, en mauve pour celle du professeur.

Picard et al. (2009) montre un autre exemple d'application très convaincant, celui des réseaux de régulation de protéines. Deux protéines sont structurellement équivalentes si elles jouent le même rôle dans le mécanisme de régulation. On peut associer un graphe à la matrice de connectivité π, en mettant une arête si deux classes q et r de protéines interagissent, i.e. si $\pi_{qr} > 0$. Comme les protéines d'une même classe peuvent interagir, c'est un graphe éventuellement à boucles. C'est une représentation de la structure du réseau qui résume les interactions globales entre les classes de protéines. On peut alors reconnaître dans la topologie du réseau des schémas universels de régulation (*feedforward loops*...) qui ont été identifiés en théorie du contrôle.

Le Stochastic Blockmodel est capable de restituer des topologies très variées, comme le montre le tableau 4.1. En particulier, les communautés sont prises en

FIGURE 4.1 – Réseau social du Zachary Karaté Club



compte, ainsi que les étoiles [1], ou les structures hiérarchiques.

Remarque sur la dénomination de modèle de mélange. SBM n'est pas à proprement parler un modèle de mélange, bien qu'on le qualifie souvent comme tel par abus de langage. Dans ce modèle les observations (en l'occurrence les arêtes) suivent effectivement une loi de mélange, mais ne sont pas indépendantes comme dans un modèle de mélange, du fait de la structure de réseau [2]. La loi de chaque observation, ici des arêtes, dépend conditionnellement des classes de ses deux extrémités, alors que dans les modèles de mélanges par exemple, la loi de chaque observation dépend d'une seule classe, qui lui est propre. Ceci posera d'ailleurs des problèmes majeurs d'inférence, que la version classique de l'algorithme EM ne résoud pas.

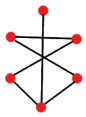
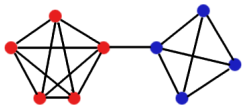
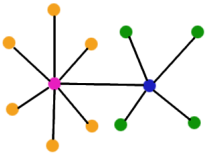
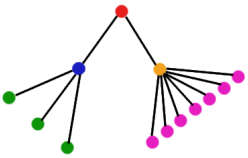
Problèmes de l'inférence

La log-vraisemblance complète du SBM à Q classes est donnée par :

1. Nous les appelons étoiles et pas concentrateurs (ou hubs), car nous avons réservé en 2.3.1 cette terminologie au cas où les noeuds de très haut degrés étaient responsables de l'émergence d'une loi des degrés sans échelle. Or la loi des degrés de SBM a une queue exponentielle et n'est pas sans échelle.

2. Voir aussi les commentaires à propos du cadre général de ce modèle en 3.3.3.

TABLE 4.1 – Exemples de SBM

$Q = 1$	$Q = 2$	$Q = 4$	$Q = 5$
			
$\pi = 0.4$	$\pi = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$	$\pi = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	$\pi = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$

$$\mathcal{L}(X, Z) = \sum_{i\in[n]} \sum_{q\in[Q]} Z_{iq} \log(\alpha_q) + \sum_{1\le i<j\le n} \sum_{q,r\in[Q]} Z_{iq} Z_{jr} \log\left(\pi_{qr}^{X_{ij}} (1 - \pi_{qr})^{1-X_{ij}}\right) \tag{4.1}$$

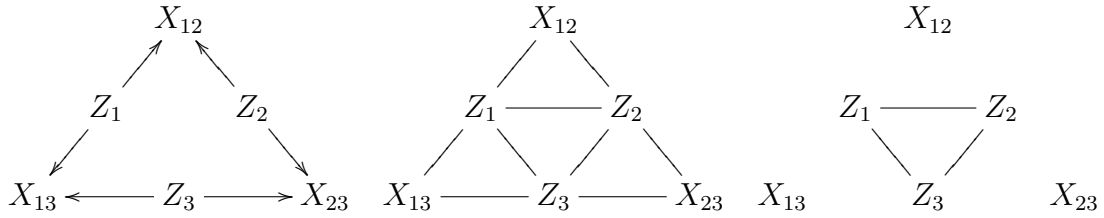
Comme dans la plupart des modèles à variables cachées, la méthode du maximum de vraisemblance n'est pas applicable en l'état. Snijders and Nowicki (1997) en particulier le remarque, et ramène toute la complexité du problème dans le cas $Q = 2$ au calcul des nombres $F_k(l, m)$ de partitions de $[n]$ en deux sous-ensembles à k et $n - k$ éléments, ayant respectivement l et m arêtes, pour tous k, l, m possibles. Ces partitions représentent les classifications possibles des noeuds. Le nombre d'opérations requises pour ce calcul est exponentiel en n, et cette méthode ne permet de traiter que de très petits graphes.

L'article envisage alors d'utiliser l'algorithme EM basique, mais montre qu'on se heurte au même calcul. Dans le cas d'une structure de dépendance intriquée des variables cachées Z sachant les observations X, l'algorithme EM n'apporte en effet aucun gain. Contrairement au cas des modèles de mélanges classiques et aux HMM, la loi conditionnelle $P(Z|X)$ n'est pas factorisable et est même la pire possible, puisque le graphe de dépendance des Z sachant X est une clique, comme l'illustre la Figure 4.2 dans le cas $n = 3$. C'est précisément dans ces cas que l'utilisation de l'approximation de champ moyen dans la formulation variationnelle de l'EM fournit des estimateurs accessibles et de bonne qualité. Nous la développons brièvement.

Application de l'approximation variationnelle

L'algorithme VEM est appliqué au Stochastic Blockmodel dans Daudin et al. (2008). Nous en donnons le principal résultat en reprenant les notations de l'algorithme EM et de la formulation variationnelle (voir 3.3 et 3.3.3 respectivement).

FIGURE 4.2 – Graphe de dépendance du SBM pour $n = 3$ (gauche), graphe moral (centre), graphe de dépendance de Z sachant X (droite).



Rappelons que la consistence de la méthode est montrée dans Alain Celisse et al. (2012).

Proposition 4.1. *(Daudin et al., 2008) Pour des paramètres (α, π) donnés, les estimateurs $(\widehat{\tau}_{iq})_{q\in[Q],i\in[n]}$ de la loi conditionnelle de Z sachant X, fournis par la maximisation de l'énergie libre $\mathcal{F}_{x,\alpha,\pi}(\tau)$, vérifient l'équation de point fixe suivante :*

$$\widehat{\tau}_{iq} \propto \alpha_q \prod_{\substack{1\leq j\neq n \\ j\neq i}} \prod_{1\leq r\leq Q} \left[\pi_{qr}^{X_{ij}}(1-\pi_{qr})^{1-X_{ij}}\right]^{\widehat{\tau}_{iq}}$$

Un algorithme de point fixe fournit ensuite les estimateurs. Rappelons que l'approximation de champ moyen consiste à trouver la loi qui optimise l'énergie libre dans l'ensemble des lois factorisables, ce qu'on peut observer dans la relation ci-dessus. Notons aussi que l'optimum prend en compte toutes les arêtes partant du noeud i, et pas seulement ses voisins dans le graphe.

Discussion sur les méthodes d'inférence

Snijders and Nowicki (1997) proposent une méthode bayesienne avec échantilonnage de Gibbs dans le cas $Q = 2$, généralisée dans Nowicki and Snijders (2001). Les auteurs font remarquer que grâce au grand nombre d'observations ($n(n-1)/2$ arêtes) pour un nombre n d'individus, l'influence du prior sur les estimateurs et les lois *a posteriori* décroît très rapidement. Cette remarque n'est pas propre à la méthode bayesienne et peut être généralisée à l'inférence dans les réseaux : le grand nombre d'observations permet une concentration très rapide des estimateurs. Cela expliquera notamment pourquoi une loi marginale du graphe aussi simple que les degrés concentre très vite et permet de réaliser toute l'inférence du SBM, ce qui est l'objet du Chapitre 6.

La principale difficulté de l'inférence dans ce type de modèle est la classification : si celle-ci est connue, l'estimation des paramètres est ensuite très basique. La méthode bayesienne et la méthode variationnelle présentent l'intérêt que classifi-

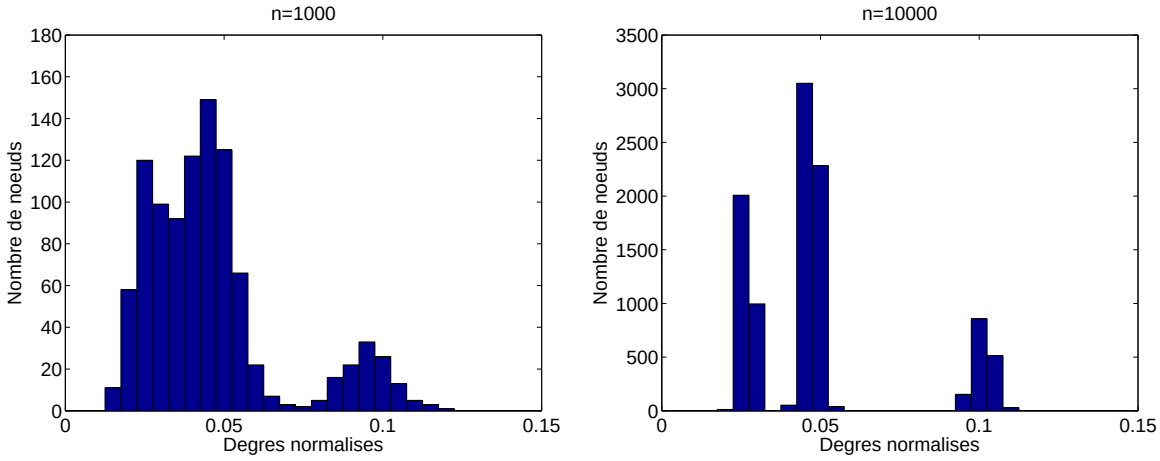
cation [3] et estimation se servent l'une et l'autre : en particulier l'estimation permet de mieux classer. Ce sont des méthodes qui seront donc très efficaces, même dans de petits graphes, et c'est d'ailleurs seulement là qu'elles pourront être utilisées, vu leur complexité algorithmique. Néanmoins l'estimation n'est pas nécessaire pour obtenir une classification consistente. Il est tout à fait possible de proposer un algorithme de classification consistent, et alors l'estimation suit sans autre condition. C'est cette démarche qui est faite dans le Chapitre 6. Rohe et al. (2010) propose aussi une méthode directe de classification consistente. Elle est très originale, car basée sur une variante du spectral clustering qui est un algorithme *a priori* étranger à ce type de réseau.

Concentration des degrés

Le Chapitre 6 montre un algorithme très simple et consistent sous certains hypothèses dans le SBM. Son avantage majeur est d'être de complexité pseudo-linéaire en nombre de noeuds n. Par comparaison, la méthode variationnelle a une complexité quadratique et le spectral clustering une complexité cubique.

La Figure 4.3 illustre la propriété de concentration des degrés dans ce modèle. Sous l'hypothèse que les moyennes des degrés normalisés conditionnelles aux groupes sont différentes, on peut retrouver les classes des noeuds en cherchant les plus grands écarts dans la distribution empirique des degrés.

FIGURE 4.3 – Histogrammes des degrés normalisés pour un SBM à $Q = 3$ classes pour $n = 1000$ et $n = 10000$ noeuds.



La conséquence est qu'asymptotiquement, la structure cachée se révèle presque

3. Estimation de la loi conditionnelle des Z sachant X à proprement parler dans ces méthodes.

d'elle-même dans ce modèle. L'algorithme du Chapitre 6 est simple, rapide, et consistent. Il donne un point de vue sur le modèle qui explique pourquoi l'inférence s'y passe finalement bien malgré sa structure de dépendance intriquée. Sa faible robustesse aux degrés très éloignés de leur moyenne peut aussi être améliorée moyennant quelques arrangements qui augmenteraient peu sa complexité. On pourrait par exemple utiliser l'histogramme des degrés et y détecter les pics, ou utiliser la somme cumulée des degrés et y détecter les plateaux qui sont le signe d'un écart entre groupes. Le gain en robustesse vient du fait qu'on ne se repose plus sur chaque noeud, mais sur des groupes de noeuds, constitués par les « bins » de l'histogramme.

Variantes du SBM

Une façon générale d'obtenir des clusters moins rigides et de relâcher la stricte équivalence structurelle en leur sein. L'Overlapping Stochastic Blockmodel par exemple, autorise l'appartenance à plusieurs classes. Il permet même le fait de n'appartenir à aucune classe, ce qui permet de tenir compte de noeuds trop atypiques pour entrer dans la classification. Latouche et al. (2011) propose une méthode variationnelle pour en faire l'inférence. Une autre variante, appelée Mixed Membership SBM, permet de modéliser le fait qu'un même individu peut jouer différents rôles sociaux au lieu d'un seul. Pour chaque noeud i, le vecteur α^i des probabilités d'appartenance à une classe est tiré aléatoirement dans une loi de Dirichlet. Puis pour chaque paire $\{i, j\}$ de noeuds, la classe de i (respectivement j) est retirée dans une multinomiale de paramètre α_i (respectivement α_j). Ici, on peut considérer que l'espace latent est le simplexe de $\mathbb{R}^Q$, qui est donc un espace continu.

4.1.2 Modèle Latent Position Cluster (LPCM)

L'idée selon laquelle un réseau serait régi par les distances sociales entre individus est née les années 1970 (voir par exemple McFarland and Brown, 1973), parallèlement au concept d'équivalence structurelle. Le concept d'espace social latent en découle naturellement ; c'est une représentation visuelle de la structure sociale sous-jacente de la population, des positions Z que les individus y tiennent et des distances [4] les séparant. Breiger et al. (1975); Faust and Wasserman (1994); Hoff et al. (2002) ont mis justement à profit la méthode de positionnement multidimensionnel (MDS, pour MultiDimensional Scaling) pour attribuer des positions aux individus dans un espace à partir de la matrice de leurs distances deux à deux. Le concept d'espace social continu et métrique a deux intérêts fondamentaux :

4. On notera que les distances ne sont pas nécessairement géographiques, ce sont plus généralement des différences au sein du milieu social.

1. Relâcher la rigidité des profils sociaux du SBM. Au lieu que les individus soient répartis en classes de stricte équivalence structurelle, chaque individu i a une position Z_i dans un espace continu qui correspond à ses caractéristiques individuelles. Ceci n'empêche pas de construire des groupes homogènes, et autorise simplement plus d'hétérogénéité en leur sein.
2. Modéliser de manière naturelle la transitivité, via l'inégalité triangulaire de la distance qui munit l'espace, ainsi que l'homophilie. Cette dernière propriété est illustrée par l'expression « qui se ressemble s'assemble », et désigne le principe sociologique selon lequel plus deux personnes ont des caractéristiques sociales proches, plus grande est leur probabilité de se connecter dans le réseau.

Le modèle Latent Position Cluster (LPCM), introduit par Handcock et al. (2007), a le but clairement affiché d'apporter un troisième intérêt à cette approche, celui du clustering. Les auteurs proposent un modèle de graphes à positions latentes dans l'espace continu et euclidien $\mathbb{R}^d$, qui s'incrit dans la famille de modèles 3.2 présentée en 3.4.1). Les variables latentes $(Z_i)_{i\in[n]}$ y suivent un modèle de mélange gaussien sphérique d-dimensionnel fini. De fait, le clustering est défini de manière inhérente au modèle, puisque les noeuds sont répartis en groupes homogènes selon la composante gaussienne qui engendre leur position dans l'espace latent. Une des tâches d'inférence sera notamment de retrouver la composante gaussienne de chaque noeud par une classification.

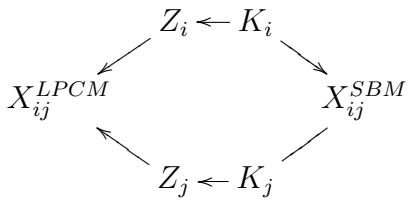
Variables latentes. On note $Q \in \mathbb{N}^*$ le nombre de composantes du mélange (classes), puis pour tout $q \in [Q]$, α_q est la probabilité d'appartenir au groupe q, μ_q la position sociale moyenne du groupe q, et σ_q^2 sa variance :

$$Z_i \text{ i.i.d. } \sim \sum_{q\in[Q]} \alpha_q \mathcal{N}_d\left(\mu_q, \sigma_q^2 I_d\right)$$

Tout se passe comme s'il y avait deux niveaux de variables latentes : un premier niveau identique au SBM originel où les classes des noeuds $(K_i)_{i\in[n]}$ sont tirées indépendamment au hasard parmi Q possibles, puis un second niveau où les variables $(Z_i)_{i\in[n]}$ sont tirées indépendamment conditionnellement à la composante obtenue au premier niveau (voir le graphe de dépendance 4.4). On ajoute donc une couche latente au SBM afin de produire plus d'hétérogénéité individuelle : sachant que le noeud i est dans la classe q, la loi de Z_i est le Dirac δ_q dans le SBM, contre la loi normale $\mathcal{N}_d(\mu_q, \sigma_q^2 I_d)$ dans le LPCM.

Graphe observé. Les variables $(X_{ij})_{i,j\in[n]}$ sont mutuellement indépendantes conditionnellement à Z. D'autre part pour effectivement inclure la transitivité et l'homophilie, la loi de l'arête X_{ij} conditionnellement aux variables latentes

FIGURE 4.4 – Graphes de dépendance du SBM et du LPCM



Z doit tenir compte de la distance entre les positions Z_i et Z_j dans l'espace euclidien $\mathbb{R}^d$. Handcock et al. (2007), à l'instar de Hoff et al. (2002), proposent une régression logistique sur la distance, avec éventuellement des covariables explicatives $(c_{ij})_{i,j\in[n]}$ observées sur les arêtes. Ce modèle est proche de l'exemple introductif sur les réseaux téléphoniques donné en 3.1.1, où les distances sont fixes et connues.

$$\eta_{ij} = \text{logodd}_\theta(X_{ij} = 1 \mid Z_i, Z_j) = \beta_0 c_{ij} - \beta_1 \|Z_i - Z_j\|$$

avec $\beta_1 \geq 0$, de sorte à pénaliser la probabilité des liens par la distance sociale.

Inférence du LPCM

Handcock et al. (2007) propose deux stratégies d'inférence. La première se fonde sur celle de Hoff et al. (2002) et s'effectue en deux étapes principales utilisant la méthode du maximum de vraisemblance et le MDS. Elle effectue une estimation des positions à isométrie près (le modèle est invariant par ces transformations) puis classe les noeuds et estime les paramètres en se basant sur l'estimation des paramètres. La seconde stratégie est totalement bayesienne.

Maximum de vraisemblance et MDS. La démarche est la suivante :

1. Estimation des positions des noeuds $(Z_i)_{i\in[n]}$ à isométrie près dans l'espace latent et du paramètre β de la régression logistique :
   (a) Estimation des distances $(\|Z_i - Z_j\|)_{i,j\in[n]}$ entre les positions des noeuds par maximum de vraisemblance
   (b) Détermination à isométrie près des positions approchées $(\widehat{Z}_i)_{i\in[n]}$ par MDS [5] à partir des distances estimées

5. MDS permet d'approcher les positions à isométrie près dans l'espace latent. Sa version

2. Algorithme EM pour le modèle de mélange gaussien, avec pour variables observées $(\widehat{Z}_i)_{i\in[n]}$ et pour variables cachées $(K_i)_{i\in[n]}$.

Estimer les positions revient à considérer les positions Z comme des paramètres du modèle. La log-vraisemblance conditionnelle du modèle sachant $Z = z$ est :

$$\mathcal{L}(X, Z = z) = \sum_{1\leq i<j\leq n} \eta_{ij} X_{ij} - \log(1 + e^{\eta_{ij}})$$

où $\eta_{ij} = \beta_0 c_{ij} - \beta_1 \|z_i - z_j\|$ pour tout $i, j \in [n]$. Mais elle n'est pas directement utilisable car elle n'a pas de propriété intéressante permettant sa maximisation par rapport à $z = (z_1, \ldots, z_n) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d$. En revanche, la log-vraisemblance conditionnelle sachant $\|Z_i - Z_j\| = d_{ij}$ pour tout $i, j \in [n]$ est simplement affine en (β, D), où $D = (d_{ij})_{i,j\in[n]}$ est une matrice symétrique positive dont les coefficients vérifient l'inégalité triangulaire, ce qui permet d'obtenir des estimateurs des distances et des paramètres de la régression.

Stratégie bayesienne Le principal défaut de la méthode précédente est qu'elle sépare complètement l'estimation des positions de la classification des noeuds et de l'estimation des paramètres du mélange, alors qu'elles sont intimement liées. En particulier, l'estimation des positions ne tient pas du tout compte de la structure en clusters de la loi des positions. Réciproquement, la méthode ne fournit pas de région de confiance pour les positions ce qui rend plus difficile l'estimation des paramètres des clusters et donc la classification.

Handcock et al. (2007) propose pour pallier ce défaut une méthode complètement bayesienne, en plaçant des priors sur les paramètres β de la régression logistique et sur les paramètres du mélange gaussien, comme en estimation bayesienne usuelle des modèles de mélange (Diebolt and Robert, 1994). Itérativement, un algorithme de Metropolis-Hastings tire des positions, et des paramètres de régression logistique, puis la loi des labels de groupe classification et des paramètres du mélange sont mises à jour en calculant leur loi *a posteriori*.

Conclusion.

Les différentes étapes de ces méthodes d'inférence sont certes standards, mais leur empilement rend complexe l'analyse théorique de la procédure globale. De plus la plupart d'entre elles utilisent des algorithmes itératifs qui ont un coût non négligeable, ce qui limite la taille des graphes qu'ils peuvent traiter à quelques milliers de noeuds. Enfin, le modèle de mélange gaussien sur les positions des

classique est celle du contexte euclidien : il consiste à utiliser l'égalité du parallélogramme pour construire une matrice de Gram à partir de la matrice des distances. Puis on réalise une Analyse en Composantes Principales sur la matrice de Gram recentrée.

noeuds est une hypothèse forte, que nous levons dans le modèle KerNet où f n'est pas paramétrique. Le clustering n'est alors plus défini par les composantes du mélange. Nous donnerons et discuterons d'une définition dans ce cadre plus général en 4.3.

4.2 Algorithmes de clustering

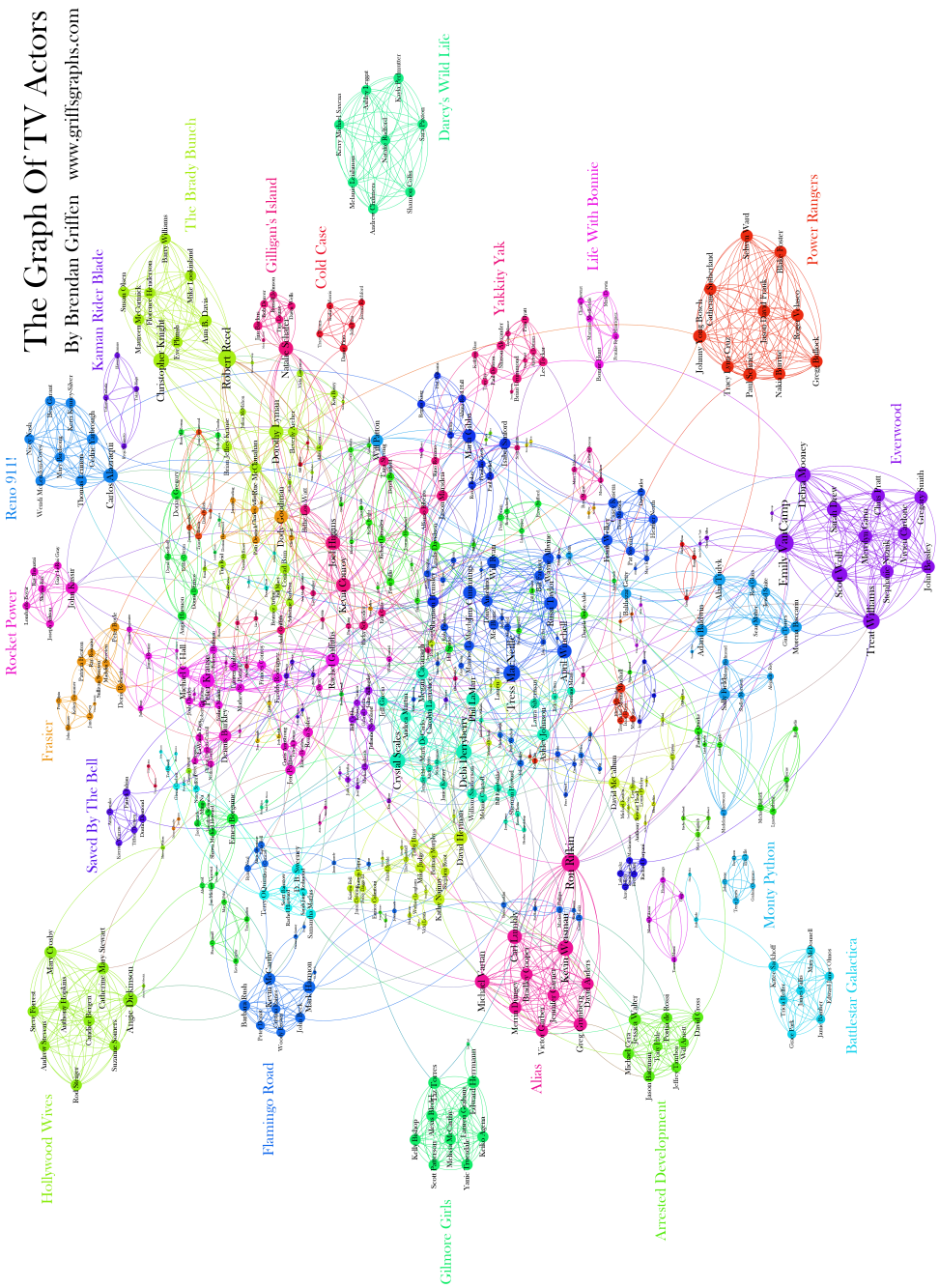
Dans cette section, nous présentons quelques exemples d'algorithmes dont le but n'est pas de retrouver les composantes d'un mélange. Ils ne sont donc généralement pas conçus pour traiter les graphes d'un modèle particulier, mais pour trouver des groupes de noeuds qui satisfont un certain critère d'homogénéité. Des exemples typiques de ce type d'algorithme sont la classification hiérarchique ou, hors du cadre des graphes, l'algorithme des centres mobiles (k-means). La classification hiérarchique a par exemple une application dans la recherche de groupes de noeuds très connectés entre eux, appelés communautés. Le lecteur peut aussi consulter Leger et al. (2013), qui établit une revue d'un grand nombre de méthodes non paramétriques dans les graphes, comparées sur des benchmarks.

La première sous-section discutera d'une classe d'algorithmes ayant pour objectif de trouver des communautés. Cette approche est l'objet de très nombreux travaux dans le domaine des graphes. Ensuite nous évoquerons la classe des algorithmes basés sur une décomposition spectrale du laplacien de la matrice d'adjacence du graphe. Selon les modalités de cette analyse, on peut étonamment construire des groupes de natures aussi variées que les communautés ou les classes du Stochastic Blockmodel. Enfin, nous présenterons le principe du K-linkage, utilisé dans le Chapitre 7.

4.2.1 Communautés

Comme annoncé au début de ce chapitre, le clustering recouvre beaucoup de définitions. Dans la littérature, l'étude de la structure en communautés des graphes est souvent appelée *clustering*. Cette structure se caractérise par la présence de sous-ensembles d'individus très denses, les communautés, peu connectées entre elles. Beaucoup de réseaux réels présentent une telle topologie, et chercher des communautés peut être utile dans les applications. Dans les réseaux sociaux, elles peuvent avoir une véritable interprétation sociologique : elles représentent des groupes de forte cohésion, comme des groupes politiques ou religieux, des groupes de chercheurs travaillant sur le même sujet, des cercles d'amis, des acteurs ayant joué dans les mêmes séries (voir Figure 4.5), etc. En biologie, les groupes dans les réseaux d'interaction protéiques représentent des ensembles de protéines ayant la même fonction cellulaire.

FIGURE 4.5 – Exemple de réseau réel présentant une structure en communautés. Les noeuds sont des acteurs de séries télévisées. Deux acteurs sont reliés s'ils ont joué dans la même série. Source : Griff's graphs, griffsgraphs.com



Comme le rappelle Girvan and Newman (2002), il subsiste parfois une confusion dans le langage entre la propriété de structure en communautés et la propriété de transitivité (parfois appelée aussi clustering [6]. En effet, on peut tout à fait avoir de la transitivité sans avoir de structure en communautés, c'est le cas des modèles LPCM et KerNet. À l'inverse, une structure en communautés a généralement un fort coefficient de regroupement. Les communautés n'ayant pas de définition rigoureuse, les objets donnés par les algorithmes de communautés sont les produits d'heuristiques. Elles forment un très vaste sujet, dont nous ne donnons ici que les grandes lignes ; le lecteur peut consulter l'article de Fortunato (2010), qui en fait une revue très exhaustive.

Classification hiérarchique. La première idée est de se munir d'une mesure de similarité idoine et d'effectuer une classification hiérarchique [7] menant à des groupes fortement connectés, comme par exemple le nombre de chemins indépendants entre deux noeuds. Des chemins entre deux mêmes noeuds sont dits indépendants par arêtes [8] s'ils n'ont aucune arête en commun. Le théorème de Menger affirme que le nombre de tels chemins entre deux noeuds i et j est égal au nombre de noeuds minimal qu'il faut retirer du graphe pour qu'il n'y existe plus de chemin les reliant. C'est donc une mesure de similarité adaptée aux communautés au sens où un grand nombre de chemins indépendants signifie que les noeuds sont fortement connectés l'un à l'autre.

Centralité. L'approche de Girvan and Newman (2002) tente d'identifier et de retirer les quelques arêtes qui séparent les communautés, afin d'extraire du graphe quelques composantes connexes fortement connectées, qui formeront les communautés. Ces arêtes sont centrales dans le graphe, au sens où l'interconnexion des deux parties du graphe situées à ses extrémités repose critiquement sur elles. Par exemple, dans la Figure 4.1 montrant des communautés (cas $Q = 2$), tous les chemins allant d'un noeud bleu à un noeud rouge passent par une seule et même arête. Ainsi l'algorithme de Girvan and Newman (2002) utilise le nombre de plus courts chemins traversant les arêtes comme mesure de leur centralité. Il procède en retirant itérativement des arêtes par centralité décroissante, en recalculant les

6. En référence au coefficient de regroupement, appelé *clustering coefficient* en anglais, voir sa définition dans les propriétés empiriques des réseaux en 2.3.1.

7. L'algorithme est le suivant : on part de groupes contenant un seul individu, puis on fusionne successivement les groupes les plus proches au sens de la similarité. On construit alors des classifications emboîtées, représentant différents niveaux de compromis entre homogénéité des groupes et nombre de clusters.

8. On peut définir de façon analogue des chemins indépendants par noeuds. Les seuls noeuds que partagent ces chemins sont alors les extrémités.

mesures à chaque étape [9] jusqu'à épuisement des arêtes. D'autres mesures de centralité d'une arête sont définies dans Newman and Girvan (2004), notamment par le biais de marches aléatoires sur le graphe.

Modularités. Newman and Girvan (2004) propose un critère de qualité d'une partition en communautés pour choisir la meilleure parmi toutes celles fournies par les itérations de l'algorithme de Girvan and Newman (2002). Ce critère, appelé modularité de Newman-Girvan, a finalement beaucoup été étudié pour lui-même, en oubliant la centralité. Il a été employé pour fournir les communautés d'un graphe par optimisation directe. Fortunato (2010) présente le cadre général des modularités comme le plus rigoureux pour l'étude des communautés. Une modularité établit une comparaison entre la densité des différents sous-graphes du graphe observé et celle qu'on obtiendrait dans un modèle de graphe d'hypothèse nulle, et s'écrit de manière générale, pour toute partition c :

$$\mathcal{Q}(c) = \frac{1}{2L} \sum_{i,j \in [n]} (X_{ij} - p_{ij}) \mathbb{1}_{c_i = c_j}$$

avec X le graphe observé, L son nombre d'arêtes, pour tous $i, j \in [n]$, p_{ij} la probabilité d'avoir une arête entre i et j dans le modèle d'hypothèse nulle. Ce principe de comparaison à un modèle d'hypothèse nulle pour détecter des communautés est aussi illustré dans Arias-Castro and Verzelen (2013), avec le modèle d'Erdős-Rényi pour modèle d'hypothèse nulle, de sorte à trouver des sous-ensembles de noeuds significativement plus denses que ceux issus d'une topologie aléatoire. Dans le cas où le modèle d'hypothèse nulle est le modèle à configuration (voir la loi des degrés dans 2.3.1), on retrouve la modularité de Newman-Girvan :

$$\mathcal{Q}_{NG}(c) = \sum_{k=1}^{n_c} \left[\frac{L_k}{L} - \left(\frac{S_k}{2L} \right)^2 \right]$$

avec n_c le nombre de communautés dans la partition c, L_k le nombre d'arêtes du sous-graphe de X induit par la communauté k, et S_k la somme des degrés dans X des noeuds de la communauté k. La différence de ces termes correspond à la comparaison entre la fraction d'arêtes d'une communauté et celle qu'aurait un graphe de topologie aléatoire avec les mêmes degrés que X.

À l'inverse des algorithmes précédents qui étaient d'une complexité accessible, l'optimisation globale d'une modularité est généralement impossible, au vu

9. Girvan and Newman (2002) indique en effet que si les communautés sont reliées par plus d'une arête, l'une d'entre elles est nécessairement traversée par un grand nombre de plus courts chemins, mais rien n'est sûr quant aux autres.

du nombre de partitions à explorer. De nombreuses techniques approchées d'optimisation ont été développées, dont on peut trouver une revue dans Fortunato (2010).

Bickel and Chen (2009) propose une modularité, interprétée comme une pseudo-vraisemblance, qui est consistente dans le SBM. La modularité de Newman-Girvan est aussi consistente dans la sous-classe des SBM correspondant à des structures en communautés [10], mais seulement dans le cas $Q = 2$. L'article fournit d'ailleurs un contre-exemple pour $Q = 3$.

4.2.2 Clustering spectral

Le clustering spectral désigne l'ensemble des méthodes fondées sur la décomposition spectrale d'un laplacien du graphe, dont le spectre est connu pour avoir des propriétés liées à celles du graphe (Mohar and Alavi, 1991). Il en existe plusieurs variantes, selon la normalisation du laplacien L et l'emploi qui en est fait ; on peut consulter Von Luxburg (2007) qui fait une revue de cette méthode avec différents points de vue. La définition originelle est la suivante :

$$L = D - X$$

où X la matrice d'adjacence du graphe et D est la matrice diagonale contenant ses degrés. Historiquement, le laplacien était un outil pour calculer le nombre d'arbres couvrants d'un graphe, via le théorème de Kirchhoff (voir par exemple Harris et al., 2008), qui affirme que le nombre d'arbres couvrants d'un graphe connexe est simplement égal au signe près à n'importe quel cofacteur de L. D'autre part, L n'admet que des valeurs propres positives, il admet 0 pour valeur propre, dont la multiplicité est égale au nombre de composantes connexes de X (Von Luxburg, 2007). Ces propriétés motivantes ont conduit Donath and Hoffman (1973) à proposer un algorithme de partitionnement de graphe basé sur le laplacien. Si Q est le nombre de clusters recherché, l'algorithme est le suivant :

Algorithme 4.1. *Clustering Spectral(X,Q)*

- *Calculer le laplacien L du graphe X*
- *Déterminer des vecteurs propres $(V_1, \ldots, V_Q)$ associés aux Q plus petites valeurs propres de L*
- *Soit V la matrice $n \times Q$ des vecteurs propres. Effectuer un algorithme des k-means dans $\mathbb{R}^Q$ sur les n lignes de V, chacune correspondant à un noeud de X.*
- *Retourner les groupes fournis par les k-means.*

10. Cas où chaque connectivité intra-classe π_{qq} est plus grande que la connectivité inter-classes $\sum_{r \neq q} \pi_{qr}$.

Le clustering spectral utilise des outils standards d'algèbre linéaire, ce qui d'une part lui confère une complexité accessible à des graphes d'une dizaine de milliers de noeuds, et permet aussi d'en faire des analyses théoriques. Von Luxburg et al. (2008) fait par exemple l'étude de la convergence des éléments propres du laplacien, interprétée comme consistence de la méthode.

L'emploi de l'algorithme basique des k-means peut paraître surprenante, mais cela souligne un aspect important du clustering. La tâche de clustering est en effet d'autant plus efficace que l'espace de représentation des données permet de mieux les séparer. Ainsi, l'hypothèse du clustering spectral est que la représentation spectrale des noeuds dans l'espace $\mathbb{R}^Q$ les sépare en général suffisament bien pour que même un algorithme aussi simple que les k-means trouve un partitionnement de qualité. C'est une idée aussi présente dans l'inférence du LPCM (voir 4.1.2), où l'une des méthodes proposées par Hoff et al. (2002) tente d'attribuer aux noeuds des positions dans un certain espace, pour ensuite faire un algorithme simple de type EM. Le procédé utilise d'ailleurs des méthodes spectrales via le MDS. Enfin, l'objet du Chapitre 6 est de montrer sous certaines hypothèses qu'à l'aide d'un algorithme très basique appliqué à une représentation même très simplifiée du graphe (celle de ses degrés normalisés), on peut retrouver les classes du SBM.

La nature des clusters fournis par le clustering spectral diffère selon la version du laplacien utilisé et l'ordre choisi pour les valeurs propres. La méthode classique présentée jusqu'ici fournit des communautés au sens défini dans le paragraphe précédent. On peut le comprendre intuitivement, avec la propriété que chaque composante connexe ajoute une unité à la multiplicité de la valeur propre nulle. Si le graphe a une structure de communautés, les premières valeurs propres sont proches de zéro et correspondent à des sous-groupes de noeuds bien connectés à l'intérieur, mais peu entre eux. D'autre part on notera le résultat très original de Rohe et al. (2010) qui affirme que l'algorithme ci-dessus fournit les classes du SBM de manière consistente quand il est appliqué au laplacien $L = D^{-1/2}XD^{-1/2}$, et avec les valeurs propres ordonnées par valeur absolue décroissante.

4.3 Modèle KerNet : définition géométrique du clustering

Ce paragraphe a un statut particulier : nous y introduisons la définition non-paramétrique du clustering qui sera utilisée dans le Chapitre 7. La définition n'est pas directement reliée aux graphes, elle est faite pour classer des individus ayant une position aléatoire dans une partie de $\mathbb{R}^d$ (Hartigan, 1975). Les clusters sont interprétés comme des régions où les individus sont très agrégés, séparées par des régions peu peuplées, vues comme des données atypiques non pertinentes. On

peut donner une définition mathématique à cette intuition dans le cas où la loi des positions admet une densité f (ici par rapport à la mesure de Lebesgue).

Définition 4.1. *Soit $t > 0$, on note $\mathcal{L}(t)$ l'ensemble de niveau t de f, soit :*

$$\mathcal{L}(t) = \{z \in \mathbb{R}^d;\ f(z) \geq t\}$$

On appelle t-cluster une composante connexe de $\mathcal{L}(t)$.

Par extension, on appelle clusters les sous-ensembles des points $(Z_i)_{i\in[n]}$ appartenant à la même composante connexe de $\mathcal{L}(t)$. Ainsi les clusters sont des groupes d'individus appartenant à la même région de forte densité. Cette définition peut être mise à profit dans les modèles de graphes aléatoires à positions latentes comme KerNet, car les noeuds ont une position aléatoire dans un espace métrique latent. Comme la définition ne concerne que l'espace latent, il faut que le graphe respecte d'une certaine façon la structure en clusters pour espérer la retrouver dans le graphe X. C'est à cela que sert l'hypothèse 7.3 de support compact du chapitre 7 pour la fonction de connexion g : ainsi les arêtes ne peuvent pas relier deux noeuds trop éloignés ni alors fusionner à tort deux clusters distincts. De manière plus générale, la vitesse de décroissance vers 0 de la fonction de connexion à l'infini doit être assez rapide.

Comparons cette définition non-paramétrique à celle fondée sur un modèle de mélange gaussien sphérique, utilisée dans le LPCM (voir 4.1.2). Tout d'abord la définition non-paramétrique des clusters exclut les points des données situés dans des régions dont la densité est inférieure à un seuil, contrairement à la définition des mélanges. Cela peut être vu comme un désavantage, puisque la classification n'est pas une partition totale des données. Cela peut au contraire être mis à profit pour se débarrasser de points résiduels nuisibles dans des données réelles. Pelletier and Pudlo (2011) par exemple en proposent cet usage. Contrairement à la définition des mélanges, ici aucune hypothèse n'est faite sur la densité, ce qui lui donne un caractère très général et assez souple. Elle tient compte donc notamment de clusters de formes géométriques beaucoup plus variées qu'une famille de mélanges fixée à l'avance. Dans les cas de formes complexes, le modèle de mélanges demanderait beaucoup de composantes pour pouvoir s'ajuster, entraînant un surparamétrage et un surpartitionnement artificiels des noeuds. La définition non- paramétrique n'a pas non plus d'*a priori* sur le nombre de clusters, alors que ce nombre fait partie des paramètres supposés connus de la définition des mélanges.

Dans les cas où les positions sont bien approchées par des normales, les deux définitions donneront des résultats similaires seulement si les clusters au sens géométrique sont bien séparés. Si les clusters sont très proches, la définition non-paramétrique ne sera capable de séparer que les points de très forte densité, jetant tous les autres points, alors que la définition des mélanges devrait donner de

meilleurs résultats puisqu'elle tient compte de la structure de la loi pour mieux classer.

En résumé, la définition non-paramétrique de clustering des noeuds d'un graphe est utile lorsqu'on n'a pas d'*a priori* sur le nombre ou la forme des clusters. Elle repose certes sur l'existence d'un modèle de graphe à positions latentes, mais dont la loi des positions est peu contrainte, ce qui est un bon compromis selon nous. C'est un cadre qui semble adapté en vue de réaliser des tests de « clustering » d'hypothèse nulle « il n'y a qu'un seul cluster » contre l'alternative « il y a plusieurs clusters ».

4.3.1 K-linkage

Dans ce paragraphe nous décrivons l'algorithme étudié au Chapitre 7. Il répond à un compromis, puisque sa consistence n'a de sens que dans un modèle, mais qui n'est pas contraint à être un modèle de mélanges. L'algorithme K-linkage consiste à d'abord élaguer le graphe observé X en retirant les noeuds de degré strictement inférieurs à K ; les clusters qu'il fournit sont les composantes connexes du graphe élagué.

Algorithme 4.2. *K-linkage(X,K)*
- *Calculer les degrés du graphe X*
- *Retirer les noeuds de degré $< K$*
- *Trouver les composantes connexes du graphe élagué par Depth First Search*
- *Retourner les composantes connexes comme clusters.*

Il sera utilisé au chapitre 7 surtout dans le but de compter le nombre de t-clusters (au sens du paragraphe précédent) de la densité des positions latentes dans le modèle KerNet. La raison pour laquelle les composantes connexes du graphe élagué sont proches des vrais clusters est que dans ce modèle, le degré d'un noeud normalisé par nh_n^d est un estimateur de la densité à la position latente de ce noeud. On peut remarquer par exemple que pour tout $i \in [n]$:

$$\mathbb{E}^{Z_i}(D_i/nh_n^d) = k_{h_n} \star f(Z_i)$$

où $\star$ est la convolution, et $k_{h_n} = k(\cdot/h_n)/h_n^d$.

En conséquence, garder seulement les noeuds de haut degré revient asymptotiquement quand n tend vers l'infini et h_n tend vers zéro, à ne garder que les noeuds situés dans des régions de forte densité. L'atout majeur de cet algorithme est sa faible complexité, au pire linéaire en nombre d'arêtes (il utilise principalement l'algorithme Depth First Search, voir en 5.4.3). Il est utilisable sur des graphes de quelques centaines de milliers de noeuds au lieu d'une dizaine de milliers au plus pour les autres algorithmes présentés dans cette section. Sa simplicité permet de

plus son analyse théorique, comme celle faite au Chapitre 7. Il a été introduit par Ling (1973), dans un contexte différent de celui de cette thèse, où en particulier les $(Z_i)_{i\in[n]}$ sont observés, et que nous discutons maintenant.

Relecture du clustering spectral et du K-linkage.

Les algorithmes de K-linkage et de clustering spectral ont été conçus à l'origine dans le but de classer des individus selon leurs positions $(Z_i)_{i\in[n]}$ observées dans un certain espace. En effet, pour classer, il faut pouvoir comparer les individus deux à deux, notamment calculer leur distance dans l'espace des données. L'objet graphe se prête très bien à cette utilisation, puisqu'il résume l'information portant sur les paires d'individus. Est née alors l'idée de construire un graphe à partir des positions, puis d'analyser ce graphe intermédiaire pour classer les individus. C'est là qu'intervient l'un des deux algorithmes.

Contrairement au cadre de cette thèse, l'objet graphe ainsi utilisé n'était donc pas la donnée elle-même initialement, mais un outil. Les arêtes du graphe « outil de classification » relient les noeuds à distance [11] inférieure à un seuil donné $r > 0$; il est appelé graphe de r-voisinage et c'est en fait un graphe géométrique (voir 3.4.2). Les algorithmes de clustering spectral ou de K-linkage présentés précédemment s'appliquent alors tels quels à ce graphe :

Algorithme 4.3. *Clustering*$((Z_i)_{i\in[n]})$
- *Calculer la matrice des distances (ou dissimilarités)* $(\|Z_i - Z_j\|)_{i,j\in[n]}$ *pour tout* $1 \leq i < j \leq n$.
- *Déterminer le graphe de* r*-voisinage :* $X = (\mathbb{1}_{\|Z_i - Z_j\|\leq r})_{i,j\in[n]}$
- *Appliquer le clustering spectral ou le* K*-linkage*

Comme le clustering spectral et le K-linkage n'utilisent que le graphe pour classer les individus et n'ont plus recours aux positions une fois le graphe construit, ces algorithmes sont des méthodes idéalement applicables à tout graphe aléatoire à variables latentes. Ainsi le cadre du modèle KerNet est en fait une relecture du problème initial de clustering présenté ici. Le modèle KerNet de paramètre h_n peut être interprété comme le graphe de h_n-voisinage, auquel on enlève des arêtes aléatoirement.

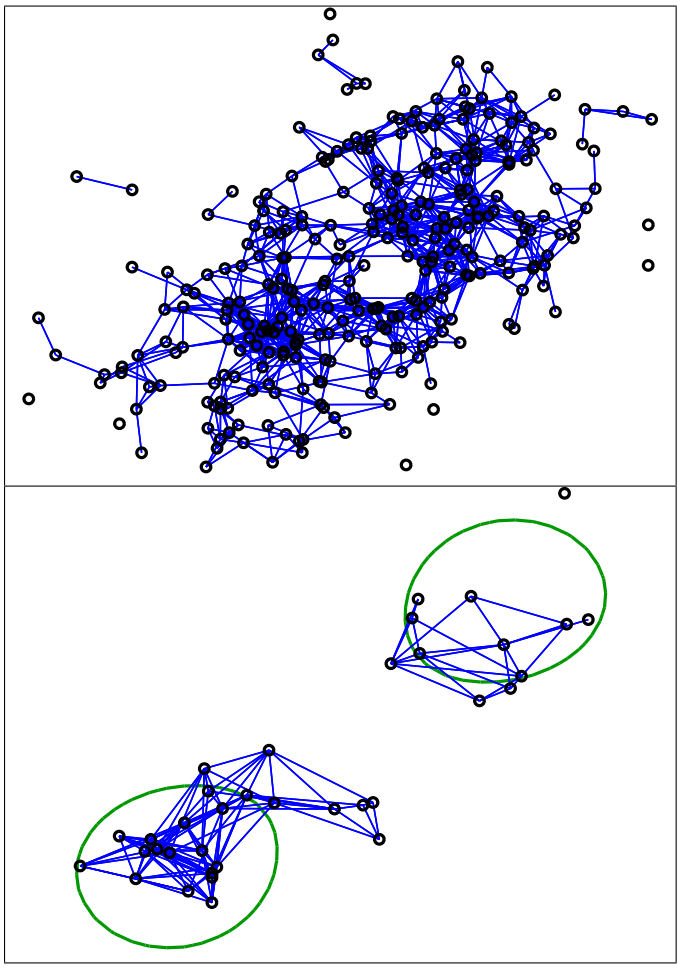
Noeuds isolés.

Comme nous le verrons dans l'étude de simulation réalisée au Chapitre 7, cet algorithme est très sensible aux petites composantes du graphe élagué, en particulier ses noeuds isolés, comme on le voit sur l'application de l'algorithme en

11. Ou plus généralement, dissimilarité.

Figure 4.6 : en plus des deux grandes composantes correspondant chacune à un cluster, il subsiste un noeud isolé. Le nombre de clusters trouvés par K-linkage augmente d'une unité quelque soit la taille des composantes connexes du graphe élagué, en particulier pour chaque noeud isolé. Or un tel noeud a peu de chances de représenter un cluster à lui tout seul. L'étude de simulation du Chapitre 7 illustrera qu'asymptotiquement, la sur-estimation du nombre de composantes dans un régime de faible densité vient en fait de ces petites composantes. Ainsi la robustesse de l'algorithme, et aussi sa vitesse de convergence peut-elle être améliorée dans le futur en retirant les petites composantes. Cette heuristique pourra être justifiée en étudiant distribution de la taille des composantes du graphe élagué.

FIGURE 4.6 – Un graphe sous le modèle KerNet avec un mélange gaussien bidimensionnel comme densité des positions (haut) et le même graphe élagué avec un seuil $K = tnh_n^2$ et la ligne de niveau t du mélange gaussien (bas)



Chapitre 5

Quelques résultats utiles

5.1 Inégalités de concentration

Les principaux résultats originaux de cette thèse sont des théorèmes de consistence de classifieurs ou d'estimateurs. Leurs énoncés sont assortis de majorants de la probabilité d'erreur de classification ou d'estimation, permettant ainsi d'avoir une idée — moins optimiste que la réalité — de la vitesse de cette consistence. Ces bornes sont obtenues à l'aide d'inégalités de concentrations pour des sommes de variables aléatoires réelles bornées et indépendantes.

L'inégalité d'Hoeffding est une inégalité de concentration très simple pour ce type de variables. Elle ne demande aucune autre hypothèse, et s'applique donc par exemple avec des suites de Bernoulli indépendantes, même de paramètres différents.

Proposition 5.1. *(Hoeffding)*
Soit $(a_i)_{i\in[n]}$ et $(b_i)_{i\in[n]}$ deux suites de réels. Soit $(X_i)_{i\in[n]}$ une suite de variables aléatoires indépendantes et telle que pour tout $i \in [n]$, $X_i \in [a_i, b_i]$ presque sûrement. Alors pour tout $t > 0$:

$$P\left(\sum_{i=1}^{n}(X_i - \mathbb{E}\left(X_i\right)) \leq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{5.1}$$

Cette inégalité est utilisée notamment dans le Chapitre 6. Cependant elle ne devrait être utilisée que lorsque la variance des variables ne peut être mieux bornée que par $\frac{(b_i-a_i)^2}{4}$, ce qui est le cas dans le chapitre mentionné. C'est en fait une borne grossière, puisque la variance d'une variable bornée ne peut de toute façon pas être plus grande, comme on le montre ici :

Proposition 5.2. *Soit X une variable aléatoire réelle bornée telle que $X \in [a, b]$*

presque sûrement. Alors :

$$\mathrm{Var}(X) \leq \frac{(b-a)^2}{4}$$

Démonstration. On pose $Y = \frac{X-a}{b-a}$. Y prend ses valeurs presque sûrement dans $[0,1]$. Soit $p = \mathbb{E}(Y)$ et Z une variable de Bernoulli de paramètre p. On a donc $\mathbb{E}(Z) = \mathbb{E}(Y) = p$. On montre qu'à espérances égales, Z a une plus grande variance que Y :

$$\begin{aligned}\mathrm{Var}(Z) - \mathrm{Var}(Y) &= \mathbb{E}\left(Z^2\right) - \mathbb{E}(Z)^2 - \mathbb{E}\left(Y^2\right) + \mathbb{E}(Y)^2 = p - \mathbb{E}\left(Y^2\right) \\ &\geq p - \mathbb{E}(Y) \ \ (\text{car } Y \leq 1 \Rightarrow \mathbb{E}\left(Y^2\right) \leq \mathbb{E}(Y)) \\ &\geq 0\end{aligned}$$

Ceci signifie que parmi toutes les variables à valeurs dans $[0,1]$ et d'espérance fixée p, c'est la loi de Bernoulli qui a la plus grande des variances. Ceci n'est pas étonnant puisqu'elle atteint cette espérance en ne prenant que les valeurs extrêmes du support qu'on s'autorise, $[0,1]$. Ce faisant, elle maximise l'écart à sa moyenne.

On rappelle que $\mathrm{Var}(Z) = p(1-p) \leq 1/4$, d'où :

$$\mathrm{Var}(X) = (b-a)^2 \mathrm{Var}(Y) \leq (b-a)^2 \mathrm{Var}(Z) \leq \frac{(b-a)^2}{4}$$

□

Dans les cas où on a un contrôle de la variance, l'inégalité de Bernstein permet d'en tirer parti pour obtenir une inégalité plus fine, comme on le verra dans le Lemme 5.1 par exemple, ou au cours du Chapitre 7. Dans le Chapitre 6, cette inégalité permettrait aussi d'améliorer les vitesses de convergence trouvées dans le cas des graphes de faible densité (voir Proposition 6.5).

Proposition 5.3. *(Bernstein)*

Soit $(X_i)_{i\in[n]}$ une suite de variables aléatoires indépendantes et telle que pour tout $i \in [n]$, $|X_i| \leq M$. Alors pour tout $t > 0$:

$$P\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{1}{2}\frac{t^2}{\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{1}{3}Mt}\right) \tag{5.2}$$

Avec les notations de l'inégalité de Hoeffding, $a_i = -M$ et $b_i = M$ pour tout $i \in [n]$. Par comparaison à celle-ci, l'inégalité de Bernstein devient plus intéressante qu'Hoeffding dès que :

$$\sum_{i=1}^{n}\mathrm{Var}(X_i) + \frac{Mt}{6} \leq \frac{\sum_{i=1}^{n}(b_i-a_i)^2}{4} = Mn^2$$

5.2 Éléments d'estimation de densité

Quelques notions d'estimation de densité seront utilisées dans le Chapitre 7. Le lecteur pourra trouver une introduction à ce sujet dans Tsybakov (2003), et de nombreux résultats dans Prakasa Rao (1983) par exemple. Certains de ces résultats sont adaptés ici aux besoins du Chapitre 7, dans lequel nous aurons besoin d'une estimation de qualité fixée, mais pas nécessairement consistente. Nous remettrons notamment en cause le traditionnel compromis entre le biais et la variance de l'estimateur à noyau : nous pourrons nous permettre dans certains cas de sacrifier le biais en se contentant de le borner sans chercher à le faire tendre vers zéro.

On suppose qu'on dispose d'un n-échantillon $(Z_i)_{i\in\mathbb{N}}$ de loi de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$. Soit g une fonction positive intégrable sur $\mathbb{R}^d$ à support inclus dans la boule unité $B(0,1)$.

On définit pour tout $h>0$ la fonction g_h par $g_h(z)=\frac{1}{h^d}g(z/h)$. L'idée intuitive des estimateurs à noyau est d'estimer la densité par une combinaison convexe de densités centrées sur les n observations (les $g_h(\cdot - Z_i)$) avec des poids égaux à $1/n$. L'estimateur à noyau de f est ainsi défini en tout $z\in\mathbb{R}^d$ par :

$$f_n(z)=\frac{1}{nh_n^d}\sum_{i\in[n]} g\left(\frac{z-Z_i}{h_n}\right)=\frac{1}{n}\sum_{i\in[n]} g_{h_n}(z-Z_i)$$

Notons que plus h est petit, plus la densité g_h est piquée, et donc plus l'estimateur sera capable de restituer des variations importantes de f.

La proposition suivante donne l'espérance de l'estimateur à noyau en tout point $z\in\mathbb{R}^d$, qui est la convolution en z de la fonction f avec g_{h_n}, qui est une version approchée de f.

Proposition 5.4.

$$\mathbb{E}\left[f_n(z)\right]=g_{h_n}\star f(z)$$

Démonstration.

$$\begin{aligned}
\mathbb{E}\left(f_n(z)\right)&=\mathbb{E}\left(\frac{1}{nh_n^d}\sum_{i=1}^n g\left(\frac{z-Z_i}{h_n}\right)\right)\\
&=\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\frac{1}{h_n^d}g\left(\frac{z-Z_i}{h_n}\right)\right)\\
&=\frac{1}{n}\sum_{i=1}^n \int_{\mathbb{R}^d}\frac{1}{h_n^d}g\left(\frac{z-u}{h_n}\right)f(u)du\\
&=\int_{\mathbb{R}^d} g_{h_n}(z-u)f(u)du
\end{aligned}$$

□

Ainsi le biais ponctuel de l'estimateur f_n en $z \in \mathbb{R}^d$ est l'écart entre $f(z)$ et sa version approchée $g_{h_n} \star f$. Dans les références données en début de section, le lecteur pourra voir que cet écart est généralement d'autant plus petit que f est régulière et que h_n est petit.

5.2.1 Bornes uniformes du biais

Pour borner uniformément le biais, nous aurons besoin de la continuité uniforme de f comme nous le verrons dans la proposition suivante. On définit s_h la variation maximale de f sur l'ensemble des boules de rayon $h > 0$:

$$s_h = \sup_{\|z-z'\| \leq h} |f(z) - f(z')|$$

L'écart entre f et sa convolution avec la fonction k_h peut être classiquement bornée (voir par exemple Prakasa Rao, 1983) en fonction de s_h :

Proposition 5.5. *Pour tout $h > 0$ et pour tout $z \in \mathbb{R}^d$:*

$$\left| g_h \star f(z) - f(z) \int_{\mathbb{R}^d} g \right| \leq s_h \int_{\mathbb{R}^d} g$$

Démonstration.

$$\begin{aligned}
\left| g_h \star f(z) - f(z) \int_{\mathbb{R}^d} g \right| &= \left| \int_{\mathbb{R}^d} g_h(u) f(z-u) du - f(z) \int_{\mathbb{R}^d} g_h \right| \\
&\leq \int_{\mathbb{R}^d} g_h(u) |f(z-u) - f(z)| du \\
&= \int_{B(0,h)} g_h(u) |f(z-u) - f(z)| du \text{ car } Supp(g_h) \subset B(0,h) \\
&\leq \sup_{\|z-z'\| \leq h} |f(z) - f(z')| \int_{B(0,h)} g_h = s_h \int_{\mathbb{R}^d} g
\end{aligned}$$

□

On suppose désormais que g est un noyau, c'est-à-dire que de plus, son intégrale est 1. La proposition affirme que s_h borne uniformément le biais sur $\mathbb{R}^d$.

Pour $\varepsilon > 0$, par uniforme continuité de f, il existe η_ε tel que pour tout $h \in [0, \eta_\varepsilon]$, $s_h \leq \varepsilon$. Dans le Chapitre 7, nous aurons simplement besoin de borner le biais par une constante petite $\varepsilon > 0$, mais qui ne tend pas nécessairement vers 0 quand n tend vers l'infini. Du fait que l'uniforme continuité implique $s_h \xrightarrow[h \to 0]{} 0$, elle assurera que s_{h_n} et donc le biais est plus petit que ε, du moment que $h_n \leq \eta_\varepsilon$ sans nécessairement supposer que la fenêtre h_n tend vers 0.

Lemma 5.1. *Soit $\varepsilon > 0$. Sous l'hypothèse $h_n \leq \eta_\varepsilon$ (défini ci-dessus), pour tout $z \in \mathbb{R}^d$:*

$$P\left(f_n(z) - f(z) > \varepsilon\right) \leq \exp\left(-\frac{1}{8}\frac{\varepsilon^2 n h_n^d}{\left(\|f\|_\infty + \frac{\varepsilon}{2}\right)\int_{\mathbb{R}^d} g^2 + \frac{\varepsilon}{3}}\right)$$

Démonstration.
$P\left(f_n(z) - f(z) > \varepsilon\right) \leq P\left(f_n(z) - \mathbb{E}\left[f_n(z)\right] > \varepsilon/2\right) + P\left(\mathbb{E}\left[f_n(z)\right] - f(z) > \varepsilon/2\right)$

Le second événement est en fait déterministe, donc sous l'hypothèse 7.2.(a), et d'après les Propositions 5.5 et 5.4 :

$$P\left(\mathbb{E}\left[f_n(z)\right] - f(z) > \varepsilon/2\right) = P\left(g_{h_n} \star f(z) - f(z) > \varepsilon/2\right) \leq P\left(s_{h_n} > \varepsilon/2\right) = 0$$

Le premier terme, lui, est majoré à l'aide de l'inégalité de Bernstein :

$$\begin{aligned} P\left(f_n(z) - \mathbb{E}\left[f_n(z)\right] > \varepsilon/2\right) &= P\left(\sum_{i=1}^n \left[g\left(\frac{z - Z_i}{h_n}\right) - \mathbb{E}\left[g\left(\frac{z - Z_i}{h_n}\right)\right]\right] > \frac{\varepsilon n h_n^d}{2}\right) \\ &\leq \exp\left(-\frac{1}{8}\frac{(\varepsilon n h_n^d)^2}{\sum_{i=1}^n \operatorname{Var} g\left(\frac{z-Z_i}{h_n}\right) + \frac{\varepsilon n h_n^d}{6}}\right) \end{aligned}$$

Maintenant nous majorons la variance qui apparaît dans la borne. On définit $\bar{g} = g^2$.

$$\begin{aligned} \operatorname{Var} g\left(\frac{z - Z_i}{h_n}\right) &= \int_{\mathbb{R}^d} g^2\left(\frac{z-u}{h_n}\right) f(u)du - \left(\int_{\mathbb{R}^d} g\left(\frac{z-u}{h_n}\right) f(u)du\right)^2 \\ &= h_n^d \bar{g}_{h_n} \star f(z) - h_n^{2d}\left(g_{h_n} \star f(z)\right)^2 \\ &\leq h_n^d \left\|\bar{g}_{h_n} \star f\right\|_\infty \\ &\leq h_n^d \left(\|f\|_\infty \int_{\mathbb{R}^d} g^2 + s_{h_n} \int_{\mathbb{R}^d} g^2\right) \text{ en utilisant la Proposition 5.5 pour } \bar{g}. \end{aligned}$$

Nous obtenons finalement :

$$P\left(f_n(z) - \mathbb{E}\left(f_n(z)\right) > \varepsilon/2\right) \leq \exp\left(\frac{1}{8}\frac{\varepsilon^2 n h_n^d}{\left(\|f\|_\infty + \frac{\varepsilon}{2}\right)\int_{\mathbb{R}^d} g^2 + \frac{\varepsilon}{6}}\right)$$

□

5.3 Éléments de géométrie différentielle

Dans cette section, nous donnons des résultats de géométrie différentielle utiles au Chapitre 7. On suppose que $f : \mathbb{R}^d \to \mathbb{R}$ est une densité de probabilité. On note :

$$M_t = \{z \in \mathbb{R}^d;\ f(z) = t\}$$

On suppose que f vérifie les hypothèses de régularité (7.1) du Chapitre 7, rappelées ici :

1. f est uniformément continue
2. f est de classe $\mathcal{C}^1$ sur un voisinage de M_t.
3. Le gradient de f ne s'annule pas sur M_t.

5.3.1 Résultats utiles

Dans le Chapitre 7, nous aurons besoin d'encadrer le volume de l'espace compris entre deux hypersurfaces de niveau :

Proposition 5.6. *Il existe ε_0 et des constantes $a, A \geq 0$ tels que pour tout $\varepsilon \in [0, \varepsilon_0]$,*

$$a\varepsilon \leq Vol(\mathcal{L}(t-\varepsilon) \backslash \mathcal{L}(t)) \leq A\varepsilon.$$

C'est pour démontrer ce résultat que nous faisons en particulier appel aux voisinages patibulaires, mais tubulaires. La démonstration se situe en 5.3.3.

5.3.2 Voisinages tubulaires

Nous rappelons quelques notions relatives aux sous-variétés de $\mathbb{R}^d$.

Définition 5.1. *Une partie M de $\mathbb{R}^d$ est une sous-variété de dimension k de $\mathbb{R}^d$ si pour tout $z \in M$, il existe un voisinage ouvert U de z dans $\mathbb{R}^d$, un voisinage ouvert V de 0 dans $\mathbb{R}^k$, et un difféomorphisme ϕ de U dans V tel que $\phi(U \cap M) = V$.*

Une sous-variété est ainsi une union de petites portions d'espaces vectoriels déformés de manière difféomorphe.

Définition 5.2. *Une submersion est une application f de classe $\mathcal{C}^1$ d'un ouvert U de $\mathbb{R}^d$ dans $\mathbb{R}^{d-p}$ dont la différentielle est surjective en tout point.*

Sous les hypothèses (7.1), le gradient de f ne s'annule pas, y compris sur un voisinage ouvert U de M_t, par continuité de la différentielle au voisinage de M_t. f est donc une submersion de U dans $\mathbb{R}$.

Proposition 5.7. *Une partie M de $\mathbb{R}^d$ est une sous-variété de dimension k de $\mathbb{R}^d$ si et seulement si pour tout $z \in M$, il existe un ouvert U de $\mathbb{R}^d$, une submersion $f : U \to \mathbb{R}^{d-k}$ et un $t \in \mathbb{R}^{d-k}$ tels que :*

$$z \in U \text{ et } U \cap M = f^{-1}(t)$$

On peut trouver cette Proposition et sa démonstration dans tout ouvrage sur la géométrie différentielle, par exemple Lafontaine (2012). Ainsi pour tout $t \geq 0$, M_t est une sous-variété de dimension $d-1$ de $\mathbb{R}^d$.

Définition 5.3. *Soit M une sous-variété de l'espace euclidien $\mathbb{R}^d$.*

- *L'espace tangent à une sous-variété M de $\mathbb{R}^d$ en un point z, noté T_zM, est l'ensemble des points m tels que $\overrightarrow{zm}$ soit tangent à M en a.*
- *Le fibré normal à M, noté $N(M)$ est l'ensemble des couples (z, v) de $M \times \mathbb{R}^d$ tels que v soit orthogonal à T_zM.*

Pour ne pas alourdir l'exposé, nous ne donnons pas les propriétés de structure du fibré normal, que le lecteur peut trouver par exemple dans Lafontaine (2012), et nous donnons directement le théorème du voisinage tubulaire, qui se suffit à lui-même. Il identifie le fibré normal à un voisinage de la sous-variété. On note pour $r > 0$:

$$\mathcal{V}_r(M) = \{z \in \mathbb{R}^d;\ \mathrm{dist}(z, M) < r\}$$

Théorème 5.1. *(Berger and Gostiaux, 1992) Soit M une sous-variété compacte de l'espace euclidien $\mathbb{R}^d$, et $N(M)$ le fibré normal à M. Pour $r > 0$, on note $N_r(M)$ la partie de $N(M)$ suivante :*

$$N_r(M) = \{(p, v) \in N(M);\ \|v\| < r\}$$

Il existe $r_0 > 0$ tel que pour tout $r < r_0$, l'application $\varphi : N(M) \subset M \times \mathbb{R}^d \to \mathbb{R}^d$ définie par $\varphi(z, v) = z + v$ est un difféomorphisme de $N_r(M)$ sur $\mathcal{V}_r(M)$. Le voisinage $\mathcal{V}_r(M)$ est alors appelé voisinage tubulaire de M de rayon r.

Ce théorème sera utilisé dans la démonstration de la Proposition 5.6, et sert de cadre au théorème suivant, qui sera aussi utilisé dans cette démonstration.

Théorème 5.2. *(Berger and Gostiaux, 1992) Soit M une sous-variété compacte de dimension k de l'espace euclidien $\mathbb{R}^d$. Si pour $r > 0$, $\mathrm{Vol}(\mathcal{V}_r(M))$ est un voisinage tubulaire, alors c'est un polynôme en r :*

$$\mathrm{Vol}(\mathcal{V}_r(M)) = \sum_{i=1}^{\lfloor k/2 \rfloor} b_{2i} r^{d-k+2i}$$

où les $(b_{2i})_{1 \leq i \leq \lfloor k/2 \rfloor}$ dépendent de d et de M.

Le terme dominant est r^{d-k}. En appliquant ce théorème à M_t qui est de dimension $d-1$, le terme dominant est donc r. En particulier, il existe des constantes $a, A \geq 0$ dépendant de d et f seulement telles que :

$$ar \leq \mathrm{Vol}(\mathcal{V}_r(M)) \leq Ar$$

5.3.3 Application : Preuve de la Proposition 5.6

La preuve de cette proposition est en partie inspirée de celle de la Proposition A.2 dans Biau et al. (2007).

Démonstration.

Étape 1. Utilisation de la continuité On montre maintenant qu'il existe un $\varepsilon_0 > 0$ tel que toutes les hypersurfaces de niveau $t - \varepsilon$ avec $\varepsilon \leq \varepsilon_0$ sont incluses dans le voisinage tubulaire de M_t. Ceci revient à montrer que pour un certain $\varepsilon_0 > 0$, tous les points de la couronne $\mathcal{L}(t-\varepsilon_0)\backslash\mathcal{L}(t)$ sont aussi proches que l'on veut de M_t. On procède par l'absurde, et on suppose donc :

$$\exists r > 0 \ \forall \varepsilon > 0 \ \exists z \in \{t - \varepsilon \leq f \leq t\};\ d(z, M_t) \geq r$$

Soit $(\varepsilon_n)_{n\in\mathbb{N}}$ une suite de réels strictement positifs tendant vers 0, et pour chaque $n \in \mathbb{N}$, $z_n \in \{t - \varepsilon_n \leq f \leq\}$ vérifiant la condition ci-dessus. Soit $m = \sup_{n\in\mathbb{N}} \varepsilon_n$. En particulier, pour tout n, on a $z_n \in \{t - m \leq f \leq t\}$.

Comme $\varepsilon_n \to 0$, $m < +\infty$ (et m est même atteint, c'est un maximum). Par continuité de f, $\{t - m \leq f \leq t\}$ est fermé. Comme f est une densité, c'est aussi un ensemble borné (sinon la mesure associée à f aurait une masse infinie). $\{t - m \leq f \leq t\}$ est donc un compact de $\mathbb{R}^d$.

En conséquence, la suite $(z_n)_{n\in\mathbb{N}}$ admet une valeur d'adhérence $a \in \{t - m \leq f \leq t\}$. À une extraction près, cette suite converge vers a, on notera encore $(z_n)_{n\in\mathbb{N}}$ cette suite extraite. On a alors :

$$\left.\begin{array}{rl} t - \varepsilon_n \leq f(z_n) \leq t \text{ d'où :} & f(z_n) \xrightarrow[n\to\infty]{} t \\ \text{Par continuité de } f : & f(z_n) \xrightarrow[n\to\infty]{} f(a) \end{array}\right\} \Rightarrow f(a) = t$$

D'où $a \in M_t$. En résumant, on obtient une contradiction, du fait que pour tout $n \in \mathbb{N}$, z_n est à distance au moins $r > 0$ de la hypersurface de niveau t, tout en convergeant vers l'élément a de cette hypersurface de niveau :

$$0 < r \leq d(z_n, M_t) \leq d(z_n, a) \xrightarrow{n\to\infty} 0$$

On obtient donc pour tout $r > 0$, l'existence d'un $\varepsilon_0 > 0$ tel que pour tout $\varepsilon \leq \varepsilon_0$:

$$\{f = t - \varepsilon\} \subset \mathcal{V}_r(M_t)$$

Étape 2. Utilisation de la structure différentielle : Théorème du voisinage tubulaire D'après ce théorème, appliqué à la sous-variété compacte de l'espace euclidien $\mathbb{R}^d$ qu'est M_t, il existe un $r_0 > 0$ tel que pour tout $0 < r < r_0$, l'application $\varphi : N(M_t) \to \mathbb{R}^d$ définie par $\varphi(z, v) = z + v$ pour tout $(z, v) \in N(M_t) \subset M_t \times \mathbb{R}^d$ établit un difféomorphisme entre $N_r(M_t)$ et $\mathcal{V}_r(M_t)$.

Or d'après l'étape 1, pour ce r_0, il existe ε_0 tel que pour tout $\varepsilon \leq \varepsilon_0$, $\mathcal{L}(t-\varepsilon) \subset \mathcal{V}_r(M_t)$. Par conséquent, pour tout $\varepsilon \leq \varepsilon_0$, on peut écrire $z \in M_{t-\varepsilon}$, comme $z = p_z + v_z$ avec $(p_z, v_z) \in N(M_t)$. De plus $v_z = -d_z e_{p_z}$ avec $d_z = d(z, M_t)$ et $\|e_{p_z}\| = 1$. En développant f en p_z, il existe $\xi > 0$ tel que :

$$f(z) = f(p_z - d_z e_{p_z}) = f(p_z) + D_{e_{p_z}} f(p_z + \xi e_{p_z}) d_z$$

ce qui équivaut à :

$$t + \varepsilon = t + D_{e_{p_z}} f(p_z + \xi e_{p_z}) d_z$$

Étape 3. Encadrement du volume des voisinages tubulaires Vu que le gradient de f ne s'annule pas sur M_t, et même sur $\mathcal{V}_r(M_t)$ pour r assez petit (par continuité de sa différentielle), on peut désormais obtenir une majoration de la distance d_z, uniforme sur les $z \in M_{t-\varepsilon}$ et en fonction de ε :

$$\frac{\varepsilon}{\sup_{z \in \mathcal{V}_{r_0}(M_t)} D_{e_{p_z}} f(z)} \leq d_z \leq \frac{\varepsilon}{\inf_{z \in \mathcal{V}_{r_0}(M_t)} D_{e_{p_z}} f(z)}$$

En conséquence, en notant $r_1(\varepsilon)$ le minorant et $r_2(\varepsilon)$ le majorant, $\mathcal{V}_{r_1(\varepsilon)} \subset M_{t-\varepsilon} \subset \mathcal{V}r_2(\varepsilon)$, et par suite :

$$\mathrm{Vol}(\mathcal{V}_{r_1(\varepsilon)}) \leq \mathrm{Vol}(\mathcal{L}(t-\varepsilon) \backslash \mathcal{L}(t)) \leq \mathrm{Vol}(\mathcal{V}_{r_2(\varepsilon)})$$

On conclut avec la Proposition 5.2 ; il existe des constantes $a, a', A, A' \geq 0$, qui ne dépendent que de d et f telles que :

$$a'\varepsilon \leq a r_1(\varepsilon) \leq \mathrm{Vol}(\mathcal{L}(t-\varepsilon) \backslash \mathcal{L}(t)) \leq A r_2(\varepsilon) \leq A'\varepsilon$$

□

5.4 Éléments de théorie des graphes

5.4.1 Monotonie de la probabilité de connexité dans les graphes d'Erdős-Rényi hétérogènes

La connexité est une propriété monotone au sens où si X et X' sont deux graphes de même taille tel que $X \subset X'$ et X est connexe, alors X' aussi. Le

fait d'ajouter des arêtes à un graphe connexe le laisse connexe. Dans les graphes aléatoires, cela se traduit par le fait que la probabilité de connexité d'un graphe aléatoire est croissante en fonction des probabilités de ses arêtes, ce que montre le lemme suivant qui sera utilisé dans 7. Soit $\mathbf{p} = (p_{ij})_{i,j\in[n]}$ et $P_{\mathbf{p}}$ la loi d'un graphe aléatoire X tel que les arêtes sont indépendantes et pour tout $i, j \in [n]$, $P_{\mathbf{p}}(X_{ij} = 1) = p_{ij}$. Une telle loi est appelée modèle de graphe d'Erdős-Rényi hétérogène noté $\mathcal{ER}(n, \mathbf{p})$ —- si tous les $(p_{ij})_{i,j\in[n]}$ sont égaux à $p \in [0, 1]$, c'est le modèle $\mathcal{ER}(n, p)$ —.

Lemma 5.2. *Soit* $\mathbf{p} = (p_{ij})_{i,j\in[n]}, \mathbf{p'} = (p'_{ij})_{i,j\in[n]} \in [0, 1]^{n\times n}$. *Si pour tous* $i, j \in [n]$, $p_{ij} \leq p'_{ij}$, *alors :*

$$P_{\mathbf{p}}(X \text{ connexe}) \leq P_{\mathbf{p'}}(X \text{ connexe})$$

Démonstration. Soit $0 < p \leq p' \leq 1$ et $X' \sim \mathcal{B}(p')$. On montre maintenant qu'il existe une variable aléatoire X de loi $\mathcal{B}(p)$ telle que $X' \geq X$. On définit la variable aléatoire X de la façon suivante :

$$X = 0 \text{ sur } \{X' = 0\} \text{ et } X|\ X' = 1 \quad \sim \mathcal{B}\left(\frac{p}{p'}\right)$$

Par construction, X est une variable de Bernoulli et de plus elle est de paramètre p :

$$P(X = 1) = P(X = 1|\ X' = 0)P(X' = 0) + P(X = 1|\ X' = 1)P(X' = 1) = p$$

Soit $X' \sim \mathcal{ER}(n, \mathbf{p'})$. Pour tous $i, j \in [n]$, $X'_{ij} \sim \mathcal{B}(p'_{ij})$. Soit $X_{ij} \sim \mathcal{B}(p_{ij})$ telle que $X'_{ij} \geq X_{ij}$ comme dans le paragraphe précédent. Vu que les X'_{ij} sont mutuellement indépendantes, et que la loi de X_{ij} définie conditionnellement à X'_{ij}, les $(X_{ij})_{i,j\in[n]}$ peuvent être choisis mutuellement indépendantes aussi. Donc la loi du graphe aléatoire X est $\mathcal{ER}(n, \mathbf{p})$. De plus $X \subset X'$, car $X'_{ij} \geq X_{ij}$ pour tous $i, j \in [n]$. Alors, si X est connexe, alors X' est aussi connexe. D'où :

$$P(X' \text{ connexe}) \leq P(X \text{ connexe})$$

□

5.4.2 Connexité asymptotique dans le modèle $\mathcal{G}(n, p)$ avec $p > 0$ fixé

Gilbert (1959) montre l'inégalité suivante, duquel nous obtenons un majorant simplifié (ci-dessous) de la probabilité qu'un graphe de $\mathcal{G}(n, p)$ soit non connexe.

Theorem 5.1. *(Gilbert, 1959) Pour tout $q \in [0,1]$, pour tout $n \in \mathbb{N}^*$:*

$$nq^{n-1}\left(1-\frac{n-1}{2}q^{n-1}\right) \leq 1-\pi_{n,p} \leq q^{n-1}\left((1+q^{(n-2)/2})^{n-1}-q^{(n-1)(n-2)/2}\right) + q^{n/2}\left((1+q^{(n-2)/2})^{n-1}-1\right)$$

Proposition 5.8.

$$1-\pi_{n,p} \leq nq^{n-1}\left(1+(n-1)q^{(n-2)/2}\exp\left((n-1)q^{(n-2)/2}\right)\right)$$

En particulier, il existe $\kappa_q > 0$ tel que pour tout $n \in \mathbb{N}^*$:

$$1-\pi_{n,p} \leq \kappa_q n q^{n-1}$$

Démonstration. On part de l'inégalité de 5.1. On commence par enlever le terme négatif de la borne supérieure du théorème 5.1, qui n'aidera pas puisqu'il est négligeable devant les autres. On s'attaque au facteur du second terme. On utilise tout d'abord l'inégalité $(1+x)^a \leq e^{ax}$ vraie pour tout $x \in \mathbb{R}$, qui donne :

$$(1+q^{(n-2)/2})^{n-1} \leq \exp\left((n-1)q^{(n-2)/2}\right)$$

Par l'inégalité des accroissements finis, on pour tout $x \geq 0$, $e^x - 1 \leq xe^x$. On obtient donc :

$$(1+q^{(n-2)/2})^{n-1}-1 \leq \exp\left((n-1)q^{(n-2)/2}\right)-1 \leq (n-1)q^{(n-2)/2}\exp\left((n-1)q^{(n-2)/2}\right)$$

En utilisant encore $(1+x)^a \leq e^{ax}$ pour le premier terme cette fois, puis en combinant cette inégalité avec celle juste au-dessus, on a :

$$1-\pi_{n,p} \leq nq^{n-1}\exp\left((n-1)q^{(n-2)/2}\right)$$

Puis en développant l'exponentielle en série, on obtient l'inégalité annoncée. □

5.4.3 Algorithme Depth First Search (DFS)

Cet algorithme reçoit en entrée un graphe et un noeud, et renvoie la composante connexe de ce noeud dans le graphe. Ses premiers développements connus sont dus au mathématicien français du XIX^{e} siècle Charles-Pierre Trémaux, qui cherchait une méthode pour sortir d'un labyrinthe sans repasser indéfiniment au même endroit (voir Even, 2011). L'algorithme est aussi appelé *Parcours en profondeur*. L'idée de l'algorithme consiste à trouver un arbre couvrant de la composante connexe du noeud donné en entrée, d'où son nom : à la fin l'algorithme a parcouru tous les noeuds de cet arbre en partant de la racine, en descendant tant qu'il peut le long des branches, et en ne remontant que quand il est obligé.

Le problème pour trouver la composante d'un noeud dans graphe réside dans ses cycles : si l'algorithme allait de noeud en noeud en suivant les arêtes sans se poser de question, il pourrait ainsi ne jamais s'arrêter s'il était pris dans un cycle. Charles-Pierre Trémaux, tout comme quiconque a joué à Mario Bros. 1 et est arrivé jusqu'à Koopa dans tous les châteaux, a très bien compris ce phénomène. Par exemple dans le graphe de la figure 5.1, l'algorithme pourrait parcourir en boucle le cycle $A-B-C$ ou $A-C-D-E$. L'algorithme marque donc les noeuds qu'il a déjà parcourus afin de ne pas y repasser indéfiniment.

L'algorithme est décrit récursivement comme suit. On donne en arguments un noeud et un graphe dans lequel chaque noeud a une étiquette *marqué* ou *non marqué*.

Algorithm 5.1. *DFS(noeud r, graphe G, étiquettes L)*
- *Changer l'étiquette du noeud r dans L en* marqué.
- *Pour chacun des voisins v non marqués de r dans G, appeler DFS(v,G,L).*
- *Retourner la concaténation des listes de noeuds données par les appels précédents avec le noeud r.*

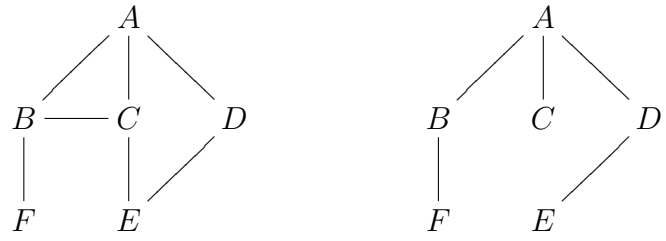
Dans un langage tel que MatLab, il est plus courant de coder itérativement, ce qu'on peut faire à l'aide d'une pile LIFO [1]. La racine est mise dans la pile, et est marquée. Puis on la ressort pour la remplacer par tous ses voisins dans le graphe, qui sont marqués à leur tour. Puis on ressort le dernier des voisins de la racine ajouté à la pile pour le remplacer par ses propres voisins non marqués, etc. À chaque fois qu'un noeud entre dans la pile, il est marqué.

Algorithm 5.2. *DFS(noeud v, graphe G)*
- *Créer une pile de longueur 1 contenant le noeud r et marquer le noeud r*
- *TANT QUE la pile n'est pas vide, FAIRE :*
  - *Trouver l'ensemble V de tous ses voisins non marqués*
  - *Enlever le dernier élément de la pile et le remplacer par V en augmentant la taille de pile si* $\mathrm{Card}(V) > 1$.
  - *Marquer tous les noeuds de V*
- *Retourner l'ensemble des noeuds marqués.*

Dans sa forme itérative, l'arbre est ainsi construit branche par branche. On descend le long d'une branche en mémorisant tous les embranchements rencontrés. Lorsqu'il a atteint une feuille, c'est-à-dire un noeud sans voisin non marqué dans le graphe, l'algorithme remonte au dernier embranchement vu et continue la construction de l'arbre à partir de là. Par exemple dans la figure 5.1, il commence à A, qui se fait marquer, puis mémorise tous les embranchements possibles B, C, D en marquant tous les noeuds. Il descend par D, puis n'ajoute que E car A est

1. LIFO : Last In First Out, le dernier élément empilé est traité et dépilé en premier.

FIGURE 5.1 – Parcours en profondeur : graphe en entrée (gauche) et son arbre couvrant (droite)



marqué, et n'a pas d'autre voisin. Il remonte alors à C, et passe tout de suite à B car A, B, E sont marqués. Il finit par B puis F. Sans marquage, il aurait tourné en boucle avec $A - D - E - C$.

Pour trouver toutes les composantes connexes d'un graphe, il suffit de stocker le vecteur contenant les étiquettes de marquage, de lancer DFS sur n'importe quel noeud non marqué du graphe, et de recommencer tant qu'il existe des noeuds non marqués. Dans le Chapitre 7, nous chercherons à compter le nombre de clusters d'une densité à l'aide du nombre de composantes connexes d'un certain graphe. Cet algorithme permettra non seulement de les compter, mais comme il construit aussi les composantes, nous pourrons aussi nous servir des composantes comme partition des noeuds du graphe.

Complexité

La principale propriété de cet algorithme est qu'il est rapide, ce qui n'est pas toujours le cas pour un algorithme sur des graphes. C'est ce qui fait aussi un des intérêts principaux de l'algorithme présenté dans le Chapitre 7, lui-même fondé sur DFS.

Proposition 5.9. *La complexité temporelle de DFS est au pire linéaire en le nombre d'arêtes. Sa complexité spatiale est linéaire en le nombre de noeuds.*

Par conséquent la complexité de DFS dépend notamment de la densité du graphe. Nous renvoyons par exemple à Leiserson et al. (2001) pour une preuve et d'autres détails.

Remarque. Il est arbitraire d'explorer l'arbre recouvrant en profondeur ; on pourrait très bien le parcourir en largeur aussi, avec le même procédé de marquage et en utilisant une queue [2] au lieu d'une pile. Cet algorithme s'appelle Breadth First Search, et a la même complexité que DFS.

2. Une queue est une pile FIFO, pour First In First Out, « premier arrivé, premier servi » : le premier élément empilé est traité et retiré en premier.

Deuxième partie

Contributions originales

Chapter 6

Classification and estimation in the Stochastic Blockmodel based on the empirical degrees

Résumé

Le Stochastic Blockmodel (Holland et al., 1983) est un modèle de mélange pour des données de réseaux hétérogènes. Contrairement au cadre statistique usuel, l'ajout d'un nouvel individu donne de l'information supplémentaire sur les autres individus de la population. Ainsi la distribution des degrés se concentre en un point conditionnellement à la la classe des noeuds. Nous montrons sous certaines hypothèses que la classification, l'estimation des paramètres et la sélection du nombre de classes peut en fait être réalisée entièrement avec les degrés. Nous proposons un algorithme de classification consistent et capable de traiter de très grands graphes, et des estimateurs des paramètres consistents. En particulier, nous montrons un majorant de la probabilité de l'existence d'un noeud mal classé, en particulier quand le nombre de classes croît.

Mots clés : Stochastic Blockmodel, Classification non supervisée, Estimation.

Ce chapitre est le fruit d'une collaboration avec Jean-Jacques Daudin et Stéphane Robin. Il a été publié dans Electronic Journal of Statistics, Volume 6 (2012), p.2574-2601. Notons que l'Appendice 6.B a été ajouté à l'article original. Il décrit un test de clustering et prouve sa consistence sous les hypothèses de l'article.

Abstract

The Stochastic Blockmodel (Holland et al., 1983) is a mixture model for heterogeneous network data. Unlike the usual statistical framework, new nodes give additional information about the previous ones in this model. Thereby the distribution of the degrees concentrates in points conditionally on the node class. We show under some assumptions that classification, estimation and model selection can actually be achieved with no more than the empirical degree data. We provide an algorithm able to process very large networks and consistent estimators based on it. In particular, we prove a bound of the probability of misclassification of at least one node, including when the number of classes grows.

Keywords: Stochastic Blockmodel, Unsupervised classification, Estimation.

This chapter is a joint work with Jean-Jacques Daudin and Stéphane Robin. It has been published in Electronic Journal of Statistics, Volume 6 (2012), p.2574-2601. Note that Appendix 6.B is added to the original article. It describes a clustering test and proves its consistency under the assumption of the article.

6.1 Introduction

Strong attention has recently been paid to network models in many domains such as social sciences, biology or computer science. Networks are used to represent pairwise interactions between entities. For example, sociologists are interested in observing friendships, calls and collaboration between people, companies or countries. Genomicists wonder which gene regulates which other. But the most famous examples are undoubtedly the Internet, where data traffic involves millions of routers or computers, and the World Wide Web, containing millions of pages connected by hyperlinks. A lot of other examples of real-world networks are empirically treated in Albert and Barabási (2002), and book Faust and Wasserman (1994) gives a general introduction to mathematical modelling of networks, and especially to graph theory.

One of the main features expected from graph models is inhomogeneity. Some articles, e.g. Bollobás et al. (2007) or Van Der Hofstad (2009), address this question. In the Erdős-Rényi model introduced by Erdős and Rényi (1959) and Gilbert (1959), all nodes play the same role, while most real-world networks are definitely not homogeneous.

In this paper, we are interested in the Stochastic Blockmodel (SBM), introduced by Holland et al. (1983) and inspired by Holland and Leinhardt (1981) and Fienberg and Wasserman (1981). This model assumes discrete inhomogeneity in the underlying social structure of the observed population: n nodes are split into Q homogeneous classes, called blocks, or more generally clusters. Then it is assumed that the distribution of the edge between two nodes, depends only on the blocks to which they belong. Thereby, within each class, all nodes have the same connection behavior: they are said to be structurally equivalent (Lorrain and White, 1971). When the class assignment is known, the social structure can possibly be visualized through the meta-graph (Picard et al., 2009), which emphasizes the role of each class. However the block structure is supposed to be not observed or *latent*. Thus the assignment Z and the model parameters must be estimated *a posteriori* through the observed graph X, which is a real challenge, especially in large networks.

Our main purpose in this paper is to present a consistent inference method under SBM, which can above all process very large graphs. Snijders and Nowicki (1997) have proposed a maximum likelihood estimate based on the EM algorithm for very small graphs with $Q = 2$ blocks. They have also proposed a Bayesian approach based on Gibbs sampling for larger graphs (hundreds of nodes), which they have extended to arbitrary block numbers in Nowicki and Snijders (2001). However the usual techniques enables the processing of only relatively small graphs, because they suffer severely from the complexity of graph structure. In particular the EM algorithm deals with the conditional distribution of the labels Z given

the observations X, whose dependency graph is actually a clique in the case of SBM (see paragraph 5.1 in Daudin et al. (2008)). Inspired by Wainwright and Jordan (2008), Daudin et al. (2008) have developed approximate methods using variational techniques in the context of SBM. From a physical point of view, the variational paradigm amounts to mean-field approximation, see Jaakkola (2000). Thus thousands of nodes can be processed with this variational EM algorithm. Lastly, Alain Celisse et al. (2012) proves the variational method to be consistent precisely under SBM.

All previous methods treat both classification and parameter estimation directly and at the same time. They are alternatively updated at each step of EM-based algorithms. Yet those tasks are actually not symmetrical, and moreover estimators are quite simple when Z is known. The classification — remaining the main pitfall thus far — can be completed first, and then the latent assignment Z just replaced with this classification by plug-in in order to estimate the parameters.

Searching for clusters from a graph is computationally difficult and has different meanings. Many algorithms, especially coming from physics and computer science, aim at detecting highly connected clusters, which are self-defined as optimizing some objective function. See Lancichinetti et al. (2009), Girvan and Newman (2002) and methods based on modularity in Newman (2006) and Bickel and Chen (2009). In contrast, the blocks under SBM have a model-based definition and do not necessarily have many inner connections (see examples in Daudin et al. (2008)). Therefore, most algorithms designed for community detection are generally not suitable in this context.

Bickel and Chen (2009), Choi et al. (2010), Alain Celisse et al. (2012) and Rohe et al. (2010) prove that it is asymptotically possible to uncover the latent structure of the graph Z. In this work, we additionally show under a separability assumption that it is possible to do so, just by utilizing degree data instead of the whole graph X. As a consequence, we can work with n variables instead of n^2, which makes classification computations much faster. The basic reason why so little information is needed — compared with other models with latent structure — is specific to SBM. The number of observed variables $(X_{ij})_{1\leq i,j\leq n}$ grows faster than the number of latent variables Z, therefore even marginal distributions of X concentrate very fast. Our algorithm actually expands the procedure introduced by Snijders and Nowicki (1997) when $Q = 2$. Like Bickel and Chen (2009), we provide probabilistic bounds for the occurrence of one error at least. Moreover we take the random assignment into account, even when the number of classes Q increases and the average degree vanishes. Related results are given in Choi et al. (2010) and Rohe et al. (2010). Nevertheless the bounds in these papers concern the rate of misclassified nodes instead, and do not prevent the number of errors

from growing to infinity. They also require the assignment Z to be fixed.

Furthermore a simulation study was carried out and shows that the method converges faster than expected from the theoretical bounds but slower than other existing methods. However it is much more computationally efficient, and does not require the storage of the whole adjacency matrix. For large networks, this trade-off might be necessary.

The paper is organized as follows. In Section 6.2, we begin by presenting the model we shall study and some notations are fixed. Above all a concentration property of the degree distribution is stated in paragraph 6.2.2, which will be very useful in proving the consistency of the method mentioned above. The classification algorithm (called LG) and the main results are presented in this section as well. In particular, Theorem 6.2 provides a bound of the error probability and Proposition 6.4 gives some convergence rates when the number of classes is allowed to grow. The consistency proof of the LG algorithm is provided in Section 6.3. Section 6.4 is devoted to deriving simple estimators of the parameters by plug-in and their consistency is also demonstrated. Section 6.5 adresses the issues related to the separability assumption and provides convergence rates of the LG algorithm as well. A simulation study in Section 6.6 illustrates the behavior of the LG algorithm, which is discussed afterwards. In Section 6.7, the model and the algorithm are more accurately studied. As an application, it is lastly proved that it is likewise possible to find out asymptotically the right number Q of blocks of the model. That completes the method relying just on degrees.

6.2 The Stochastic Blockmodel

6.2.1 Model

We first recall the SBM. For all integers $n \geq 1$, $[n]$ denotes the set $\{1, \ldots, n\}$. The undirected binary graphs with n nodes are defined by the pair $([n], X)$ where X is a symmetric binary square matrix of size n. X is called the adjacency matrix of the graph. Let $Q \geq 1$ be the number of blocks.

- $Z = (Z_i)_{i \in [n]}$ denotes the *latent* vector of $[Q]^n$ such that $Z_i = q$ if the node i is q-labeled. Let $\alpha = (\alpha_1, \ldots, \alpha_Q)$ be the vector of the block proportions in the whole population.

$$Z = (Z_i)_i \text{ i.i.d.} \sim \mathcal{M}(1; \alpha)$$

- Conditionally on the labels Z, the variables $\{X_{ij}, i, j \in [n]\}$ are independent Bernoulli variables. Conditionally on $\{Z_i = q, Z_j = r\}$, the parameter of X_{ij} is π_{qr}.

$$(X_{ij} | Z_i = q, Z_j = r) \sim \mathcal{B}(\pi_{qr})$$

π_{qr} is the connection probability between any q-labeled node and any r-labeled node. Noting $\pi = (\pi_{qr})_{q,r\in[Q]}$ the connection matrix, the parameters of the model are (α, π). This model will be denoted by $\mathcal{G}(n, \pi, \alpha)$. Note that in the sequel n will be often removed in the notations for the sake of simplicity.

This is a classical problem in mixture models: the block labeling is naturally not identifiable. The content of the blocks remains unchanged by permutating labels. But equivalence classes are identifiable as soon as $n \geq 2Q$, see Alain Celisse et al. (2012).

6.2.2 Degree distribution

For all $i \in [n]$, let $D_i^n = \sum_{j\neq i} X_{ij}$ the degree of the node i, that is the number of neighbors of this node.

Proposition 6.1. *For all $q \in [Q]$, let $\overline{\pi}_q = \sum_{r\in[Q]} \alpha_r \pi_{qr}$. D_i^n is a binomial distributed random variable conditionally on $Z_i = q$ with parameters $(n-1, \overline{\pi}_q)$.*

$(D_i^n)_{i\in[n]}$ is therefore a sample of a mixture of binomial distributed random variables with parameters $(n-1, \overline{\pi}_q)_{q\in[Q]}$ and proportions $(\alpha_q)_{q\in[Q]}$.

These variables are correlated. Thus we are not in the validity range of the usual algorithms for mixtures like EM. But there is only one edge shared by any pair of nodes and the degrees are consequently not heavily correlated. Using the EM algorithm would make sense for practical purposes. Nevertheless we have chosen to use a faster one-step algorithm, unlike EM which is iterative.

A concentration inequality for binomial random variables

The following inequality will be useful throughout the article. This will especially account for the fast concentration of the degree distribution. It is a straightforward consequence of Hoeffding's inequality for bounded variables.

Theorem 6.1. *(Hoeffding) Let $n \geq 1$, $p \in]0, 1[$ and $(Y_i)_{i\in[n]}$ a sequence of independent identically distributed Bernoulli random variables with parameter p. Let $S_n = \sum_{i=1}^n Y_i$. Then for all $t > 0$:*

$$P\left(\left|\frac{S_n}{n} - p\right| > t\right) \leq 2e^{-2nt^2} \tag{CCT}$$

Concentration property of the normalized degrees

Define the normalized degree of node $i \in [n]$:

$$T_i^n = \frac{D_i^n}{n-1}$$

$(T_i^n)_{i\in[n]}$ cluster around their average conditionally on the node class when n is increasing, according to (CCT):

$$P(|T_i^n - \overline{\pi}_q| > t | Z_i = q) \leq 2e^{-2nt^2} \tag{6.1}$$

Hence normalized degrees corresponding to q-labeled nodes gather around $\overline{\pi}_q$. Consequently, in the degree distribution, nodes from different classes split up into groups centered around $\overline{\pi}_q$, provided that all conditional averages $(\overline{\pi}_q)_{q\in[Q]}$ are different. From now on, we will assume that they are:

Assumption

$$\forall q, r \in [Q] \quad q \neq r \Rightarrow \overline{\pi}_q \neq \overline{\pi}_r \tag{H}$$

Also define δ the size of the smallest gap between two distinct conditional averages (Assumption (H) amounts to $\delta > 0$):

Definition 6.1.

$$\delta = \min_{q\neq r} |\overline{\pi}_q - \overline{\pi}_r|$$

Because of the concentration, a larger gap is expected between normalized degrees of nodes from different classes than nodes from the same class. The LG algorithm relies on this remark. It consists in building Q blocks by finding the $Q-1$ largest gaps formed by two consecutive normalized degrees.

The smaller δ is, the closer the degrees are and so the harder the separation of the classes between them is: δ can be regarded as separability parameter of the model. Given δ, n must be large enough so that the classes are clearly separated. This issue is explicitly discussed in Section 6.5.

Note that this assumption rules out some models, for example the case of π_{qq} equal for all q and π_{qr} equal for all $q \neq r$ and equal proportions which was studied in Decelle et al. (2011) with a physical point of view.

6.2.3 Largest Gaps Algorithm

If $(u_i)_{i\in[n]}$ is a sequence of real numbers, $(u_{(i)})_{i\in[n]}$ denotes the same sequence but sorted in increasing order.

Algorithm

– Sort the sequence of the normalized degrees in increasing order:

$$T_{(1)} \leq \cdots \leq T_{(n)}$$

– Calculate every gap between consecutive normalized degrees:

$$T_{(i+1)} - T_{(i)} \text{ for all } i \in [n-1]$$

– Find the indexes of the $Q-1$ largest gaps: $i_1 < \cdots < i_{Q-1}$, such that for all $k \in [Q-1]$ and for all $i \in [n] \setminus \{i_1, \ldots, i_{Q-1}\}$:

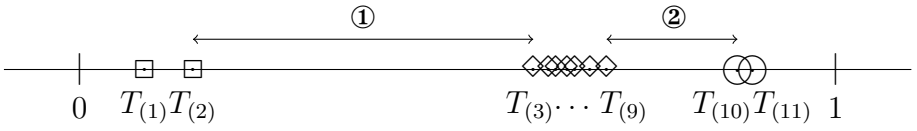
$$T_{(i_{k+1})} - T_{(i_k)} \geq T_{(i+1)} - T_{(i)}$$

– Noting $(i_0) = 0$ and $(i_Q) = n$, associate with each index (i) a class number: $i \mapsto k$ such that $(i_{k-1}) < (i) \leq (i_k)$.

Example

On the figure below, the largest gaps correspond to the intervals $[T_{(2)}, T_{(3)}[$, denoted by ①, and $[T_{(9)}, T_{(10)}[$, denoted by ②. Nodes (1) and (2) are therefore classified in class 1, nodes from (3) to (9) in 2, nodes (10) and (11) in 3.

Figure 6.1: Repartition of the normalized degrees
□: Class 1, ◇: Class 2, ◯: Class 3



This algorithm has all the qualities mentioned in Introduction and makes good use of the concentration, which makes the consistency easy to prove. Whereas variational EM algorithms runs as many quadratic steps as needed to reach convergence and classical spectral clustering runs in cubic time, this algorithm is especially fast. Indeed the sorting runs in quasilinear time and although the computation of the degrees is quadratic, this is a very basic operation which is very quickly performed. Note that Condon and Karp (2001) gave an algorithm running in linear time and consistent under SBM — called planted ℓ-partition model in this paper —, but provided that the weights of the blocks are equal.

6.2.4 Main result

The true (respectively estimated) partition of $[n]$ in classes is denoted by the set $\{\mathcal{C}_q^n\}_{q\in[Q]}$, (resp. by $\{\widehat{\mathcal{C}}_q^n\}_{q\in[Q]}$) and the cardinality of the true q-labeled class by N_q^n (resp. by $\widehat{N}_q^n$). We expect the estimated partition to be almost surely the

true partition when n is large enough. Define E_n as the event "The LG algorithm makes at least one mistake", that is:

$$E_n = \left\{ \{\widehat{\mathcal{C}}_q^n\}_q \neq \{\mathcal{C}_q^n\}_q \right\}$$

Definition 6.2. *$\{\widehat{\mathcal{C}}_q^n\}_{q\in[Q]}$ is said to be consistent if*

$$P_{\alpha,\pi}^n(E_n) \xrightarrow[n\to\infty]{} 0$$

Let us define α_0 the smallest proportion of the model:

$$\alpha_0 = \min_{q\in[Q]} \alpha_q$$

Theorem 6.2. *Under Assumption* (H),

$$P_{\alpha,\pi}^n(E_n) \leq 2ne^{-\frac{1}{8}n\delta^2} + Q(1-\alpha_0)^{n+1}$$

The proof of this theorem is given in the paragraph 6.3.3. Note that the bound is uniform over all models with the same δ, even though these do not behave exactly the same way. In particular the intraclass variability has a certain effect on the concentration of the node degrees of the class. Sparse models concentrate faster than models with a medium density for example.

6.3 Consistency proof of the LG algorithm

6.3.1 An ideal event for the algorithm

The LG algorithm delivers the true partition especially when none of the classes is empty, and the spreading of the normalized degrees is small compared with the minimal gap δ. A_n denotes the event "No true class is empty", that is

$$A_n = \bigcap_{q\in[Q]} \{\mathcal{C}_q^n \neq \varnothing\} = \bigcap_{q\in[Q]} \{N_q^n = 0\}$$

Definition 6.3. *We call maximal intraclass distance (or spreading) the random variable d_n defined by:*

$$d_n = \max_{q\in[Q]} \sup_{i\in\mathcal{C}_q^n} |T_i^n - \overline{\pi}_q|$$

This is the maximal distance between the normalized degree of a node and its own conditional mean, over all nodes and all classes. This is basically a measurement of the within-class spreading of the normalized degrees.

Proposition 6.2. *Under Assumption* (H)*, the following inclusion holds for all* $\varepsilon > 0$*:*

$$A_n \cap \left\{ d_n \leq \frac{\delta}{4+\varepsilon} \right\} \subset \overline{E}_n$$

Proof. Suppose that $A_n \cap \{d_n \leq \frac{\delta}{4+\varepsilon}\}$ is true. For all $i, j \in [n]$ and $q, r \in [Q]$:

– If nodes i and j have label q, then:

$$|T_i - T_j| \leq |T_i - \overline{\pi}_q| + |T_j - \overline{\pi}_q| \leq \frac{2\delta}{4+\varepsilon}$$

– Inversely, if they have different labels, respectively q and r, then:

$$\begin{aligned} |T_i - T_j| &\geq |T_j - \overline{\pi}_q| - |T_i - \overline{\pi}_q| \\ &\geq |T_j - \overline{\pi}_q| - \frac{\delta}{4+\varepsilon} \\ &\geq |\overline{\pi}_r - \overline{\pi}_q| - |T_j - \overline{\pi}_r| - \frac{\delta}{4+\varepsilon} \\ &\geq \delta - \frac{\delta}{4+\varepsilon} - \frac{\delta}{4+\varepsilon} = \frac{2+\varepsilon}{4+\varepsilon}\delta > \frac{2\delta}{4+\varepsilon} \end{aligned}$$

As a conclusion of this alternative, i and j are in the same class if and only if $|T_i - T_j| \leq \frac{2\delta}{4+\varepsilon}$. Notice moreover that there exists exactly $Q-1$ intervals among the set $([T_i, T_j[)_{i,j}$ strictly greater than $\frac{2\delta}{4+\varepsilon}$ on this event. Hence the $Q-1$ largest intervals lie between groups of normalized degrees from different classes; whereas all others lie between degrees of the same class. In this case the algorithm returns the true partition.

□

6.3.2 Bound of the probability of large spreading

In this paragraph we shall show that the dispersion d_n converges to 0 thanks to the subgaussian tail of the binomial distributions. This is a basic result of this article, because all others require controlling the dispersion.

Proposition 6.3. *For all* $t > 0$*:*

$$P(d_n > t) \leq 2ne^{-2nt^2}$$

Proof. It consists in conditioning by the class of each node, in order to apply the concentration inequality (CCT), and of a union bound. Since $D_i^n \sim \mathcal{B}(n, \overline{\pi}_q)$, (CCT) gave the inequality (6.1):

$$P(|T_i - \overline{\pi}_q| > t | Z_i = q) \leq 2e^{-2nt^2}$$

Hence:

$$\begin{aligned} P(d_n > t) &= \mathbb{E}\left(P(d_n > t|Z)\right) \\ &= \mathbb{E}\left(P\left(\cup_{q\in[Q]} \cup_{i\in\mathcal{C}_q} \{|T_i - \overline{\pi}_q| > t\}|Z\right)\right) \\ &\leq \mathbb{E}\left(\sum_{q\in[Q]} \sum_{i\in\mathcal{C}_q} P(|T_i - \overline{\pi}_q| > t|Z)\right) \\ &\leq \mathbb{E}\left(\sum_{q\in[Q]} \sum_{i\in\mathcal{C}_q} P(|T_i - \overline{\pi}_q| > t|Z_i = q)\right) \\ &\leq 2ne^{-2nt^2} \end{aligned}$$

□

Remark. Furthermore d_n almost surely converges to 0 because the upper bound is summable, by applying a usual consequence of the Borel-Cantelli lemma.

6.3.3 Bound of the error probability (proof of Theorem 6.2)

Thanks to the bound of the probability of large spreading, one can easily conclude that the ideal event $A_n \cap \{d \leq \frac{\delta}{4+\varepsilon}\}$ is actually strongly likely for n large enough and for all $\varepsilon > 0$:

Proof. First we have $A_n \cap \{d \leq \frac{\delta}{4+\varepsilon}\} \subset \overline{E}_n$ according to Proposition 6.2, hence:

$$P(E_n) \leq P\left(\overline{A_n \cap \{d_n \leq \frac{\delta}{4+\varepsilon}\}}\right) \leq P\left(d_n > \frac{\delta}{4+\varepsilon}\right) + P(\overline{A_n})$$

On the one hand, Proposition 6.3 implies that:

$$P\left(d_n > \frac{\delta}{4+\varepsilon}\right) \leq 2\exp\left(-2n\left(\frac{\delta}{4+\varepsilon}\right)^2\right)$$

On the other hand $\overline{A}_n$ corresponds to "There exists an empty class". For all $q \in [Q]$, $N_q \sim \mathcal{B}(n, \alpha_q)$, hence:

$$\begin{aligned} P(\overline{A}_n) &= P\left(\cup_{q\in[Q]}\{N_q = 0\}\right) \\ &\leq \sum_{q\in[Q]} P(N_q = 0) = \sum_{q\in[Q]} (1-\alpha_q)^n \leq Q(1-\alpha_0)^n. \end{aligned}$$

Once the both previous inequalities have been put together, we have an upper bound of $P(E_n)$ which depends on ε. The limit of the upper bound when ε tends to zero yields the bound of the Theorem. □

6.4 Consistency of the plug-in estimators

If the true classes were known, the usual moment estimators would be enough to estimate (α, π). Indeed the empirical proportions estimate α and the connection frequencies estimate the connection probabilities. We first prove that if we knew the classes, we would obtain a consistent estimate. However those variables are not observed but latent. That is why we plug the partition delivered by any consistent classification algorithm into these estimators. Notice that it does not depend on the choice of the consistent algorithm.

Notations For all q, r in $[Q]$, $\mathcal{C}_{qr}$ denotes $\mathcal{C}_q \times \mathcal{C}_r$, and N_{qr} its cardinality. If $q \neq r$, $N_{qr} = N_q N_r$ and if $q = r$, $N_{qq} = \frac{N_q(N_q-1)}{2}$. We define the following estimators:

$$\widetilde{\alpha}_q = \frac{N_q}{n} \text{ and } \widetilde{\pi}_{qr} = \frac{1}{N_{qr}} \sum_{(i,j)\in\mathcal{C}_{qr}} X_{ij}$$

Recall that all of these variables are hidden thus far.

6.4.1 Estimation with revealed classes

Theorem 6.3. *$(\widetilde{\alpha}, \widetilde{\pi})$ is a consistent estimator of (α, π).*

Proof. For all $q \in [Q]$, N_q is the sum of n independent Bernoulli random variables with parameter α_q. Applying directly the concentration inequality, we get for all $t > 0$ and $q \in [Q]$: $P\left(\left|\frac{N_q}{n} - \alpha_q\right| > t\right) \leq 2e^{-2nt^2}$. Applying the concentration inequality (CCT) conditionally on N_{qr} and then taking the expectation, we get for all $t > 0$:

$$P\left(|\widetilde{\pi}_{qr} - \pi_{qr}| > t\right) = \mathbb{E}\left[P\left(|\widetilde{\pi}_{qr} - \pi_{qr}| > t | N_{qr}\right)\right] \leq 2\mathbb{E}\left(e^{-2N_{qr}t^2}\right)$$

Define:

$$\alpha_{qr} = \alpha_q \alpha_r \text{ if } q \neq r \text{ and } \alpha_{qq} = \frac{\alpha_q^2}{2} \text{ if } q = r.$$

Let (r_n) be a non-negative sequence tending to infinity. We split up the support of the expectation into two pieces, depending on the values of N_{qr}. On the one hand the exponential term inside the expectation is bounded on the first piece of the support by a deterministic sequence. On the other hand, the probability of the support of the second piece of the expectation $\{|N_{qr} - \alpha_{qr}n^2| > r_n\}$ is accurately controlled by using the concentration inequality derived from (CCT) in Appendix 6.9.

$$\begin{aligned}
\mathbb{E}\left[\exp(-2N_{qr}t^2)\right] &= \mathbb{E}\left[\exp(-2N_{qr}t^2)\mathbb{1}_{\{|N_{qr}-\alpha_{qr}n^2|\leq r_n\}}\right. \\
&\qquad \left. + \exp(-2N_{qr}t^2)\mathbb{1}_{\{|N_{qr}-\alpha_{qr}n^2|>r_n\}}\right] \\
&\leq \mathbb{E}\left[\exp(-2t^2(\alpha_{qr}n^2 - r_n))\right] + P(|N_{qr} - \alpha_{qr}n^2| > r_n) \\
&\leq \exp(-2t^2(\alpha_{qr}n^2 - r_n)) + P\left(\left|\frac{N_{qr}}{n^2} - \alpha_{qr}\right| > \frac{r_n}{n^2}\right) \\
&\leq \exp\left[-r_n t^2\left(\frac{n^2\alpha_0^2}{r_n} - 1\right)\right] + 4\exp\left(-\frac{1}{2}\frac{r_n^2}{n^3}\right) \qquad \text{(B)}
\end{aligned}$$

In order to have a vanishing bound (B), we just have to choose (r_n) such that:

$$\varliminf_{n\to+\infty} \frac{\alpha_0^2 n^2}{r_n} > 1 \text{ and } \frac{r_n^2}{n^3} \xrightarrow[n\to+\infty]{} +\infty$$

For example, $r_n = n^{7/4}$, hence:

$$\mathbb{E}\left[\exp(-2N_{qr}t^2)\right] \leq \exp\left[-n^{7/4}t^2\left(n^{1/4}\alpha_0^2 - 1\right)\right] + 4\exp\left(-\frac{1}{2}\sqrt{n}\right)$$

Then we conclude with a union bound:

$$P(\|\widetilde{\pi} - \pi\|_\infty > t) \leq 2Q^2\left(e^{-n^{7/4}t^2\left(n^{1/4}\alpha_0^2-1\right)} + 4e^{-\frac{1}{2}\sqrt{n}}\right)$$

Finally we conclude for all parameters:

$$P(\|(\widetilde{\pi},\widetilde{\alpha}) - (\alpha,\pi)\|_\infty > t) \leq 2Q^2\left(e^{-n^{7/4}t^2\left(n^{1/4}\alpha_0^2-1\right)} + 4e^{-\frac{1}{2}\sqrt{n}}\right) + 2Qe^{-2nt^2}$$

□

6.4.2 Estimation with hidden classes

We now assume that we have got a partition of the nodes $\{\widehat{\mathcal{C}}_q\}_q$ returned by any classification algorithm. The estimators $\widehat{\alpha}$ and $\widehat{\pi}$ are defined by plug-in with the estimated partition $\{\widehat{\mathcal{C}}_q\}_q$ instead of the true one $\{\mathcal{C}_q\}_q$. If the classification is right, then estimators both with hat and with tilde are equal.

$$\widehat{\alpha}_q = \frac{\widehat{N}_q}{n} \text{ and } \widehat{\pi}_{qr} = \frac{1}{\widehat{N}_{qr}} \sum_{(i,j)\in\widehat{\mathcal{C}}_{qr}} X_{ij}$$

Theorem 6.4. *If $\{\widehat{\mathcal{C}}_q\}_q$ is consistent, then $(\widehat{\alpha},\widehat{\pi})$ is a consistent estimator of (α,π).*

Proof. For all $t > 0$, let $B_t^n = \{\|(\widehat{\alpha}, \widehat{\pi}) - (\alpha, \pi)\| > t\}$.

$$\begin{aligned} \forall t > 0 \quad P(B_t^n) &= P(B_t^n \cap \overline{E}_n) + P(B_t^n \cap E_n) \\ &\leq P(B_t^n \cap \overline{E}_n) + P(E_n) \end{aligned}$$

On the event $\overline{E}_n$, the equality $(\widehat{\alpha}, \widehat{\pi}) = (\widetilde{\alpha}, \widetilde{\pi})$ holds, hence:

$$\forall t > 0 \; P(B_t^n) \leq P\left(\|(\widetilde{\alpha}, \widetilde{\pi}) - (\alpha, \pi)\| > t\right) + P(E_n).$$

The first term converges to 0 according to Theorem 6.3 and the second one as well, provided the algorithm is consistent (see Theorem 6.2). □

6.4.3 Conclusions

The previous paragraphs did not depend on the algorithm chosen. Now putting together the results of the previous section and the results concerning the LG algorithm, we get:

Theorem 6.5. *For all* $t > 0$

$$\begin{aligned} P(\|(\widehat{\pi}, \widehat{\alpha}) - (\alpha, \pi)\|_\infty > t) \leq 2Q^2 \left(e^{-n^2t^2\left(\alpha_0^2 - n^{-1/4}\right)} + 4e^{-\frac{1}{2}\sqrt{n}}\right) + 2Qe^{-2nt^2} \\ + 2ne^{-\frac{1}{8}n\delta^2} + Q(1-\alpha_0)^n \end{aligned}$$

Note that the estimation procedure requires larger graphs to achieve consistency than does the classification procedure with the LG algorithm alone. This is basically due to the variability of the empirical proportions.

Since the upper bound is summable, a usual consequence of the Borel-Cantelli lemma implies the strong consistency of these estimators.

6.5 Using LG algorithm under weak separability

The case of a weak separation of the classes is now considered, that is when δ vanishes or is exactly zero.

6.5.1 Convergence rates of the LG algorithm

Here the separability parameter δ is supposed to be vanishing when n is increasing. This amounts to remove asymptotically the assumption (H). Moreover the number of classes Q is supposed to be growing with n. It is actually connected because if Q is growing and all of the $\overline{\pi}_q$ are distinct at the same time, then δ

is necessarily vanishing. Convergence rates ensuring LG to be consistent are provided for δ, Q and α_0, in order to illustrate up to where, at least, the algorithm theoretically works.

In this subsection only, another asymptotic framework is chosen. The parameters (α, π) are assumed to be functions of n. Consistency does not mean convergence under the distribution of $\mathcal{G}(n, \alpha, \pi)$ anymore, but under $\mathcal{G}(n, \alpha^n, \pi^n)$, with $\alpha^n = (\alpha_1^n, \ldots, \alpha_{Q_n}^n)$ and $\pi^n = (\pi_{qr}^n)_{1 \leq q,r \leq Q_n}$. It is assumed that:

$$\delta_n \xrightarrow[n\to\infty]{} 0,\ \alpha_0^n \xrightarrow[n\to\infty]{} 0 \text{ and } Q_n \xrightarrow[n\to\infty]{} +\infty$$

Proposition 6.4. *The classification procedure with LG algorithm is still consistent under the following assumptions:*

(a) $\underline{\lim}_{n\to+\infty}\ \delta_n\sqrt{\dfrac{n}{\ln n}} > 2\sqrt{2}$, *implying* $Q_n = O\left(\sqrt{\dfrac{n}{\ln n}}\right)$

(b) $\underline{\lim}_{n\to+\infty} -\dfrac{n\ln(1-\alpha_0^n)}{\ln Q_n} > 1$

For example, if $Q_n = 1 + \left\lfloor\sqrt{\dfrac{n}{\ln n}}\right\rfloor$, *it is sufficient that:* $\alpha_0^n \geq \dfrac{\ln n}{2n}$.

Proof. Assumption (a) implies that there exists $C > 2\sqrt{2}$ such that for n large enough:

$$\delta_n\sqrt{\frac{n}{\ln n}} \geq C \text{ and then } \frac{n\delta_n^2}{\ln n} - 8 \geq C^2 - 8 > 0$$

Therefore

$$\begin{aligned} n\exp\left(-\frac{1}{8}n\delta_n^2\right) &= \exp\left[-\frac{1}{8}\ln n\left(\frac{n\delta_n^2}{\ln n} - 8\right)\right] \\ &\leq \exp\left[-\frac{1}{8}\ln n\left(C^2 - 8\right)\right] \xrightarrow[n\to+\infty]{} 0 \end{aligned}$$

Secondly the model requires $(Q_n - 1)\delta_n \leq 1$ as a necessary condition. Hence, applying (a):

$$Q_n \leq 1 + \frac{1}{\delta_n} = O\left(\sqrt{\frac{n}{\ln n}}\right)$$

According to Assumption (b), there exists $C' > 1$ such that for n large enough:

$$-\frac{n\ln(1-\alpha_0^n)}{\ln Q_n} > C', \text{ so that:}$$

$$\begin{aligned} Q_n(1-\alpha_0^n)^n &= \exp\left[\ln Q_n + n\ln(1-\alpha_0^n)\right] \\ &= \exp\left[-\ln Q_n\left(\frac{-n\ln(1-\alpha_0^n)}{\ln Q_n} - 1\right)\right] \\ &\leq \exp\left(-\ln Q_n\,(C'-1)\right) \xrightarrow[n\to+\infty]{} 0 \end{aligned}$$

Thus it has been just proved that the two terms of the bound of the theorem 6.2 were vanishing, which finishes the proof.

□

Large graphs are more and more sparse as n increases, which results in the decrease in the connectivity defined by $\overline{\overline{\pi}}_n = \mathbb{E}_{\alpha^n,\pi^n}(T_1^n)$. Convergence rates are now given when sparsity increases.

Proposition 6.5. *The LG algorithm is still consistent in the following cases:*

1. $\overline{\overline{\pi}}_n = O\left(\left(\frac{\ln n}{n}\right)^{3/2}\right)$, *if* Q_n *is bounded.*
2. $\overline{\overline{\pi}}_n = O\left(\sqrt{\frac{\ln n}{n}}\right)$, *if* $Q_n \sim \sqrt{\frac{n}{\ln n}}$.

Proof. We sketch the proof with the following inequality, where the right hand side estimates the connectivity of the sparsest model:

$$\overline{\overline{\pi}}_n = \sum_{q=1}^{Q_n} \alpha_q \overline{\pi}_q \geq \sum_{q=1}^{Q_n} \alpha_q^n (q-1)\delta_n \geq \alpha_0^n \frac{Q_n(Q_n-1)}{2}\delta_n$$

□

6.5.2 Separation of mixed classes

In this paragraph, it is supposed to be known that two average normalized degrees are equal, so that $\delta = 0$. there are Q classes and $\overline{\pi}_q = \overline{\pi}_r$ for some q and r. For the sake of simplicity, all other conditional averages are assumed to be pairwise distinct.

The LG algorithm can be previously applied to the graph with the input parameter $Q-1$. The $Q-1$ groups returned by LG are asymptotically the true classes, except classes q and r, which are mixed together in one group of nodes, denoted by $M \subset [n]$.

We shall briefly explain a procedure to separate this group, using the concentration of some additional binomial variables, namely the number of common neighbors of each pair of nodes (or number of paths of length 2 between each pair

of nodes). Since there is a quadratic number of node pairs, this is not as fast as our procedure using degrees only.

Note that the paths of length 2 have been considered in the stochastic block model in some papers, for spectral clustering in Rohe et al. (2010) or for parameter estimates in Ambroise and Matias (2011). More general motifs are also studied in Bickel et al. (2011).

Notation. Define $\underline{\alpha}$ the diagonal matrix the diagonal coefficients of which are $(\alpha_q)_{q\in[Q]}$ and the bilinear map on $\mathbb{R}^Q$:

$$\langle \cdot, \cdot \rangle_\alpha : \ (X, Y) \mapsto {}^tX \underline{\alpha} Y$$

which is a scalar product, as soon as α_q is non-negative for all q. $\| \cdot \|_\alpha$ denotes the associated norm.

For all pairs of nodes $(i, j) \in M \times M$, define

$$D_{ij} = \sum_{k \neq i,j} Y_{ijk}, \text{ where } Y_{ijk} = X_{ik} X_{jk}.$$

Y_{ijk} is a Bernoulli distributed variable, that equals one if and only if i *and* j are both connected to k. Its parameter conditionally depends on each class of nodes i and j:

- If i and j both belong to the q-labeled class:

$$P(Y_{ijk} = 1 | Z_i = Z_j = q) = \sum_{l=1} \alpha_l \pi_{ql}^2 = \|\pi_q\|_\alpha^2$$

  where π_q is the row vector $(\pi_{ql})_l$. Symmetrically, if they both belong to the r-labeled class, the parameter is $\|\pi_r\|_\alpha^2$.
- Otherwise, if they belong to distinct classes $q \neq r$:

$$P(Y_{ijk} = 1 | Z_i = q, Z_j = r) = \sum_{l=1}^{Q} \alpha_l \pi_{ql} \pi_{rl} = \langle \pi_q, \pi_r \rangle_\alpha$$

The behavior of the new variables D_{ij} looks like that of the degrees; they once more quickly concentrate around their average value as a consequence of the concentration of binomial variables. There are three groups of node pairs, concentrating around $\|\pi_q\|_\alpha^2$, $\|\pi_r\|_\alpha^2$, or $\langle \pi_q, \pi_r \rangle_\alpha$. The first two contain only pairs of nodes of the same membership, whereas the last one is made up of pairs of nodes of different memberships.

Up to a label switch, it can be supposed that $\|\pi_q\|_\alpha \leq \|\pi_r\|_\alpha$. The following lemma shows that the group with pairs of nodes of different memberships is well separated from one of the other two. This will be sufficient to separate classes q and r.

Lemma 6.1.

$$0 \leq \langle \pi_q, \pi_r \rangle_\alpha < \|\pi_r\|_\alpha^2$$

Proof. First of all $\langle \pi_q, \pi_r \rangle_\alpha \geq 0$ because this was defined as a probability. Then, by applying the Cauchy-Schwarz inequality:

$$\langle \pi_q, \pi_r \rangle_\alpha \leq \|\pi_q\|_\alpha \|\pi_r\|_\alpha \leq \|\pi_r\|_\alpha^2$$

The case of equality in the Cauchy-Schwarz inequality cannot arise; if it did, then π_q and π_r would be collinear vectors. Noting c the constant of collinearity, it would yield $\overline{\pi}_q = c\overline{\pi}_r$. But $\overline{\pi}_q$ and $\overline{\pi}_r$ are assumed to be equal in this section; hence $c = 1$. π_q and π_r would be equal. This is not allowed by the model for identifiability reasons. The inequality is eventually strict. □

Now the LG algorithm is applied to the set of variables $(D_{ij})_{i,j \in M}$ with $Q = 2$ as input parameter. Define W as the set of the pairs which are returned in the second group — the groups being sorted in increasing order — and F as the set of nodes, which are involved in those pairs. Let K be the graph defined by (F, W).

Note that K has no obvious relation to the observed graph X. An edge between $i \in M$ and $j \in M$ just means that the pair (i, j) has been classified in the second group by LG.

Proposition 6.6. *In the graph K there are edges only between nodes from the same class with high probability when n is large enough. As a consequence K is asymptotically made of one or two cliques and each clique of K is made of all nodes from either class q or class r.*

Proof. There are two major cases, depending on the relative position of $\|\pi_q\|_\alpha^2$, $\|\pi_r\|_\alpha^2$ and $\langle \pi_q, \pi_r \rangle_\alpha$.

- If $\|\pi_q\|_\alpha^2 \leq \langle \pi_q, \pi_r \rangle_\alpha < \|\pi_r\|_\alpha^2$, the gap between $\|\pi_q\|_\alpha^2$ and $\langle \pi_q, \pi_r \rangle_\alpha$ is actually strictly smaller than the gap between $\langle \pi_q, \pi_r \rangle_\alpha$ and $\|\pi_q\|_\alpha^2$:

$$\begin{aligned} \|\pi_r\|_\alpha^2 - \langle \pi_q, \pi_r \rangle_\alpha - \left(\langle \pi_q, \pi_r \rangle_\alpha - \|\pi_q\|_\alpha^2 \right) \\ = \|\pi_q\|_\alpha^2 + \|\pi_r\|_\alpha^2 - 2\langle \pi_q, \pi_r \rangle_\alpha = \|\pi_q - \pi_r\|_\alpha^2 > 0 \end{aligned}$$

  As a consequence, LG selects asymptotically the gap between $\langle \pi_q, \pi_r \rangle_\alpha$ and $\|\pi_r\|_\alpha^2$ as the largest one. Then the second group returned by LG is asymptotically made up of the node pairs concentrated around $\|\pi_r\|_\alpha^2$, i.e. the pairs of nodes from class r. K forms asymptotically one clique, which is made up of all nodes from class r.
- If $\langle \pi_q, \pi_r \rangle_\alpha < \|\pi_q\|_\alpha^2 \leq \|\pi_r\|_\alpha^2$, LG selects asymptotically either the gap between $\|\pi_q\|_\alpha^2$ and $\|\pi_r\|_\alpha^2$ and then there is only one clique as in the previous case, or the gap between $\langle \pi_q, \pi_r \rangle_\alpha$ and $\|\pi_q\|_\alpha^2$ and then the second group returned is made up of the node pairs concentrated around $\|\pi_q\|_\alpha^2$ and $\|\pi_r\|_\alpha^2$. There are two cliques and each one corresponds to one class.

□

Since the content of one of the two classes is known, the node group M which contains nodes with mixed memberships can be separated for large enough n.

Remark. Here it is supposed to be known that $\overline{\pi}_q = \overline{\pi}_r$. However we do not provide here any procedure to know if the averages degrees are really equal. Further developments would be needed to test this hypothesis, using the size of the tail of the observed distribution of the variables $(D_{ij})_{ij\in M}$ for instance. Indeed these variables concentrate around only one value when there is only one class, and around several values when there are more than one class.

6.6 Simulation study

Our main purpose in this study is to figure out how the LG algorithm behaves in practice, and above all, to check whether the bounds of Theorem 6.2 are pessimistic or not. The empirical frequency of the graphs with no error would be of great interest, because that is the quantity the bound concerns. But actually this frequency has no smooth evolution: it suddenly shifts from 0 to almost 1. We shall use two types of classification error rates: a global one and one for each class, so as to examine more accurately the results given by the algorithm.

Moreover the results of LG are compared with these of the variational method (Daudin et al., 2008), which is available online in the packages MixNet [1], MixeR [2] and WMixnet [3]. The latter has been chosen in the current simulation study. In WMixnet the variational EM-algorithm (VEM) is initialized by a spectral clustering algorithm (Rohe et al., 2010). VEM can be additionally run several times with multiple reinitializations in order to prevent from getting caught in a local maximum. WMixnet also proproses a smoothing option working the following way. VEM algorithm is run with several values of Q. As soon as the likelihood is nonincreasing or the ICL criterion is not convex with respect to Q, the VEM is run once more for the problematic values of Q. It is basically reinitialized with the classification returned by VEM either for $Q-1$ classes after having split one class or for $Q+1$ classes after having merged two classes.

The results will be given with and without smoothing.

1. See at http://stat.genopole.cnrs.fr/logiciels/mixnet .
2. See at http://cran.r-project.org/web/packages/mixer/index.html .
3. See at http://ssbgroup.fr/mixnet/wmixnet.html .

6.6.1 Simulation design

The parameters used in the simulation are:

$$\alpha = (0.3\ 0.55\ 0.15) \quad \pi = \begin{pmatrix} 0.03 & 0.02 & 0.045 \\ 0.02 & 0.05 & 0.09 \\ 0.045 & 0.09 & 0.25 \end{pmatrix}$$

Hence $\overline{\pi} = (0.0267\ 0.047\ 0.1005)$ and $\delta = 0.0203$. The parameters have been chosen so that the graphs are relatively sparse.

200 graphs are drawn from the model $\mathcal{G}(n, \alpha, \pi)$ for n from 200 to 11000. Then both LG and WMixnet are applied to each graph so as to obtain the node classification and the parameter estimators. Above 3000 nodes (respectively 5600) WMixnet with smoothing (resp. without) turned out too slow to be run in reasonable time. However it has already converged from $n = 2200$ nodes (resp. $n = 5200$).

The evolutions of the classification error rates and the estimators with respect to the number of nodes n are averaged over the 200 graphs and displayed from 200 to 11000 nodes for the LG algorithm, and to 5600 nodes for the variational method.

Error rates First of all, the global error rate g_n is defined as the proportion of node pairs (i, j), either classified in distinct classes whereas their true labels are identical, or classified together whereas their true labels are different. That is, denoting $\widehat{Z}$ the label vector returned by the classification algorithm:

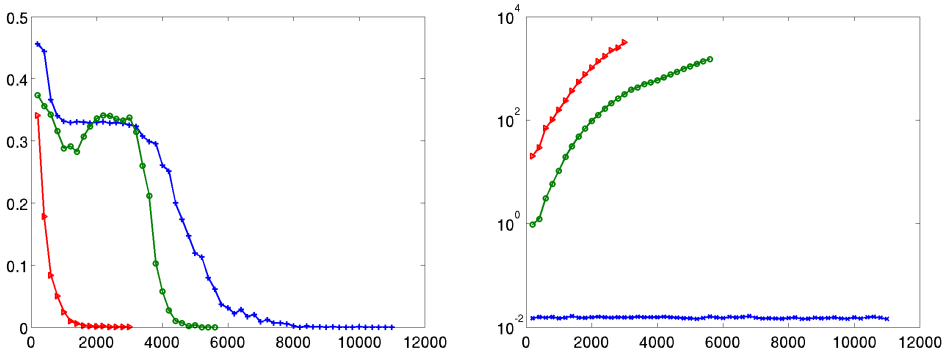
$$g_n(Z, \widehat{Z}) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \left(\mathbb{1}_{Z_i = Z_j} \mathbb{1}_{\widehat{Z_i} \neq \widehat{Z_j}} + \mathbb{1}_{Z_i \neq Z_j} \mathbb{1}_{\widehat{Z_i} = \widehat{Z_j}} \right)$$

Secondly, we also propose error rates per class. Define I_q, resp. M_q, the rate of intruders (or false positive rate) in the class q predicted by the algorithm, resp. the rate of missing nodes of the true class q (or false negative rate):

$$I_q^n(Z, \widehat{Z}) = \frac{1}{\widehat{N}_q} \sum_{i \in \widehat{\mathcal{C}}_q} \mathbb{1}_{Z_i \neq q} \text{ and } M_q^n(Z, \widehat{Z}) = \frac{1}{N_q} \sum_{i \in \mathcal{C}_q} \mathbb{1}_{\widehat{Z}_i \neq q}$$

blackLabels will be allocated to the nodes in order of increasing degree in the classification algorithms. Indeed the true labels are expected to be sorted this way, because $\overline{\pi}_1 < \overline{\pi}_2 < \overline{\pi}_3$. This partially solves the label switching problem which arises when trying to identify the true labels instead of the equivalence classes.

Figure 1: Error rates g_n and running time as functions of the graph size n
$+$: LG ; $\circ$: WM ; $\triangleright$: WMS



Number of nodes

6.6.2 Results

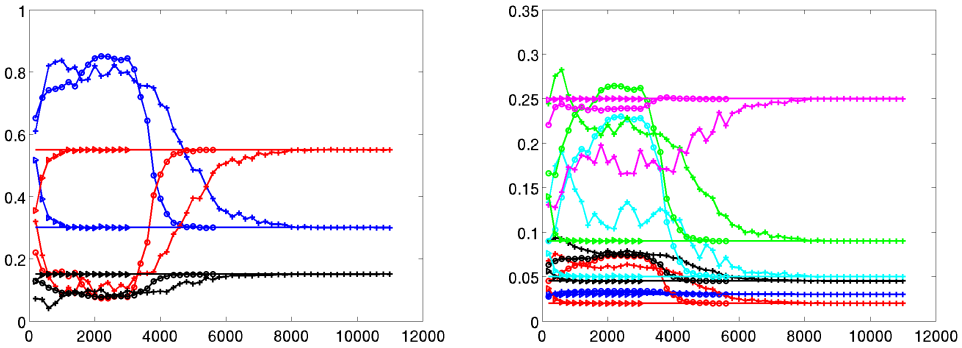
The evolution is quite satisfactory because the global error rate g_n of LG completely vanishes from $n = 8600$ nodes, which is even earlier than expected from the bound of Theorem 6.2. Indeed this bound predicted that the probability of at least one error would not be less than 0.05 earlier than $n = 300000$. The bound seems to be pessimistic, basically because of the union bound, used in the proof of Proposition 6.3. Note nevertheless that (see also the remark following the theorem 6.2) since the model is relatively sparse, the classification is not as intricate as for models with medium density. For instance we have also tried a model with $\delta = 0.02$ and average normalized degrees close to 0.6, and the global error rate vanished only from $n = 40000$ nodes (not shown here).

WMixnet with smoothing (WMS) converges so fast that its error rate completely vanished from $n = 2400$ nodes, much earlier than LG. Up to $n = 3000$, both WMixnet without smoothing (WM) and LG return poor and very similar results. Then the error rate of WM suddenly vanishes from $n = 5200$ nodes. Thus there is a gap between $n = 5600$ and $n = 8600$ where WMixnet is hardly usable and LG does not provide good results.

The running time of LG seems to be constant with respect to n, because the asymptotical regime (quasilinear) has not been reached yet, whereas these of the WMixnet algorithms are dramatically increasing.

Transitional phase of LG Now the behavior of LG alone is more accurately discussed. After a dramatic decrease of the error rate of LG at the beginning, its evolution encounters a slight stagnation between $n = 1000$ and $n = 3000$ nodes

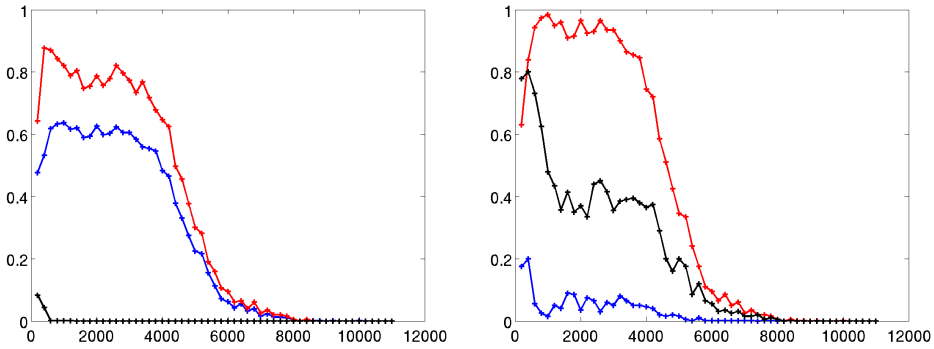
Figure 2: Means of the estimators
$+$: LG ; $\circ$: WM ; $\triangleright$: WMS
— : Class 1 ; — : Class 2 ; — : Class 3 — : 1-1 ; — : 1-2 ; — : 1-3
— : 2-2 ; — : 2-3 ; — : 3-3



Number of nodes

(see Figure . An interpretation of this transitional phase of LG is given using the error rates per class.

Figure 3: Error rates I_q^n and M_q^n as functions of n for LG
— : Class 1 ; — : Class 2 ; — : Class 3



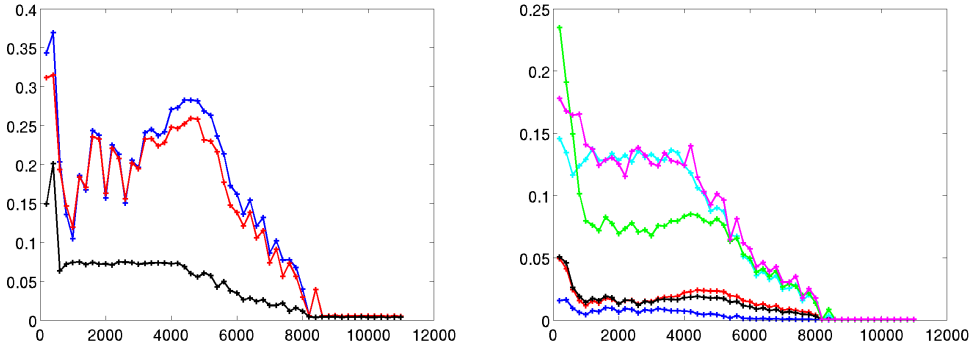
Number of nodes

The third class is much better detected even at small graph sizes, unlike class 1 and class 2. Indeed it is sufficient that the maximal intraclass distance d_n is less than $(\overline{\pi}_3 - \overline{\pi}_2)/4$ to detect this class, whereas the other two are not supposed to be separated before

$$d_n < \frac{\overline{\pi}_2 - \overline{\pi}_1}{4} = \frac{\delta}{4} < \frac{\overline{\pi}_3 - \overline{\pi}_2}{4}$$

according to our previous study. That is the reason why the global error rate dramatically decreases until reaching $n = 1000$ nodes, and why it does not decrease

Figure 4: Standard deviations of the estimators for LG
— : Class 1 ; — : Class 2 ; — : Class 3 — : 1-1 ; — : 1-2 ; — : 1-3
— : 2-2 ; — : 2-3 ; — : 3-3



Number of nodes

anymore before reaching $n = 3000$. Note that the bound of Theorems 6.3 and 6.2 had not predicted this before reaching $n = 39000$ and $n = 317000$ respectively.

In short, as long as the tails of the normalized degree distribution are overlapping, the classes are mixed and cannot be properly detected. The curves show in particular that many nodes of class 2 seem to be caught by class 1, since there are many missing nodes in class 2 (the biggest class) and many intruders in class 1. The missing nodes of class 3 must be caught by class 2 as well. As a consequence, the proportion of classes 2 and 3 are underestimated in the transitional phase, whereas the proportion of class 1 is overestimated. Thus the inversion of classes 1 and 2 is clearly shown again on the graphic of the proportions estimates (see Figure 2). The graphic of the connectivities estimates (see Figure 2 and Figure 4) also shows that as long as there is a lot of missing nodes in a class, the intraconnectivity estimator of the class is not good (see the curve of 2-2 and 3-3, unlike 1-1). However the interconnectivity 1-2 is well estimated, because nodes from class 1 and 2 must be often permuted.

6.6.3 Conclusions

As a conclusion of this practical comparison, the LG algorithm should be used only for very large graphs, when nothing else is possible. LG can deal with millions of nodes on the same computers we used for the current simulation. For small graphs, other techniques provide better results.

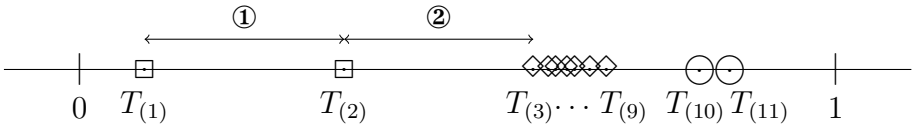
This algorithm lacks robustness because it takes every normalized degree into account and each one carries the same weight, even if it is isolated and not statistically representative. In the worst case, one untypical node is sufficient to trick the

algorithm, making the classification wrong by a majority (see Figure 6.2). This often arises at small graph size in generated data and may occur at any size also in real data.

Example

On the figure below, the largest gaps correspond to the intervals $[T_{(1)}, T_{(2)}[$, denoted by ①, and $[T_{(2)}, T_{(3)}[$, denoted by ②. Node (1) is classified in class 1, (2) in class 2, and all remaining nodes from (3) to (11) in class 3. Just because of the misplacement of $T_{(1)}$ or $T_{(2)}$, the real class 3 (circled nodes) is not anymore separated from the real class 2 (diamond nodes).

Figure 6.2: Repartition of the normalized degrees
□: Class 1, ◇: Class 2, ◯: Class 3



6.7 Model selection

Up to this section, the number of classes was supposed to be known and was an input parameter of the LG algorithm. Our main purpose hereafter is to examine more accurately the sequence of the gaps sorted in increasing order and then the sequence of the intervals between the means of the groups given by the LG algorithm, depending on the selected number of classes Q for the model. As an application of this study, we finally show that degrees are once more sufficient to select the right number of classes for large enough n.

6.7.1 Study of the gap sequence

We will use the same notations as in the last section. Moreover Q_0 denotes the true number of classes, and Q the current input parameter of the LG algorithm. We will often use the event $B_n = A_n \cap \{d_n \leq \frac{\delta}{5}\}$, where no class is empty and the dispersion d_n is so small that the $Q_0 - 1$ largest intervals separate the true classes (see Proposition 6.2 with $\varepsilon = 1$). Then we can affirm that two normalized degrees are in the same class if and only if their distance is less than $2d_n$.

Let $(G_q^n)_{q\in[n-1]}$ be the sequence of the distances between consecutive normalized degrees $(T_{(i+1)}^n - T_{(i)}^n)_{i\in[n-1]}$, but sorted in decreasing order:

$$G_1^n \geq G_2^n \geq \cdots \geq G_{n-1}^n$$

The $Q_0 - 1$ largest gaps in the LG algorithm have lengths $G_1, \ldots, G_{Q_0-1}$. Define also $(\gamma_q)_{q\in[Q_0-1]}$ the sequence $(\overline{\pi}_{(q+1)} - \overline{\pi}_{(q)})_{q\in[Q_0-1]}$, sorted in decreasing order. This is called the sequence of the theoretical gaps. The following theorem states that largest empirical gaps converge to the corresponding theoretical gaps, which enforces our intuition about the model.

Theorem 6.6. *For all $q < Q_0$, $G_q \xrightarrow[n\to+\infty]{} \gamma_q$ a.s.*

Refer to Appendix 6.A to see the proof. One can easily realize that the only gap (among the $Q_0 - 1$ largest) lying between $\overline{\pi}_{(q)}$ and $\overline{\pi}_{(q+1)}$ converges to $\overline{\pi}_{(q+1)} - \overline{\pi}_{(q)}$. However the index of this interval is random and depends on n. This interesting but technical problem is solved in the second part of the proof. For the moment we provide a weaker version of this theorem, the proof of which is much simpler. Its conclusion is sufficient for our purposes.

Theorem 6.7. *For all $q < Q_0$, $\underline{\lim}_{n\to+\infty} G_q > 0$*

Proof. If $q < Q_0$: on the event B_n, the $Q_0 - 1$ largest intervals necessarily lie between normalized degrees from different classes. There exists $i \in \mathcal{C}_r$ and $j \in \mathcal{C}_s$, where $s \neq r$ such that $G_q = |T_i - T_j|$. But $|T_i - \overline{\pi}_r| \leq d_n$ and $|T_j - \overline{\pi}_s| \leq d_n$, hence

$$G_q \geq |\overline{\pi}_r - \overline{\pi}_s| - 2d_n \geq \delta - \frac{2}{5}\delta = \frac{3}{5}\delta > 0$$

Namely $B_n \subset \{G_q \geq \frac{3}{5}\delta\}$.

$$P\left(G_q < \frac{3}{5}\delta\right) \leq P\left(\overline{B}_n\right) \leq 2e^{-\frac{2}{25}n\delta^2} + Q_0(1-\alpha_0)^n$$

As the upper bound is summable, according to the Borel-Cantelli lemma,

$$P\left(\overline{\lim_{n\to+\infty}}\{G_q < \frac{3}{5}\delta\}\right) = 0$$

Therefore $\underline{\lim}_{n\to+\infty} G_q \geq \frac{3}{5}\delta > 0$ almost surely.

□

All further gaps lie between degrees of nodes of the same class and then converge to zero. The next theorem gives an estimation of the convergence rate.

Theorem 6.8. *For all $\beta \in]0,1[$, the triangular array*

$$\{n^{\frac{1-\beta}{2}} G_q^n ; Q_0 \leq q \leq n-1\}$$

converges uniformly w.r.t. q and a.s. to zero when n tends to infinity.

Proof. First of all, recall that for all n,

$$G_{Q_0}^n \geq G_{Q_0+1}^n \geq \cdots \geq G_{n-1}^n \geq 0$$

Therefore we can just prove that $n^{\frac{1-\beta}{2}} G_{Q_0} \xrightarrow[n\to+\infty]{} 0$, and the uniform convergence will follow.

On the event B_n, the $Q_0 - 1$ largest intervals lie between normalized degrees from different classes. The next intervals lie between degrees from the same class, and the distance to their corresponding conditional mean is at most d_n. As G_{Q_0} is one of these, $G_{Q_0} \leq 2d_n$. Hence, for all $0 < t < \frac{\delta}{5}$:

$$\begin{aligned} P\left(n^{\frac{1-\beta}{2}} G_{Q_0} > t\right) &= P(n^{\frac{1-\beta}{2}} G_{Q_0} > t \cap B_n) + P\left(n^{\frac{1-\beta}{2}} G_{Q_0} > t \cap \overline{B}_n\right) \\ &\leq P(2n^{\frac{1-\beta}{2}} d_n > t) + P\left(\overline{B}_n\right) \\ &\leq 2(e^{-\frac{1}{2}n^\beta t^2} + e^{-\frac{2}{25}n\delta^2}) + Q_0(1-\alpha_0)^n \end{aligned}$$

□

6.7.2 Study of the intervals between estimated classes

By distances between estimated classes, we mean distances between empirical averages of the normalized degrees of each class, provided by the LG algorithm. Define m_q to be the average of the normalized degrees of the q-labeled class estimated by the algorithm:

$$m_q = \frac{1}{N_q} \sum_{i \in \widehat{\mathcal{C}}_q} T_i$$

The sequence of the gaps between consecutive averages $(m_{(q+1)} - m_{(q)})_{q\in[Q-1]}$ is sorted in order of decreasing length, just as the sequence of the gaps $(T_{(i+1)} - T_{(i)})_{i\in[n-1]}$ is in the previous paragraph. This new sequence is denoted by $(H_q^n)_{q\in[Q-1]}$. Of course it depends on the current Q, whereas $(G_q)_q$ does not.

When $Q = Q_0$, H_q and G_q are very close for all $q \leq Q_0 - 1$. On the contrary, when $Q < Q_0$, some of the $(H_q)_{q\in[Q_0-1]}$ stretch over several classes and include more than one of the G_q. As a result, there is at least one q such that H_q differs from G_q for large enough n.

Theorem 6.9.

1. *If* $Q = Q_0$, *then* $\sum_{q=1}^{Q-1} (H_q - G_q) \xrightarrow[n\to+\infty]{a.s.} 0$
2. *If* $Q < Q_0$, *then* $\varliminf_{n\to+\infty} \sum_{q=1}^{Q-1} (H_q - G_q) > 0$ *a.s.*

Proof. Let $(J_q)_{q\in[Q_0-1]}$ the $Q_0 - 1$ largest intervals between consecutive normalized degrees, hence for all q, $|J_q| = G_q$. Define also $J'_0 = [0, \min_{i\in[n]} T_i[$ and $J'_Q = [\max_{i\in[n]} T_i, 1[$. The union of $J'_0, J_1, \ldots, J_{Q-1}, J'_Q$ partially covers the interval $[0, 1[$. These intervals are separated and the distance between the bounds of consecutive intervals is at most $2d_n$. As a result:

$$1 - 2Q_0 d_n \leq \sum_{q=1}^{Q_0-1} G_q + H_0 + H_Q \leq 1 = \sum_{q=0}^{Q} H_q$$

$Q = Q_0$ Subtracting the right-hand side (which actually equals 1), we deduce from both previous inequalities that:

$$-2Q_0 d_n \leq \sum_{q=1}^{Q_0-1} (G_q - H_q) \leq 0$$

The first assertion follows directly from this inequality; for all $t > 0$:

$$P\left(\left|\sum_{q=1}^{Q_0-1} (H_q - G_q)\right| > t\right) \leq P(2Q_0 d_n > t)$$
$$\leq 2\exp\left(-2n\left(\frac{t}{2Q_0}\right)^2\right) = 2\exp\left(-\frac{1}{2Q_0^2} n t^2\right)$$

$Q < Q_0$ Subtracting the right-hand side from the second inequality only yields this time:

$$\sum_{q=Q}^{Q_0-1} G_q \leq \sum_{q=1}^{Q-1} (H_q - G_q)$$

But as shown in Theorem 6.7, the lower limit of G_q is non-negative for all $q \leq Q_0 - 1$. *A fortiori*, the second assertion of the theorem 6.9 stands as well.

□

6.7.3 Application to model selection

The summed differences $\sum_{q=1}^{Q-1}(H_q - G_q)$ examined in the last paragraph have an interesting property regarding model selection: when Q is the right number of classes, it converges to zero, and when Q is too small, it converges to a non-negative value, because one of the H_q does not match G_q. Thus this quantity measures the risk of underestimating the number of classes.

However, its minimization over all $Q \in \{2, \dots, n\}$ yields the unexpected solution $Q = n$, for all Q_0. Therefore we have to penalize overly small gaps between normalized degrees. We chose to use an *ad hoc* penalty, that can be easily inferred from our previous study, in order to have a correct estimate of Q_0. Define for all $Q \in \{2, \dots, n\}$:

$$f_Q = \sum_{q=1}^{Q-1}(H_q - G_q) + \frac{1}{n^{\frac{1-\beta}{2}} G_{Q-1}} \in [0, +\infty] \text{ where } \beta \in]0, 1[.$$

Theorem 6.10.

1. *If* $Q = Q_0$, *then* $f_Q \xrightarrow[n\to+\infty]{a.s.} 0$
2. *If* $Q < Q_0$, *then* $\varliminf_{n\to+\infty} f_Q > 0$ *a.s.*
3. *If* $Q > Q_0$, *then* $f_Q \xrightarrow[n\to+\infty]{a.s.} +\infty$

It follows that $\widehat{Q} = \arg\min_{2\le Q\le n} f_Q \xrightarrow[n\to+\infty]{} Q_0$ *a.s.*

Proof. **If** $Q = Q_0$ Applying Theorem 6.9, the sum $\sum_{q=1}^{Q-1}(H_q - G_q)$ converges a.s. to 0. According to Theorem 6.7, $\varliminf_{n\to+\infty} G_{Q_0-1} > 0$ almost surely. Therefore:

$$\frac{1}{n^{\frac{1-\beta}{2}} G_{Q-1}} \xrightarrow[n\to+\infty]{a.s.} 0, \text{ and then } f_Q \xrightarrow[n\to+\infty]{a.s.} 0$$

If $Q < Q_0$ According to the second assertion of Theorem 6.9, the lower limit of the first term is non-negative. There is no change by adding the second term, because it is positive. Hence:

$$\varliminf_{n\to+\infty} f_Q > 0$$

If $Q > Q_0$ The sum $\sum_{q=1}^{Q-1} H_q - G_q$ is lower bounded by -1 (notice that it is even positive), and according to the second assertion of Theorem 6.8, $(n^{\frac{1-\beta}{2}} G_{Q-1})_n$ uniformly converges to 0, as soon as $q \ge Q_0$. The last assertion follows.

□

6.8 Conclusions

Unlike most of the methods known thus far, the LG algorithm is able to process very large graphs. In fact it provides good results only for such graphs. Nevertheless, according to the simulation study, the algorithm is efficient even for smaller graphs than theoretically expected. Moreover it is self-sufficient: it provides consistent methods for node clustering, parameter estimation and model selection. It performs every task using the degree data alone. Lastly, this algorithm is free from any preliminary setting. There is need neither for any prior knowledge nor for multiple runnings of the algorithm. Thus it can quickly provide initialization values for other algorithms which depend severely on them.

However other techniques provide better results for small graph size and it does not seem to be a practical method for real data, because of the lack of robustness above all.

As a conclusion, the LG algorithm is a good theoretical tool which proves this statement: for large enough n, when the average degrees are separated enough, the degree data alone is a sufficient statistics to achieve all of the statistical inference under SBM.

Acknowledgements We thank the reviewers for their helpful comments and remarks, and we also thank Jean-Benoist Léger for the help to use his package WMixnet.

6.9 Concentration inequality for products of binomial distributed variables

Proposition 6.7. *Let X (respectively Y) be a sum of n independent bernoulli distributed variables with parameter p, respectively q. Then for all $t > 0$*

$$P\left(\left|\frac{XY}{n^2} - pq\right| > t\right) \leq 4\exp(-\frac{1}{2}nt^2)$$

Proof.

$$\begin{aligned} P\left(\left|\frac{XY}{n^2} - pq\right| > t\right) &= P\left(\left|\left(\frac{X}{n} - p\right)\frac{Y}{n} + \left(\frac{Y}{n} - q\right)p\right| > t\right) \\ &\leq P\left(\left|\frac{X}{n} - p\right|\frac{Y}{n} > \frac{t}{2}\right) + P\left(\left|\frac{Y}{n} - q\right|p > \frac{t}{2}\right) \\ &\leq P\left(\left|\frac{X}{n} - p\right| > \frac{t}{2}\right) + P\left(\left|\frac{Y}{n} - q\right| > \frac{t}{2}\right) \\ &\leq 2 \times 2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) = 4\exp(-\frac{1}{2}nt^2) \end{aligned}$$

The last line is obtained by applying the usual concentration inequality (CCT) to both X and Y. □

With a similar proof, we prove that for all $t \in]0, 1/4]$:

$$P\left(\left|\frac{X(X-1)}{2n^2} - \frac{\alpha^2}{2}\right| > t\right) \leq 4\exp\left(-2nt^2\right)$$

Appendix

6.A Proof of Theorem 6.6

Let us define $(J_i)_{i\in[n]}$ the sequence of the intervals $[T_{(i)}, T_{(i+1)}[$ sorted in order of decreasing length, hence for all $i \in [n]$, $|J_i| = G_i$. We suppose hereafter that the sequence $(\overline{\pi}_q)_q$ is sorted in increasing order: $\overline{\pi}_1 < \cdots < \overline{\pi}_Q$.

Proof. On the event B_n, among the $Q_0 - 1$ largest intervals, we can associate with each $\overline{\pi}_q$ the only one lying between $\overline{\pi}_q$ and $\overline{\pi}_{q+1}$. Namely the only J_i with $i \in [Q_0 - 1]$ such that $J_i \cap]\overline{\pi}_q, \overline{\pi}_{q+1}[\neq \varnothing$. $S(q)$ denotes the index in $[Q_0 - 1]$ corresponding to this unique interval.

Moreover, $s(q)$ denotes one of the indexes $s \in [Q_0 - 1]$ such that $\gamma_s = \overline{\pi}_{q+1} - \overline{\pi}_q$, chosen so that s is injective. Let us point out that S is a random permutation whereas s is deterministic. In order to simplify notations, we silently make the deterministic index change $r = s(q)$. Thereby $(\gamma_q)_q$ still denotes the sequence $(\gamma_{s(q)})_q$, and S the permutation $S \circ s^{-1}$.

Notice that on B_n and especially when $d_n \leq \frac{\delta}{5}$:

$$[\overline{\pi}_q + d_n, \overline{\pi}_{q+1} - d_n] \subset J_{S(q)} \subset [\overline{\pi}_q - d_n, \overline{\pi}_{q+1} + d_n]$$
$$\text{Hence } |G_{S(q)} - \gamma_q| \leq 2d_n. \tag{6.2}$$

1. We first prove that the gap $G_{S(q)}$ converges to the theoretical gap γ_q. For all $t > 0$:

$$\begin{aligned} P(|G_{S(q)} - \gamma_q| > t) &= P(|G_{S(q)} - \gamma_q| > t \cap B_n) + P(|G_{S(q)} - \gamma_q| > t \cap \overline{B}_n) \\ &\leq P(2d_n > t) + P(\overline{B}_n) \\ &\leq 2(e^{-\frac{1}{2}nt^2} + e^{-\frac{2}{25}n\delta^2}) + Q_0(1 - \alpha_0)^n \end{aligned} \tag{6.3}$$

2. Secondly, none of the $Q_0 - 1$ largest intervals permute anymore expect for those having the same theoretical values. It follows from the inequality (6.2) that for all $q, r \in [Q_0 - 1]$,

$$\gamma_q - \gamma_r - 4d_n \leq G_{S(q)} - G_{S(r)} \leq \gamma_q - \gamma_r + 4d_n$$

Define $\eta = \frac{1}{5}(\min_{q\in[Q]}(\gamma_q - \gamma_{q+1}) \wedge \delta)$, a threshold designed to distinguish distances converging to one value from those converging to another. On the event $d_n \leq \eta$, the previous inequality yields:

$$\gamma_q - \gamma_r - 4\eta \leq G_{S(q)} - G_{S(r)} \leq \gamma_q - \gamma_r + 4\eta$$

- If $\gamma_q - \gamma_r < 0$, then $\gamma_q - \gamma_r + 4\eta < 0$ is also true by the definition of η. As a result of the inequality just above, $G_{S(q)} - G_{S(r)} < 0$.
- If $\gamma_q - \gamma_r > 0$, then $\gamma_q - \gamma_r - 4\eta > 0$, and $G_{S(q)} - G_{S(r)} > 0$.

If $(u_i)_{1\leq i\leq m}$ is a sequence, we write $i \sim_u j$ if and only if $u_i = u_j$. $\sim_u$ is an equivalence relation. Applying the Lemma 6.2 stated and proved afterwards, if $d_n \leq \eta$, there exists $r \sim_\gamma q$ such that $q = S(r)$. Notice furthermore that the sequence $(\gamma_q)_{q\in[Q_0-1]}$ is constant on the $\sim_\gamma$-equivalence classes. The term $|G_q - \gamma_q|$ is necessarily in the sum $\sum_{r\sim q} |G_{S(r)} - \gamma_r|$. Finally, define

$$\begin{aligned}
P(|G_q - \gamma_q| > t) &= P(|G_q - \gamma_q| > t \cap B_n) + P(|G_q - \gamma_q| > t \cap \overline{B}_n) \\
&\leq P\left(\sum_{r\sim q} |G_{S(r)} - \gamma_r| > t\right) + P(\overline{B}_n) \\
&\leq \sum_{r\sim q} P\left(|G_{S(r)} - \gamma_r| > \frac{t}{Q_0}\right) + P(\overline{B}_n) \\
&\leq 2Q_0(e^{-\frac{1}{2Q_0^2}nt^2} + e^{-\frac{2}{25}n\delta^2}) + 2e^{-2n\eta^2} \text{ according to (6.3).}
\end{aligned}$$

□

Lemma 6.2. *Let $(u_i)_{1\leq i\leq m}, (v_i)_{1\leq i\leq m}$ be two real decreasing sequences. Let p be the number of $\sim_u$-equivalence classes and σ one permutation of $\{1,\dots,m\}$. We especially assume that for all $i, j \in \{1,\dots,m\}$,*

- *$u_i < u_j \Rightarrow v_{\sigma(i)} < v_{\sigma(j)}$*
- *$u_i > u_j \Rightarrow v_{\sigma(i)} > v_{\sigma(j)}$*

Then $\sigma = \sigma_1 \circ \cdots \circ \sigma_p$ where the support of σ_i is the i^{th} $\sim_u$-equivalence class.

Proof. Since u is decreasing, the $\sim_u$-equivalence classes are just sets of consecutive natural integers. Define recursively $(r_i)_{1\leq i\leq p}$ the increasing sequence of indexes j when the value of u_j changes:

- Let $r_1 = 1$.
- For $i \geq 1$, let r_{i+1} be the smallest integer $j > r_i$ such that $u_{r_i} = \cdots = u_{j-1} > u_j$.

The construction of $(r_i)_i$ implies that for all $j < r_i$, all $r_i \leq l < r_{i+1}$ and all $k \geq r_{i+1}$: $u_j < u_k < u_l$, and furthermore $v_{\sigma(j)} < v_{\sigma(k)} < v_{\sigma(l)}$ as well. As v decreases, $\sigma(\{r_i,\dots,r_{i+1}-1\}) = \{r_i,\dots,r_{i+1}-1\}$. The result follows directly from this.

□

6.B Clustering test under the Stochastic Blockmodel with the degrees

Assumption H is still made in this section. We would also like to complete our procedure based on the degrees with a test of the clustering structure. It is actually an extension of the remark at the end of the subsection 6.5.2. We would like to test by detecting a large gap in the degree distribution, whether there is only one class, which will be the null hypothesis, or more than one. Therefore it can be regarded as a clustering test, i.e. a test of the type "is not clustered" versus "clustered", or in other words, "the nodes are homogeneous" versus "the nodes are inhomogeneous". Under the Stochastic Blockmodel, it amounts to wonder whether the model is Erdős-Rényi or a mixture with more than one component.

6.B.1 Presentation of the test

The hypotheses are formally defined as follows:

Null hypothesis H_0: " $Q = 1$ " vs. Alternative hypothesis H_1: "$Q > 1$"

The decision of the test is based on the significance of the size of the largest gap. Under H_0, all normalized degrees should concentrate around a single value, and all gaps between them should be small. If the largest gap is rather "significantly" large, then it is unlikely that the degrees come from a model with only one class, and H_0 will be rejected.

The test statistic is:

$$G_n = \max_{i \in [n-1]} \left(T_{(i+1)} - T_{(i)} \right)$$

We would like to reject the null hypothesis when the statistic G_n is too large. Thus the region of rejection is $]t, 1]$ with some threshold $t \geq 0$, which depends on the significance level fixed by the user. For each $n \in \mathbb{N}^*$, the test defined by:

$$\phi_n^t(X) = \mathbb{1}_{G_n > t}$$

is called LG test. Note that both H_0 and H_1 are composite hypotheses. The error of the first kind of ϕ_n^t is then defined as the supremum of the probability that H_0 is rejected, i.e. G_n is larger than t, over all Erdős-Rényi models with n nodes, that is:

$$\chi_n^I(t) = \sup_{\pi \in [0,1]} P_\pi(G_n > t)$$

The error of the second kind is the supremum of the probability that H_0 is accepted over all Stochastic Blockmodels with more than one class:

$$\chi_n^{II}(t) = \sup \left\{ P_{\alpha,\pi}(G_n \leq t);\ Q \geq 2,\ \alpha \in Spl_Q,\ \pi \in \mathcal{S}_Q([0,1]) \right\}$$

where Spl_Q is the Q-dimensional simplex and $\mathcal{S}_Q([0,1])$ is the set of symmetric matrices with coefficients in $[0,1]$. Remind that a test sequence — all tests of the sequence having significance level a — is consistent if the error of the second kind is vanishing when n tends to infinity, or equivalently, if the power is tending to one.

For each $n \in \mathbb{N}$, let $t_n(a) > 0$ be such that $\phi_n^{t_n(a)}$ has significance level a. Define also for each $n \in \mathbb{N}^*$, $\beta_n(a)$ the error of the second kind of the LG test under the constraint that the significance level is a, which amounts to:

$$\beta_n(a) = \chi_n^{II}(t_n(a))$$

The main result of this section is the consistency of the LG test joined with the convergence rate:

Theorem 6.11. *Let $a \in]0,1]$. The LG test sequence with significance level a, $(\phi_n^{t_n(a)})_{n\in\mathbb{N}}$, is consistent and:*

$$\beta_n(a) \leq a \exp\left(-\frac{n\delta^2}{2}\left(1 - \frac{2}{\delta}\sqrt{\frac{2\ln\left(\frac{2n}{a}\right)}{n}}\right)\right) + Q\alpha_{max}^n$$

where $\alpha_{max} = \max_{q\in[Q]} \alpha_q$

The consistency follows from the upper bound of $\beta_n(a)$, which follows from Propositions 6.10 and 6.9:

Proposition 6.8.

$$\chi_n^I(t) \leq 2n \exp\left(-\frac{1}{2}nt^2\right)$$

Proposition 6.9. *Let for each $n \in \mathbb{N}^*$ and for all $a \in]0,1]$:*

$$t'_n(a) = \sqrt{\frac{2\ln\left(\frac{2n}{a}\right)}{n}}$$

Then $t_n(a) \leq t'_n(a)$ and the significance level of the test $\phi_n^{t'_n(a)}$ is at most a.

Proposition 6.10. *For all $t < \delta$:*

$$\chi_n^{II}(t) \leq 2n \exp\left(-\frac{1}{2}n(\delta - t)^2\right) + Q\alpha_{max}^n$$

6.B.2 Proof of Theorem 6.11

Proof. Indeed, as $P_{\alpha,\pi}(G_n \leq t)$ is increasing with respect to t, and using the inequality $t_n(a) \leq t'_n(a)$ from Proposition 6.9, we first have:

$$\beta_n(a) = P_{\alpha,\pi}(G_n \leq t_n(a)) \leq P_{\alpha,\pi}(G_n \leq t'_n(a))$$

Then using this inequality, the upper bound of the error of the second kind from Proposition 6.10 and the expression of $t'_n(a)$:

$$\begin{aligned}
\beta_n(a) &\leq 2n \exp\left(-\frac{n}{2}(\delta - t'_n(a))^2\right) + Q\alpha_{max}^n \\
&\leq 2n \exp\left(-\frac{n}{2}\left(\delta^2 - 2\delta\sqrt{\frac{2\ln\left(\frac{2n}{a}\right)}{n}} + \frac{2}{n}\ln\left(\frac{2n}{a}\right)\right)\right) + Q\alpha_{max}^n \\
&\leq a \exp\left(-\frac{n\delta^2}{2}\left(1 - \frac{2}{\delta}\sqrt{\frac{2\ln\left(\frac{2n}{a}\right)}{n}}\right)\right) + Q\alpha_{max}^n
\end{aligned}$$

□

6.B.3 Error of the first kind (proof of Propositions 6.8 and 6.9)

The exact distribution of the test statistic G_n is not known under H_0, but the tail of the distribution can be easily bounded from Proposition 6.3. The following proposition gives an upper bound of the probability that G_n exceeds a level t under the null hypothesis H_0.

Proof of Proposition 6.8

Proof. When $Q = 1$, the random graph model is Erdős-Rényi. The sole parameter of the model is the uniform connection parameter between nodes, denoted here by π. This is also the mean of the normalized degrees $(T_i)_{i\in[n]}$.

It is first proved that if the largest gap G_n is larger than t, then at least one of

the normalized degrees is actually far from the average π.

$$\begin{aligned}\{G_n > t\} &\subset \bigcup_{i\in[n-1]} \{|T_{(i)} - T_{(i+1)}| > t\} \\ &\subset \bigcup_{i\in[n-1]} \{|T_{(i)} - \pi - (T_{(i+1)} - \pi)| > t\} \\ &\subset \bigcup_{i\in[n-1]} \{|T_{(i)} - \pi| > t/2\} \cup \{|T_{(i+1)} - \pi| > t/2\} \\ &\subset \bigcup_{i\in[n]} \{|T_i - \pi| > t/2\} =\subset \left\{\max_{i\in[n]} |T_i - \pi| > t/2\right\} = \{d_n > t/2\}\end{aligned}$$

As a consequence, for all $\pi \in [0,1]$:

$$P_\pi(G_n > t) \leq P_\pi(d_n > t/2)$$

Now we are using Proposition 6.3:

$$P_\pi(d_n > t/2) \leq 2n \exp\left(-\frac{1}{2}nt^2\right)$$

As the upper bound holds for all $\pi \in [0,1]$, we can take the supremum over π in the left hand side and finally establish the proposition.

□

Since the exact distribution of G_n is not known, we can compute neither the exact error of the first kind, nor the exact threshold $t_n(a)$ to construct a test of given significance level $a \in [0,1]$, because they are actually linked to each other. Instead of this, we are deriving an upper bound of $t_n(a)$, denoted by $t'_n(a)$, from the upper bound of the tail of the distribution of G_n, so that the significance level of the test $\phi_n^{t'_n(a)}$ is not exactly a, but at most a.

Proof of Proposition 6.9

Proof. By definition of $t_n(a)$, and using Proposition 6.8:

$$a = P_\pi(G_n > t_n(a)) \leq 2n \exp\left(-\frac{1}{8}\frac{nt_n(a)^2}{\pi(1-\pi) + \frac{t_n(a)}{6}}\right)$$

Therefore:

$$t_n(a) \leq \sqrt{\frac{2\ln\left(\frac{2n}{a}\right)}{n}} = t'_n(a)$$

As $P_\pi(G_n > t)$ is decreasing with respect to t, we obtain:

$$a = P_\pi(G_n > t_n(a)) \geq P_\pi(G_n > t'_n(a))$$

The error of the first kind of $\phi_n^{t'_n(a)}$ is smaller than a, which exactly means that the significance level of this test is at most a. □

6.B.4 Error of the second kind (proof of Proposition 6.10)

Proof.

1. Conditioning on the event that two classes are present Under the hypothesis H_1, there are two or more classes in the model. However, if all nodes belong to the same class, the empirical degree distribution looks like there was only one class, and the test cannot detect any representative gap within the distribution. First of all, we thus condition on the event B_n that all nodes have not been drawn from the same class, i.e. at least two distinct classes are present in the label sample $(Z_i)_{i\in[n]}$. B_n can be written as follows:

$$B_n = \bigcap_{q\in[Q]} \{N_q < n\}$$

Note also that:

$$\{G \leq t\} = \bigcap_{i\in[n-1]} \{T_{(i+1)} - T_{(i)} \leq t\}$$

On the event B_n, since two distinct classes are present, there is necessarily one of the gaps $(T_{(i+1)} - T(i))_{i\in[n]}$ such that the end nodes are from distinct classes. That is: there is a $i_0 \in [n-1]$ such that nodes (i_0) and (i_0+1) are not in the same class, i.e. $Z_{(i_0)} \neq Z_{(i_0+1)}$, in spite of the fact that the distance of their normalized degrees is smaller than t. Therefore:

$$\{G \leq t\} \cap B_n \subset \bigcup_{i\in[n]} \{T_{(i+1)} - T_{(i)} \leq t\} \cap \{Z_{(i)} \neq Z_{(i+1)}\}$$

2. Normalized degree of one of the nodes is far from its own mean We now prove that conditionally on B_n, the normalized degree of one of the two nodes involved in this gap, either (i_0) or (i_0+1), is necessarily far from the conditional mean of its class. More precisely the distance to the class mean is no smaller than $\frac{\delta-t}{2}$. For all $i \in [n-1]$:

$$\begin{aligned} T_{(i+1)} - T_{(i)} &= |T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}} - (T_{(i)} - \overline{\pi}_{Z_{(i)}}) + \overline{\pi}_{Z_{(i+1)}} - \overline{\pi}_{Z_{(i)}}| \\ &\geq |\overline{\pi}_{Z_{(i+1)}} - \overline{\pi}_{Z_{(i)}}| - |T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}}| - |T_{(i)} - \overline{\pi}_{Z_{(i)}}| \end{aligned}$$

Note that for all $i \in [n-1]$, if $Z_{(i+1)} \neq Z_{(i)}$, then $|\overline{\pi}_{Z_{(i+1)}} - \overline{\pi}_{Z_{(i)}}| \geq \delta$. If in addition to that, $T_{(i+1)} - T_{(i)} \leq t$, then:

$$\begin{aligned} t \geq T_{(i+1)} - T_{(i)} &\geq \delta - |T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}}| - |T_{(i)} - \overline{\pi}_{Z_{(i)}}| \\ |T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}}| + |T_{(i)} - \overline{\pi}_{Z_{(i)}}| &\geq \delta - t \end{aligned}$$

Moreover, since $t < \delta$, the following alternative holds:

$$\text{either } |T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}}| \geq \frac{\delta - t}{2} \text{ or } |T_{(i)} - \overline{\pi}_{Z_{(i)}}| \geq \frac{\delta - t}{2}$$

These inequalities imply:

$$\{T_{(i+1)} - T_{(i)} \leq t\} \cap \{Z_{(i)} \neq Z_{(i+1)}\} \subset \left\{|T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}}| \geq \frac{\delta - t}{2}\right\} \cup \left\{\left|T_{(i)} - \overline{\pi}_{Z_{(i)}}\right| \geq \frac{\delta - t}{2}\right\}$$

As a conclusion:

$$\begin{aligned} \{G_n \leq t\} \cap B_n &\subset \bigcup_{i \in [n-1]} \{T_{(i+1)} - T_{(i)} \leq t\} \cap \{Z_{(i)} \neq Z_{(i+1)}\} \\ &\subset \bigcup_{i \in [n-1]} \left\{|T_{(i+1)} - \overline{\pi}_{Z_{(i+1)}}| \geq \frac{\delta - t}{2}\right\} \cup \left\{\left|T_{(i)} - \overline{\pi}_{Z_{(i)}}\right| \geq \frac{\delta - t}{2}\right\} \\ &\subset \bigcup_{i \in [n]} \left\{|T_i - \overline{\pi}_{Z_i}| \geq \frac{\delta - t}{2}\right\} \\ &\subset \left\{\max_{i \in [n]} |T_i - \overline{\pi}_{Z_i}| \geq \frac{\delta - t}{2}\right\} = \left\{d_n \geq \frac{\delta - t}{2}\right\} \end{aligned}$$

which was announced at the point 2 above.

3. Using the bound of probability of large spreading from Proposition 6.3 For all $\alpha \in Spl_Q, \pi \in \mathcal{S}_Q([0,1])$,

$$\begin{aligned} P_{\alpha,\pi}(G_n \leq t) &= P_{\alpha,\pi}(G_n \leq t \cap B_n) + P_{\alpha,\pi}\left(\{G_n \leq t\} \cap \overline{B_n}\right) \\ &\leq P_{\alpha,\pi}(G_n \leq t \cap B_n) + P\left(\overline{B_n}\right) \\ &\leq P_{\alpha,\pi}\left(d_n \geq \frac{\delta - t}{2}\right) + P\left(\overline{B_n}\right) \end{aligned}$$

$\overline{B_n}$ is the event that there is only one class which is present in the label sample, and can be written as follows:

$$\overline{B_n} = \bigcup_{q \in [Q]} \{N_q = n\}$$

Since $N_q \sim \mathcal{B}(n, \alpha_q)$, we thus have for all $\alpha \in Spl_Q, \pi \in \mathcal{S}_Q([0,1])$:

$$\begin{aligned} P_{\alpha,\pi}\left(\overline{B_n}\right) &= P_{\alpha,\pi}\left(\bigcup_{q \in [Q]} \{N_q = n\}\right) \\ &\leq \sum_{q \in [Q]} P_{\alpha,\pi}(N_q = n) = \sum_{q \in [Q]} \alpha_q^n \\ &\leq Q\alpha_{max}^n \end{aligned}$$

Using also Proposition 6.3, we can get the final bound:

$$P_{\alpha,\pi}(G_n \leq t) \leq 2n \exp\left(-\frac{1}{2} n(\delta - t)^2\right) + Q\alpha_{max}^n$$

This bound is uniform with respect to (α, π), therefore we can take the supremum again and the inequality of the proposition follows. □

Chapter 7

Clustering in a random graph model with latent positions

This chapter is a joint work with Jean-Jacques Daudin and Sté phane Robin. It is a project which will be submitted within next months.

Many scientific domains interested in the global behavior of a group, a population or an organization, do not collect data on each individual or entity only, but also on pairwise interactions between them. Interaction data are represented by a graph: nodes represent individuals, and ties between pairs of nodes represent interaction between individuals. Ties can be quantitative, modeling flow or traffic intensity through the network. This chapter deals with undirected binary graphs, just indicating the presence or absence of interaction: for example, collaboration between researchers, friendship in a social network, transmission of information or infection between people.

Real-world networks are inhomogeneous in terms of degree of the nodes (i.e. their number of neighbors), hierarchical or community structure, etc. Latent space models provide a general framework to account for such an heterogeneity (Bollobás et al., 2007). The general model is as follows (see also Section 3.4.1 in Chapter 3). For a graph with n nodes, each node $i \in [n]$ is associated with an (unobserved) position Z_i in a latent space. The positions $(Z_i)_{1 \leq i \leq n}$ are supposed to be i.i.d. with common distribution f. Conditionally on the positions $Z = (Z_i)_{1 \leq i \leq n}$, the edges $(X_{ij})_{1 \leq i,j \leq n}$ are independent Bernoulli random variables and the probability of presence of the edge between i and j, — the parameter of X_{ij} — is $\kappa_n(Z_i, Z_j)$. Unlike the positions Z, the network described by the edges in matrix X is supposed to be completely available in the data.

In many settings, two individuals interact when they look alike, meaning that their position are close. This property is often called homophily (see Hoff et al. (2002); Handcock et al. (2007) and references therein) and is related to graphs embedded in a space equipped with a distance or more generally a dissimilarity measure. In

this chapter, the probability of connection of nodes i and j conditionally on Z_i and Z_j, $\kappa_n(Z_i, Z_j)$, is thus supposed to decrease with respect to the euclidean distance between Z_i and Z_j. Note that through the triangular inequality, the distance also brings transitivity ('friends of my friends are more likely to be my friends as well'), which is also a desirable property in many contexts.

Clustering brings together a lot of various methods generally aiming at splitting the population into distinct homogeneous subgroups, see Hartigan (1975); Von Luxburg and Ben-David (2005). However this notion is not universally defined and strongly depends on the homogeneity criterion. In many papers, these subgroups, called clusters, are implicitly defined, often as optimizers of some criterion coming from an heuristic idea, see Newman (2006) for instance. But most of the time, it is hard to figure out what exactly are these clusters. In this chapter, we provide an explicit definition of these objects, and a practical method which estimates their number in the context of graphs.

Some random graph models are designed for a clustering purpose, and clusters are defined by the model itself. In mixture-based models, each component of the mixture is a cluster. For examples in networks, in the Stochastic Block Model (see 3.4.2, Chapter 6 and references therein), where the connection probability between two nodes depends only on the node colors, the homogeneity is based on the similarity of average connection behavior: two nodes are regarded as similar if they connect to the same colors with the same probabilities. Closer to our setting than the SBM is the Latent Position Cluster Model (see Subsection 4.1.2 in Chapter 3 and Handcock et al., 2007): the unknown common density f of the latent positions is supposed to be a Gaussian mixture, and two nodes are in the same cluster if they have been drawn from the same component of the mixture. However this model supposes rigid constraints on the cluster shape. In particular it is impossible to fit a unique complex-shaped cluster with only one Gaussian distribution, even though this is required to test whether the distribution is clustered. Furthermore in the case of the estimation of the cluster number, many Gaussian mixtures may be needed to fit complex-shaped clusters and the number of clusters may be dramatically overestimated. From this perspective, it seems desirable to make as few assumptions as possible about the shape of the distribution f, as we try in this chapter.

In this chapter, unlike in mixture-based models, clusters are not defined by the model. Positions $(Z_i)_{i\in[n]}$ are only supposed to have a common density f on $\mathbb{R}^d$, and no more assumption is made on f except for regularity. In this non-parametric setting, a common definition of clustering is that of Hartigan (1975), which was designed for generic data, that is independent pointwise data in $\mathbb{R}^d$: positions are directly observed (no network information is provided). Hartigan (1975) describes clusters as connected regions — in a topological sense — of high density separated

from each other by regions of low density. For all $t > 0$, a cluster of f at level t is a topological connected component of $\{f \geq t\}$. The whole clustering structure of f is described by the set of all clusters, when t takes all possible values. This definition has become popular in the field of non-parametric statistics, see Cuevas et al. (2000, 2001); Azzalini and Torelli (2007). It is also introduced in Section 4.3, Chapter 4.

In this non-parametric setting for observed positions, strategies based on proximity graphs have been developed to find clusters in generic data. Clusters are typically identified as connected components of a proximity graph built from the pointwise data (ε-graphs or kNN-graphs, see Subsection 4.3.1 in Chapter 4 and see Hartigan (1981); Penrose (2003); Biau et al. (2007); Pelletier and Pudlo (2011); Chaudhuri and Dasgupta (2010)). Besides, in the point of view of Hartigan (1975), two points are regarded as similar at level t if they belong to the same cluster of f, that is if there is a continuous path between them through the same high density region, and without crossing a low density region. Loosely speaking, edge paths through the proximity graph mimic topological paths in the data space. Therefore this kind of strategy makes a link between both topological and graphical notions of connectedness. The link comes from the homophily: the fact that only close nodes are connected is a critical property so that the generation process of the graph preserves the cluster structure of f in some sense.

The general idea of the present chapter is to borrow graphical techniques developed for non-parametric clustering for observed positions and apply them to network clustering, where latent positions are unobserved. The main question addressed here is the estimation of the number of t-clusters for a given level $t > 0$. Biau et al. (2007) provides such an estimate in the framework of observed positions. The algorithm consists in first estimating the density at the node positions, via an usual estimator (kernel estimator for instance), then removing nodes where the estimated density is low and finally counting the number of connected components of the pruned graph.

Unlike Biau et al. (2007), our density estimator cannot be based on the positions Z as they are not observed. Therefore normalized node degrees of the observed graph will be used as a surrogate estimator of the density at the node positions. Thus our procedure simply computes the normalized degrees of the graph, removes low degree nodes and counts connected components of this pruned graph. This procedure amounts to weak K-linkage algorithm (see introduction of Penrose, 2003 or Ling, 1973) with K equal to t times the normalization constant of the degrees. It is especially fast and is able to process very large graphs.
Moreover the Depth First Search algorithm actually yields the components of the pruned graph, which provide an estimate of the number of clusters as well as a natural classifier of the high density nodes into clusters. The quality of the resulting

classification will be also studied in this chapter.

To give more details about the connection between usual non-parametric clustering and Thus we reinterpret techniques developed in usual non-parametric clustering for observed positions, but from the point of view of latent variables and apply them to network clustering, where positions are unobserved. When the positions are observed, the graph is a proximity graph and is built just as a tool to perform clustering: it is actually not the original data, whereas in our framework the graph is the only original observed data. The observed graph under the proposed model can also be thought of as a proximity graph (ε-graph) whose edges are randomly removed, such that the probability of a tie between two nodes is even larger when they are close. Finally, we slightly improve the result of Biau et al. (2007), since we show that the graph is sufficient to do the same job: positions are actually not needed to design a consistent estimate of the number of clusters.

In Section 7.1, the model is introduced from a mathematical point of view and some of its probabilistic properties like invariances by transformations of the latent space and the different density regimes are highlighted. The algorithm and what it is attempting to do is also presented in that section. Then Section 7.2 proves the asymptotic non-underestimation of the number of clusters, essentially based on density estimation tools. Section 7.3 provides a partial result about the asymptotical non-overestimation, which involves elements of both geometry and percolation. Some extra proof elements towards consistency are provided in the appendices. Moreover connected components of the pruned graph are more accurately studied in Section 7.4. Results of this section follow from elements of the two previous sections which are just reinterpreted to emphasize their properties from a classification point of view. A simulation study is carried out in Section 7.5 to illustrate the estimator of the number of clusters and the classification procedure. Some other empirical statistics are also presented to analyze the behavior of the algorithm. Previous sections are only theoretical and in this section practical issues are discussed as well, in particular the issue of the choice of t.

7.1 The model

7.1.1 Random graph model

For all $n \in \mathbb{N}^*$, $[n]$ denotes the set $\{1, \dots, n\}$. Binary graphs with n nodes are represented by the adjacency matrix $X = (X_{ij})_{i,j\in[n]}$, where $X_{ij} = 1$ if there is an edge between nodes i and j, $X_{ij} = 0$ otherwise. Graphs are assumed to be undirected and with no self-loops, then X is symmetric and its diagonal is zero. We consider the following model:

– Let $Z = (Z_i)_{i\in[n]}$ be independent and identically distributed variables drawn

from a density f with respect to the Lebesgue measure on $\mathbb{R}^d$:

$$Z = (Z_i)_{i\in[n]} \text{ i.i.d.} \sim f.$$

– Conditionally on the latent positions Z, the edges $(X_{ij})_{i,j\in[n]}$ are independent Bernoulli random variables such that for each pair of nodes $\{i,j\} \subset [n]$:

$$(\ X_{ij} \mid Z_i, Z_j\ ) \sim \mathcal{B}\left(k\left(\frac{Z_i - Z_j}{h_n}\right)\right)$$

with $h_n > 0$ and $k : \mathbb{R}^d \to [0,1]$ an isotropic and decreasing function with respect to the euclidean norm [1].

k is called *connection function*. It gives the probability of connection for a given distance. Homophily has been addressed in Introduction: the fact that k is decreasing models the fact that nodes are more likely connected if they are close in the latent space. Isotropy seems to be a natural assumption because there is no apparent reason why homophily would be stronger in a specific direction than in another one. Note also that $k(0) = 1$ is not necessary.

h_n is called *connection radius*. The distance between nodes is scaled by this parameter and then it affects their probability of being connected to each other. In this chapter, this parameter has a double interpretation we are discussing now.

From the model point of view, if the support of the connection function k is included in the unit ball (see Assumption 7.3.(a)) any two nodes have a non-zero probability of being connected if and only if their distance in the latent space is less than h_n. On the other hand, when h_n tends to zero, the probability of an edge between any two nodes tends to zero. Thus parameter h_n controls the density of the graph. Note that in particular, the average number of edges scales with $n^2 h_n^d$. Therefore $n^2 h_n^d \to +\infty$ should be assumed, so that the graph does not get empty when the number of nodes n tends to infinity and so as to avoid trivial models.

From the point of view of density estimation, h_n is a bandwidth parameter. The degrees of the nodes of X will be normalized by nh_n^d to mimic a kernel density estimator (see Section 7.1.5). Indeed these normalized degrees will be used as an estimator of the latent position density f at the node positions. One major theoretical difference with Biau et al. (2007) is that we cannot use a bandwidth parameter other than h_n in our estimator, as we can use only the graph for estimation and h_n is imposed by the graph model.

Note also that, although the algorithm theoretically studied in this chapter requires h to compute the normalized degrees, the parameter h_n is not identifiable from the graph (see Subsection 7.1.4). If h_n is not known, which must be most of the cases with real-world networks, we also propose a practical algorithm in Subsection 7.5.4.

1. It is equivalent to say that k can be written the following way: $k = \tilde{k}(\|\cdot\|/h)$ with $\tilde{k} : \mathbb{R}^+ \to [0,1]$ decreasing.

7.1.2 Cluster definition and other notations

We use the definition of Hartigan (1975). Let $t > 0$. The t-level set of the density f is:

$$\mathcal{L}(t) = \{x \in \mathbb{R}^d;\ f(x) \geq t\}.$$

Each connected component of $\mathcal{L}(t)$ is called a *geometric t-cluster* or just a t-cluster. The set of all geometric t-clusters is denoted by $\{\mathcal{C}_i\}_{i\in[Q(t)]}$ where $Q(t)$ denotes the number of such t-clusters. The distribution f is said to be *clustered* at level t if there is more than one t-cluster, that is if $Q(t) > 1$.

For any subset C of $\mathbb{R}^d$, let us define α_C as the probability that $Z_i \in C$ for any $i \in [n]$:

$$\alpha_C = P(Z_i \in C) = \int_C f(z)dz.$$

The probability that any Z_i is in the t-level set is also denoted by $\mu_f(t)$, the probability that it is in the i-th t-cluster is denoted by $\alpha_i(t)$ for all $i \in [Q(t)]$, and $\alpha_0(t)$ denotes the smallest one, that is:

$$\mu_f(t) = P(Z_1 \in \mathcal{L}(t)), \quad \alpha_i(t) = P(Z_1 \in \mathcal{C}_i(t)), \quad \alpha_0(t) = \min_{i\in[Q(t)]} \alpha_i(t), \tag{7.1}$$

so $\mu_f(t) = \sum_{i\in[Q(t)]} \alpha_i(t)$.

δ_t further denotes the smallest distance between any two t-clusters, and is called minimal inter-cluster distance:

$$\delta_t = \min\{\|z - z'\|;\ z \in \mathcal{C}, z' \in \mathcal{C}' \text{ where } \mathcal{C} \text{ and } \mathcal{C}' \text{ are distinct } t\text{-clusters of } f\}.$$

Recall also the following definition of Chapter 5, for all $h > 0$:

$$s_h = \sup_{\|z-z'\|\leq h} |f(z) - f(z')|.$$

Let A be a subset of $\mathbb{R}^d$. Define the subgraph of X *induced by* A and denoted by X_A as the subgraph induced by the nodes which are located in A. If X is represented by $([n], E)$ with $[n]$ the node set and E the edge set, X_A is represented by (V_A, E_A) where $V_A = \{i \in [n]; Z_i \in A\}$ and $E_A = E \cap \mathcal{P}(V_A)$ with $\mathcal{P}(J)$ the set of the pairs of elements from J. The cardinality of V_A, that is the number of nodes located in A, is denoted by n_A.

The set of the high density nodes is denoted by $J_n(t) = \{i \in [n];\ f(Z_i) \geq t\}$. In this chapter, the degrees will be normalized by nh_n^d, hence the normalized degree T_i of any node $i \in [n]$ is given by:

$$T_i = \frac{1}{nh_n^d} \sum_{j\neq i} X_{ij}$$

and the set of high degree nodes by $\widehat{J}_n(t) = \{i \in [n];\ T_i \geq t\}$. The subgraph induced by this set of nodes is denoted by $\widehat{X}(t)$.

For all points $z, z' \in \mathbb{R}^d$ we write $z \sim_t z'$ if they both are in $\mathcal{L}(t)$ and in the same t-cluster (i.e. if there is a continuous path in $\mathcal{L}(t)$ between the two points). For all nodes $i, j \in [n]$, we also write $i \widehat{\sim}_t j$ if they both are in $\widehat{J}_n(t)$ and in the same connected component of $\widehat{X}(t)$ (i.e. if there is a path in graph $\widehat{X}(t)$ between them), so as to emphasize the fact that paths in the graph $\widehat{X}(t)$ are somehow an approximation of paths in the latent space.

Finally, we will often use this notation for the conditional probability: if Y is a random variable or an event, and A is an event, $P^Y(A)$ is the probability of A conditionally on Y, and $\mathbb{E}^Y$ is the expectation conditionally on Y.

7.1.3 Graph density

In this paragraph, the link between the connection radius $h_n = h$ and the average graph density $\zeta_n(h)$ is emphasized. First, it is clear that for a fixed distance between two nodes, the probability of having an edge grows when h decreases, because k is a decreasing function. It is even possible to establish an equivalent of the average density for small connection radii.

Define the average density of the graph X, that is the expected fraction of edges over the number of node pairs:

$$\zeta_n(h) = \mathbb{E}\left(\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} X_{ij}\right).$$

The following proposition (proved in Appendix 7.A) gives an equivalent when h tends to zero.

Proposition 7.1. *For any fixed $n \in \mathbb{N}^*$,*

$$\zeta_n(h) \sim_{h \to 0} h^d \int_{\mathbb{R}^d} f^2$$

It is often noticed in real-world networks that graphs become sparser as the size n increases, i.e. that the graph density vanishes. Even though $h = h_n \xrightarrow[n \to \infty]{} 0$ is not a necessary condition to have consistency of the proposed method, it can be interesting to make this assumption to have sparsity when n tends to infinity. And in that case, we can derive an equivalent of $\zeta_n(h_n)$ when n tends to infinity with the same proof as Proposition 7.1:

$$\zeta_n(h_n) \sim_{n \to \infty} h_n^d \int_{\mathbb{R}^d} f^2.$$

7.1.4 Model invariance up to similarity transformations of the latent space

Here we show that the model is invariant up to any similarity transformation of the latent space, which entails that the positions are not identifiable, like in Hoff et al. (2002).

Definition 7.1. *Let $\lambda > 0$. A similarity transformation R of $(\mathbb{R}^d, \|\cdot\|)$ with scale factor λ is a bijective map $\mathbb{R}^d \to \mathbb{R}^d$ such that:*

$$\forall z, z' \in \mathbb{R}^d, \ \|R(z) - R(z')\| = \lambda \|z - z'\|.$$

The graph model is invariant up to similarity transformations of the latent space, as k is isotropic. Indeed, if the latent space is transformed by some similarity transformation R with scale factor λ, then the probability of an edge between any two nodes $i, j \in [n]$ does not change, that is:

$$k\left(\frac{R(Z_i) - R(Z_j)}{\lambda h}\right) = k\left(\frac{Z_i - Z_j}{h}\right).$$

Note that here h turns into λh, because h is the maximal distance between two connected nodes in the latent space, and the transformation R changes all distances with the scale factor λ.

The latent distribution and the connection radius (f, h), are not identifiable, because any similarity transformation with scale factor λ would change both f and h and would give the same graph model. If $R = H \circ I$ with H any homothetic transformation with scale factor λ, and I any affine isometric transformation, in the latent space transformed by R, the density of the positions turns into:

$$\frac{1}{\lambda^d} f\left(\frac{R^{-1}(\cdot)}{\lambda}\right).$$

and the connection radius is λh.

Issue of the choices of t and h Therefore h cannot be estimated from the graph X. Note also that such a transformation of the latent space change the density level t into $\frac{t}{\lambda^d}$, and $Q(t)$ into $Q(t/\lambda^d)$. As a consequence, there is no proper value of t prescribed *a priori* or inferable from the graph data. In fact the issue of the choices of t and h are connected, since the product th^d is invariant up to similarity transformations. We will use this property to propose a practical algorithm in Subsection 7.5.4.

7.1.5 Density estimation from the degrees

Recall the definition of f_n the density estimator of f based on the kernel k:

$$f_n(z) = \frac{1}{nh^d} \sum_{j \in [n]} k\left(\frac{z - Z_j}{h}\right).$$

The degree of a node $i \in [n]$ is denoted by $D_i = \sum_{j \in [n]} X_{ij}$. The density at the position Z_i of the node i will be estimated by the normalized degree of this node, we are recalling:

$$T_i = \frac{1}{nh^d} \sum_{j \in [n]} X_{ij}.$$

The following calculation shows that the expected normalized degree of a node i, conditionally on the positions Z, turns out to be the kernel estimator of the density at the position Z_i of the node:

$$\mathbb{E}^Z(T_i) = \frac{1}{nh^d} \sum_{j \in [n]} \mathbb{E}^Z(X_{ij}) = \frac{1}{nh^d} \sum_{j \in [n]} k\left(\frac{Z_i - Z_j}{h}\right) = f_n(Z_i).$$

7.1.6 Algorithm

In Biau et al. (2007) and Ling (1973), in order to find out the clustering structure of the density of some *observed* position data, a deterministic graph is constructed from the data just by putting edges between nodes whose distance is less than a threshold h. Both authors run an algorithm on this instrumental graph, which can be summarized in four steps: (1) estimate the density at each data point (i.e. node), (2) remove low density nodes, (3) extract the subgraph made of the remaining nodes (4) count the connected components of the pruned graph. Our algorithm is similar to this procedure, step by step:

Algorithm 7.1.

1. *Compute the normalized degrees $(T_i)_{i \in [n]}$ of the observed graph X.*
2. *Threshold normalized degrees: find the set of the nodes such that the normalized degree exceeds t, denoted by $\widehat{J}_n(t) = \{i \in [n];\ T_i \geq t\}$.*
3. *Extract from X the subgraph induced by $\widehat{J}_n(t)$, that is $\widehat{X}(t)$.*
4. *Compute and count connected components of $\widehat{X}(t)$ with Depth First Search.*

The procedure basically amounts to K-linkage (Ling, 1973), with $K = \lceil tnh^d \rceil$. Biau et al. (2007) keeps only the nodes $i \in [n]$ such that $f_n(Z_i) \geq t$, where f_n is any usual density estimator. The key feature is that the normalized degree is

used here as a surrogate density estimator instead of an usual density estimator. The first one is based only on the graph data X, whereas the latter one, used in Biau et al. (2007), is based on the position data Z, which is supposed to be not available in our setting. The objective of this paper could be thought of as showing that there is no need for knowing the position data, in particular to estimate the number of clusters: this can be actually done with the graph alone.

7.1.7 Assumptions

Our main results address theoretical asymptotical properties of this algorithm. Their proofs require the following assumptions.

Assumption 7.1. ***(Regularity)*** *The density f is:*

(a) uniformly continuous on $\mathbb{R}^d$

(b) continuously differentiable on a neighborhood of $\{f = t\}$

(c) the gradient of f is non-zero at all $x \in \{f = t\}$

Assumption 7.1 will be made all along this chapter and is also made in Biau et al. (2007). Note that an integrable and uniformly continuous function vanishes at infinity and then for all $t > 0$, $\mathcal{L}(t)$ is bounded (see the proof in Appendix 7.A.2). Furthermore, we can also use the following lemma under Assumption 7.1.

Lemma 7.1. ***(Geometric Lemma)*** *There exists $\varepsilon_0 > 0$ such that for all $\varepsilon \in [0, \varepsilon_0]$, the intersection of each connected component of $\mathcal{L}(t - \varepsilon)$ with $\mathcal{L}(t + \varepsilon)$ is non-empty and each connected component of $\mathcal{L}(t + \varepsilon)$ is the intersection of a connected component of $\mathcal{L}(t - \varepsilon)$ with $\mathcal{L}(t + \varepsilon)$.*

Its proof is given in Appendix 7.A. This ε_0 will also be used all along the chapter. Note that it depends on t. The main consequence of the lemma is that $Q(t - \varepsilon) = Q(t) = Q(t + \varepsilon)$ for all $\varepsilon \in [0, \varepsilon_0]$.

Provided that f is uniformly continuous, for each $\varepsilon > 0$, there exists $\eta_\varepsilon > 0$ such that $s_{\eta_\varepsilon} \leq \varepsilon/2$. Hence we can make the following assumption:

Assumption 7.2. ***(Density estimation)***

(a) $h \leq \eta_{\varepsilon_0}$

(b) the integral of k is 1, so that k is a kernel.

Assumption 7.2.(a) enables the density to be estimated from the graph degrees with a sufficient quality, see Section 7.1.5 and the discussion in Section 7.2.6. Note that it is fulfilled for n large enough whenever $h = h_n \to 0$.

The connection function k is involved in some density estimator, see paragraph 7.1.5. Assumption 7.2.(b) is made so that levels of the density estimators can be interpreted as levels of the density f too. As a matter of fact, it is rather an assumption for interpretation than a theoretical limit.

Assumption 7.3. ***(Cluster separation)***

(a) the support of k is the unit ball $B(0,1)$

(b) $h < \delta_{t-\varepsilon_0}$

The main consequence of Assumption 7.3.(a) is that two nodes have a non-zero probability of being connected if and only if their distance in the latent space is less than h. Assumption 7.3.(b) will prevent nodes of disjoint clusters from being connected and turns out to be useful when showing non-underestimation of the number of clusters, see Proposition 7.3. Note that Assumption 7.3.(b) is again fulfilled for n large enough whenever $h_n \to 0$.

7.2 Cluster separation: non-underestimation

In this section, we will bound from above the probability that, for a given $t > 0$, the algorithm underestimates $Q(t)$, when the connection radius h_n is small enough.

Theorem 7.1. *Under Assumptions 7.2 and 7.3:*

$$P\left(\widehat{Q}_n(t) < Q(t)\right) \leq 6n \exp\left(-\frac{1}{32}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right) + Q(t)(1-\alpha_0(t))^n.$$

7.2.1 Intermediate results

We will need the following definitions and propositions to prove the theorem. The proofs of the propositions are given in Appendix 7.B.

Location of the nodes of $\widehat{X}(t)$: control of the density overestimation outside clusters

Let us define the maximal estimation error as:

$$U_n = \sup_{1 \leq i \leq n} |T_i - f(Z_i)|$$

We first study the location of the nodes of $\widehat{X}(t)$: control of the density overestimation outside clusters.

Let us define $E_n^\varepsilon(t)$ as the event where each connected component of $\widehat{X}(t)$ is completely included in one $(t-\varepsilon)$-cluster.

The following proposition states that whenever the maximal overestimation error is smaller than a given level ε, nodes of $\widehat{X}(t)$ are necessarily located in the $(t-\varepsilon)$-level set, basically the geometric t-level set plus a small safety margin.

Proposition 7.2. *For all $\varepsilon > 0$,*

$$\{U_n \leq \varepsilon\} \subset \left\{\widehat{J}_n(t) \subset J_n(t-\varepsilon)\right\}.$$

Gap between clusters

We then study the gaps between clusters. As geometric t-clusters are separated by regions of low density, there is a gap distinct clusters, called minimal intercluster distance, denoted by δ_t (defined in 7.1.2). Moreover, recall that nodes cannot be connected, whenever they are further than the connection radius h_n from each other, because of Assumption 7.3.(a) on the support of k. Therefore, if h_n is basically smaller than the intercluster distance (Assumption 7.3.(b)), nodes from distinct clusters cannot be connected to each other, and thus the components of $\widehat{X}(t)$ are well separated. This is the meaning of the next proposition.

Proposition 7.3. *Under Assumption 7.3, the following inclusion is satisfied:*

$$\left\{\widehat{J}_n(t) \subset J_n(t-\varepsilon)\right\} \subset E_n^{\varepsilon}(t).$$

7.2.2 Presence of nodes of $\widehat{X}(t)$ in every cluster: control of the density underestimation inside clusters

Now, the graph $\widehat{X}(t)$ has to contain at least one node in each cluster to make sure that the graph completely covers the level set to estimate the right number of clusters. Let $B_n^{\varepsilon}(t)$ the event that there is at least one node in each $(t+\varepsilon)$-cluster which is not removed by the algorithm, that is:

$$B_n^{\varepsilon}(t) = \bigcap_{i=1}^{Q(t+\varepsilon)} \{V_{\mathcal{C}_i(t+\varepsilon)} \cap \widehat{J}_n(t) \neq \varnothing\}.$$

Let $A_n(t)$ be the event that there is at least one node in each t-cluster:

$$A_n(t) = \bigcap_{i=1}^{Q(t)} \{V_{\mathcal{C}_i(t)} \neq \varnothing\}.$$

Proposition 7.4 shows that if there is at least one node of X in each cluster, and the density at this node is not too much underestimated by its degree, then clusters are properly covered by nodes of $\widehat{X}(t)$.

Proposition 7.4. *For all $\varepsilon > 0$,*

$$\{U_n \leq \varepsilon\} \cap A_n(t+\varepsilon) \subset B_n^{\varepsilon}(t).$$

7.2.3 Non-underestimation

Finally, Proposition 7.5 shows that $\widehat{Q}_n(t)$ does not underestimate $Q(t)$ if basically:

- there is no node of $\widehat{X}(t)$ in low density regions and no edge across such regions (see event $E_n^{\varepsilon}(t)$),
- there is at least one node of the graph $\widehat{X}(t)$ in each cluster (see event $B_n^{\varepsilon}(t)$).

Proposition 7.5. *If ε_0 is chosen according to Lemma 7.1, then:*

$$E_n^{\varepsilon_0}(t) \cap B_n^{\varepsilon_0}(t) \subset \left\{\widehat{Q}_n(t) \geq Q(t)\right\}$$

7.2.4 Upper bounds

Now we give the upper bounds required to conclude from the inclusions provided above, that is upper bounds of the probability for having (1) an empty cluster and (2) a large density estimation error.

Proposition 7.6.

1. $P\left(\overline{A_n(t)}\right) \leq Q(t)(1-\alpha_0(t))^n$.
2. *Under Assumption 7.2,*

$$P(U_n > \varepsilon_0) \leq 6n \exp\left(-\frac{1}{32}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right).$$

7.2.5 Proof of Theorem 7.1

Combining Propositions 7.2, 7.3 and 7.4, we first have:

$$\{U_n \leq \varepsilon_0\} \cap A_n(t+\varepsilon_0) \subset B_n^{\varepsilon_0}(t) \cap E_n^{\varepsilon_0}(t).$$

Then using Proposition 7.5, we also have:

$$\{U_n \leq \varepsilon_0\} \cap A_n(t+\varepsilon_0) \subset \left\{\widehat{Q}_n(t) \geq Q(t)\right\}.$$

The consequence of this inclusion is that:

$$\begin{aligned} P\left(\widehat{Q}_n(t) > Q(t)\right) &\leq P\left(\{U_n > \varepsilon_0\} \cup \overline{A_n(t+\varepsilon_0)}\right) \\ &\leq P(U_n > \varepsilon_0) + P\left(\overline{A_n(t+\varepsilon_0)}\right) \end{aligned}$$

Proposition 7.6 provides an upper bound for each of these terms.

7.2.6 Discussion about the connection radius h_n

Firstly we examine the role of h_n in this section. It plays two key roles: it is the largest possible distance between two connected nodes as well as a bandwidth parameter of an underlying kernel density estimator. Assumption 7.3 deals with the first role: the connection radius needs to be small enough to prevent $\widehat{X}(t)$ from having "wrong" edges between nodes from distinct clusters. Besides Assumption 7.2.(a) is related to the quality of the underlying kernel density estimator: h_n needs to be small enough so that the bias is not too large, that is, the estimation error is not too large.

Note that the assumption $h_n \to 0$ is actually not necessary at all. Indeed the bias of the underlying kernel estimator is proportional to h_n under regularity assumption on f (Tsybakov, 2003), and in usual density estimation, it is thus supposed to be vanishing to derive an asymptotically unbiased estimator. But here the separation between clusters arises whenever the bias is smaller than some threshold, which depends on t. Hence h_n just needs to be small enough to achieve separation.

Nevertheless h_n can also be assumed to be vanishing, but for another reason mentioned in Section 7.1.3: namely the graph density is proportional to h_n^d, and in real-world networks, this variable is precisely vanishing when the number of nodes grows. Thus note that according to Theorem 7.1, the probability of underestimation is still vanishing in that case whenever

$$\frac{nh_n^d}{\ln n} \to +\infty.$$

7.3 Connectedness inside clusters: non-overestimation

The main theorem of this section provides a partial result about the non-overestimation of $Q(t)$. Unlike Theorem 7.1, it does not deal with the graph returned by the algorithm, but with the subgraph of X induced by $\mathcal{L}(t)$, namely $X_{\mathcal{L}(t)}$. This is actually the graph which would be returned by the algorithm, conditioned on the fact that there is no mistake in the thresholding step.

Theorem 7.2. *Let $\widetilde{Q}_n(t)$ the number of connected components of $X_{\mathcal{L}(t)}$. There exists K_0,K_1 non-negative constants depending on d, f, k and t such that:*

$$P\left(\widetilde{Q}_n(t) > Q(t)\right) \leq K_0 n \exp(-K_1 n h_n^d)$$

This theorem essentially tells that the subgraph of X induced by each cluster $\mathcal{C}_i(t)$, that is $X_{\mathcal{C}_i(t)}$, is connected. The proof is divided in two steps. We first prove *locally* the connectedness of the subgraphs of X, that is, subgraphs induced by

small subsets of the clusters are connected with high probability when n tends to infinity (Proposition 7.7), and secondly, we prove with tilings of the clusters that the connectedness spreads over each whole cluster w.h.p. when n tends to infinity.

Note that supplementary material about the non-overestimation issue is provided in Appendix 7.D. In particular it is shown there that w.h.p. when n tends to infinity, $X_{\mathcal{L}(t+\varepsilon_n)}$ equals the graph the nodes of which are selected by the algorithm *and* are in $\mathcal{L}(t+\varepsilon_n)$, with no more than the tools used in this section. Thus the gap to the complete consistency actually consists in controlling nodes in the neighborhood of the boundary of the level set (see also comments in the cited appendix).

7.3.1 Local Connectedness

In this subsection, we give an upper bound of the probability of disconnectedness of the subgraph X_C induced by a small compact set.

Let C a subset of $\mathbb{R}^d$. Conditionally on Z, X_C is a fully parametrized random graph with parameters $\mathbf{p} = (p_{ij})_{i,j\in V_C}$, where for all $i, j \in V_C$:

$$p_{ij} = P^Z(X_{ij} = 1) = k\left(\frac{Z_i - Z_j}{h_n}\right)$$

Let $\gamma \in [0,1[$ and define: $p = \min_{B(0,\gamma)} k$, and $q = 1-p$. Note that $p > 0$; indeed k is a continuous function on the compact set $B(0,\gamma)$ and the support of k is $B(0,1)$, therefore k reaches its non-negative minimum p on $B(0,\gamma)$.

The following lemma gives an upper bound of the probability of disconnectedness of X_C conditionally on Z. The key point of the proof (see Appendix 7.C) first consists in showing that conditionally on Z, X_C somehow 'contains' an Erdős-Rényi random graph with parameters (n_C, p). Then the bound is derived by using the monotonicity of connectedness (see Lemma 7.5).

Lemma 7.2. *Let C be a bounded subset of $\mathbb{R}^d$ with diameter smaller than γh. Then there exists $p > 0$ such that:*

$$P^Z(X_C \text{ is disconnected}) \leq P_{n_C,p}(\text{disconnectedness}).$$

Now we give a bound of the probability of an Erdős-Rényi random graph with parameters (n,p) to be disconnected, adapted from Theorem 1 in Gilbert (1959) and derived from Proposition 2.2.

Lemma 7.3. *Let $p \in [0,1]$ and $q = 1-p$,*

$$P_{n,p}(\text{disconnectedness}) \leq \kappa_q n q^{n-1} \quad \text{where} \quad \kappa_q = \exp\left(\frac{-2e^{-1}}{\sqrt{q}\ln(q)}\right).$$

By combining both previous Lemmas 7.2 and 7.3, we obtain a bound conditional on the number of nodes in C. The final bound of the probability of local connectedness, provided in Proposition 7.7, essentially follows by computing the expectation of both sides of the inequality.

Proposition 7.7. *For all bounded set C with diameter smaller than γh:*

$$P(X_C \text{ is disconnected}) \leq \kappa_q n \alpha_C \exp\left(-p\alpha_C(n-1)\right)$$

where $\kappa_q = \exp\left[-2/(e\sqrt{q}\ln(q))\right]$.

Proof. Using Lemma 7.2, and then Lemma 7.3:

$$\begin{aligned} P(X_C \text{ is disconnected}) &= \mathbb{E}\left(P^Z(X_C \text{ is disconnected})\right) \\ &\leq \mathbb{E}\left(P_{n_C,p}(\text{disconnectedness})\right) \\ &\leq \mathbb{E}\left(\kappa_q n_C q^{n_C-1}\right) = \kappa_q H'_C(q) \end{aligned}$$

where H_C is the probability-generating function of n_C. As n_C is a binomial random variable with parameters (n, α_C), we have, for all $q \in [0,1]$, $H_C(q) = (\alpha_C q + (1-\alpha_C))^n$, so

$$\begin{aligned} P(X_C \text{ is disconnected}) &\leq \kappa_q n \alpha_C \left(\alpha_C q + (1-\alpha_C)\right)^{n-1} \\ &\leq \kappa_q n \alpha_C \left(1 - \alpha_C(1-q)\right)^{n-1} \\ &\leq \kappa_q n \alpha_C \exp\left(-p(n-1)\alpha_C\right). \end{aligned}$$

□

Comments Note that C should be neither too small nor too big so as to prevent X_C from being disconnected. If the diameter is larger than h_n, the connection probability between some nodes may be zero, and there is no non-zero uniform lower bound p for the connection probabilities $(p_{ij})_{i,j\in V_C}$ of the fully parametrized random graph X_C. On the opposite, if the diameter of C vanishes too quickly when n tends to infinity, then α_C vanishes as well. However the probability of connectedness of the Erdős-Rényi model $\mathcal{G}(n,p)$, for some fixed p, tends to one when the number of nodes n tends to infinity. If the average number of nodes of X_C, $n\alpha_C$, does not tend to infinity, there may be not enough nodes in C to make connectedness of X_C guaranteed when n tends to infinity.

7.3.2 From local to global connectedness

In this subsection, we will show that the local connectedness spreads over each t-cluster with high probability for n large enough. That is: the subgraph of

X induced by any t-cluster of f is connected with high probability. Thus each connected component of $X_{\mathcal{L}(t)}$ — the subgraph of X induced by $\mathcal{L}(t)$ — can be mapped to a t-cluster. Therefore, w.h.p., there are less connected components in $X_{\mathcal{L}(t)}$ than geometric t-clusters. This is what the algorithm tries to capture: some relation between both topological and graphical notions of connectedness. Basically, we will show that if nodes are well spread over a t-cluster — i.e. if there is at least one node in each element C of a cover of $\mathcal{L}(t)$ — and X_C is connected, then the subgraph of X induced by this t-cluster is connected.

Our strategy for this proof is to define a convenient cover of $\mathcal{L}(t)$ made of subsets, each included in one component of $\mathcal{L}(t)$ and containing a sufficient number of points to use the local results from Section 7.3.1. Then we combine the results on each subset to get a result for the component as a whole. We first define tilings.

Definition 7.2. *A tiling $\mathcal{T}$ of a compact subset A of $\mathbb{R}^d$ is a cover of A such that the interior of the elements of $\mathcal{T}$ are pairwise disjoint. The elements of tilings (and covers, by abusing the vocabulary) are called tiles.*

The next proposition shows that a tiling with the desirable properties exists w.h.p. All proofs are given in Appendix 7.C.

Proposition 7.8. *There exists a tiling sequence $(\mathcal{T}_n^{\oplus})_n$ and a cover $(\mathcal{T}_n^{\bigcirc})$ such that for n large enough on the intersection of the following events of the list below, there is at most one component of the graph $X_{\mathcal{L}(t)}$ in each topological connected component of $\mathcal{L}(t)$, and hence $\widetilde{Q}_n(t) \leq Q(t)$:*

- *there is at least one of the $(Z_i)_{i\in[n]}$ in each tile of $\mathcal{T}_n^{\oplus}$:*

$$\forall c \in \mathcal{T}_n^{\oplus},\ V_c = c \cap \{Z_i, i \in [n]\} \neq \varnothing;$$

- *X_C is connected for every tile $C \in \mathcal{T}_n^{\bigcirc}$.*

Substantial details about the construction of $\mathcal{T}_n^{\oplus}$ and $\mathcal{T}_n^{\bigcirc}$ are given from Appendix 7.C.2 to 7.C.5. More specifically, according to Lemma 7.7 "Puzzle", they satisfy the following conditions:

- for all $c \in \mathcal{T}_n^{\oplus}$, $\mathrm{Vol}(c) \geq (\rho_n/2)^d$ and
- for all $C \in \mathcal{T}_n^{\bigcirc}$, $\rho_n^d \leq \mathrm{Vol}(C) \leq (3\rho_n/2)^d$

where $\rho_n = \frac{2\gamma h_n}{3\sqrt{d}}$ and $\gamma \in [0, 1[$. In addition to that, any tile of $\mathcal{T}_n^{\oplus}$ shares at least one face with some other tile of $\mathcal{T}_n^{\oplus}$. Thus the proof of Proposition 7.8 is essentially the same as that of Proposition 7.20.

Note also that, on the intersection of the events of the proposition, if two nodes are in the same geometric t-cluster, then they necessarily belong to the same connected component of $X_{\mathcal{L}(t)}$, as there is no more than one component per cluster. This remark will be fruitful for the study of the connected components as a classification in Section 7.4.

7.3.3 Bounds

In the proof of Theorem 7.2, we will need some bounds, which are stated here. Their proof is given in Appendix 7.C.6.

The first two propositions provide a control of both the probability for a subtile from $\mathcal{T}_n^{\oplus}$ to be empty (Prop. 7.9) and the probability for the subgraph induced by a tile of $\mathcal{T}_n^{\bigcirc}$ to be disconnected (Prop. 7.10). The two following ones provide upper bounds of the number of (sub-)tiles in the sets $\mathcal{T}_n^{\oplus}$ (Prop. 7.11) and $\mathcal{T}_n^{\bigcirc}$ (Prop. 7.12), constructed in Lemma 7.7 and the properties of which have just been exposed in the previous subsection.

Probability of a subtile from $\mathcal{T}_n^{\oplus}$ being empty

Proposition 7.9. *For all $c \in \mathcal{T}_n^{\oplus}$:*

$$P(V_c = \varnothing) \leq \exp\left(-\frac{t}{2^d} n \rho_n^d\right)$$

Probability of disconnectedness of the subgraph induced by a tile of $\mathcal{T}_n^{\bigcirc}$

Proposition 7.10. *For all $C \in \mathcal{T}_n^{\bigcirc}$ and n large enough:*

$$P(X_C \text{ is disconnected}) \leq \kappa_q t n \rho_n^d \exp\left(-p \|f\|_\infty \left(\frac{3}{2}\right)^d (n-1)\rho_n^d\right).$$

Number of subtiles in $\mathcal{T}_n^{\oplus}$

Proposition 7.11. *For some $K_2 > 0$, for n large enough:*

$$\mathrm{Card}(\mathcal{T}_n^{\oplus}) \leq K_2 \left(\frac{\rho_n}{2}\right)^{-d}$$

Number of tiles in the auto-cover $\mathcal{T}_n^{\bigcirc}$

Proposition 7.12. *For n large enough:*

$$\mathrm{Card}(\mathcal{T}_n^{\bigcirc}) \leq K_2 \left(\frac{\rho_n}{2}\right)^{-d}$$

7.3.4 Proof of Theorem 7.2

We are now using an union bound to conclude. According to Proposition 7.8, it is first established that:

$$\begin{aligned} P(\widetilde{Q}_n(t) > Q(t)) &\leq P\left[\left(\bigcup_{c\in\mathcal{T}_n^{\oplus}}\{V_c = \varnothing\}\right)\bigcup\left(\bigcup_{C\in\mathcal{T}_n^{\circ}}\{X_C \text{ is disconnected}\}\right)\right] \\ &\leq \sum_{c\in\mathcal{T}_n^{\oplus}} P(V_c = \varnothing) + \sum_{C\in\mathcal{T}_n^{\circ}} P(X_C \text{ is disconnected}) \end{aligned}$$

The probabilities inside the both sums are uniformly bounded using Propositions 7.9 and 7.10. The cardinalities of $\mathcal{T}_n^{\oplus}$ and $\mathcal{T}_n^{\circ}$ can be bounded using Propositions 7.11 and 7.12, so we get:

$$\begin{aligned} P\left(\widetilde{Q}_n(t) > Q(t)\right) &\leq \mathrm{Card}(\mathcal{T}_n^{\oplus})\exp\left(-\frac{t}{2^d}n\rho_n^d\right) \\ &\quad + \mathrm{Card}(\mathcal{T}_n^{\circ})\kappa_q t n\rho_n^d \exp\left(-p\left\|f\right\|_{\infty}\left(\frac{3}{2}\right)^d(n-1)\rho_n^d\right) \\ &\leq K_2\left(\frac{\rho_n}{2}\right)^{-d}\exp\left(-\frac{t}{2^d}n\rho_n^d\right) \\ &\quad + K_2\left(\frac{\rho_n}{2}\right)^{-d}\kappa_q t n\rho_n^d \exp\left(-p\left\|f\right\|_{\infty}\left(\frac{3}{2}\right)^d(n-1)\rho_n^d\right) \\ &\leq K_0 n\exp(-K_3 n\rho_n^d) \end{aligned}$$

where K_0 and K_3 are non-negative constants which both depend on d, f, k and t. Moreover $\rho_n = \frac{\gamma h_n}{\sqrt{d}}$, and the inequality derived above holds for all $\gamma \in [0,1[$. However note that $p = \min_{B(0,\gamma)} k \ > 0$ depends on γ and $p \xrightarrow[\gamma\to 1]{} 0$. $\gamma \in [0,1[$ can be taken such that $p\gamma^d$ is maximal for example. Finally it is established that:

$$P\left(\widetilde{Q}_n(t) > Q(t)\right) \leq K_0 n\exp(-K_1 n h^d)$$

where K_1 a non-negative constant depends on d, f, k and t.

7.4 Classification via the connected components of $\widehat{X}(t)$

7.4.1 Notations and assumptions

Connected components of a graph provide a natural partition of the nodes. In this section we use the components of $\widehat{X}(t)$ given by the algorithm as a classification

of the high-density nodes. The support of the classification $J_n(t)$ is also random, and is estimated by $\widehat{J}_n(t)$. In addition to misclassification errors on node pairs of $\mathcal{L}(t)$, there may be errors in the support. Some nodes may be classified even though they are not in the t-level set, and conversely some nodes from the t-level set may be removed by the thresholding. We then address two kinds of errors:

- thresholding error, quantifying the fact that a node might be in $\widehat{J}_n(t)$ but not in $J_n(t)$ (*false positive*), or might be on the contrary not in $\widehat{J}_n(t)$ but in $J_n(t)$ (*false negative*).
- cluster classification error, quantifying the fact that two nodes selected by the algorithm might be either in the same graphic t-component but not in the same geometric t-cluster (*false positive*), or in distinct graphic t-components but in the same geometric t-cluster (*false negative*).

It is actually hard to deal directly with these errors. We will give some results about them, but most of the results deal rather with approximate errors.

7.4.2 Thresholding error

Definition 7.3. *Let $\varepsilon \geq 0$. A node $i \in [n]$ is said to be a ε-false positive for the thresholding error if $i \in \widehat{J}_n(t) \backslash J_n(t-\varepsilon)$, and ε-false negative if $i \in J_n(t+\varepsilon) \backslash \widehat{J}_n(t)$. The ε-thresholding error is the number of ε-false positives and ε-false negatives.*

The main theorem of this section is devoted to bound from below the probability that the ε-thresholding error is zero. We will also deduce from the proof a result about the probability that a given node is a false positive or negative, but no result about thresholding error, or equivalently, 0-thresholding error. However we will assume that $\varepsilon_n \to 0$ to approximate the probability that there is no thresholding error. We will need the following assumption:

Assumption 7.4. *There exists $\mu \in]0,1]$ such that from n large enough:*

$$s_{h_n} \leq (1-\mu)\varepsilon_n$$

Assumption 7.4 is true for example if f is Lipschitz continuous with Lipschitz constant L, and $\varepsilon_n \geq Lh_n/(1-\mu)$. Indeed we have $s_{h_n} \leq Lh_n \leq (1-\mu)\varepsilon_n$. Note that ε_n cannot thus vanish quicker than h_n. The interpretation of this fact is that the maximal bias of the estimation density by the degrees cannot decrease quicker than the bandwidth h_n. This was expected, because in usual pointwise density estimation, the bias is proportional to h_n (under the assumption of $\mathcal{C}^1$ regularity).

For all $i \in [n]$, let $M_n^{i,\varepsilon}(t)$ be the event "there is no ε-thresholding error on i"

and $M_n^{\varepsilon}(t)$ "the ε-thresholding error equals zero", that is:

$$M_n^{i,\varepsilon}(t) = \underbrace{(\{T_i \geq t\} \cap \{Z_i \in \mathcal{L}(t-\varepsilon)\})}_{\text{" } i \text{ is no } \varepsilon\text{-false positive"}} \cup \underbrace{(\{T_i < t\} \cap \{Z_i \notin \mathcal{L}(t+\varepsilon)\})}_{\text{" } i \text{ is no } \varepsilon\text{-false negative"}},$$

$$M_n^{\varepsilon}(t) = \bigcap_{i \in [n]} M_n^{i,\varepsilon}(t)$$

Theorem 7.3. *Under Assumption 7.4, from n large enough:*

$$P(M_n^{\varepsilon_n}(t)) \geq 1 - \frac{2}{(1-\mu_f(t)) \bigvee \mu_f(t+\varepsilon_0)} n \exp\left(-\frac{\mu^2}{2} \frac{n h_n^d \varepsilon_n^2}{\|f\|_\infty + \varepsilon_n}\right) \tag{7.2}$$

And for all $i \in [n]$, for some positive constant K_4:

$$P(M_n^{i,0}(t)) \geq 1 - \left(\frac{2}{(1-\mu_f(t)) \bigvee \mu_f(t+\varepsilon_0)} \exp\left(-\frac{\mu^2}{2} \frac{n h_n^d \varepsilon_n^2}{\|f\|_\infty + \varepsilon_n}\right) + K_4 \varepsilon_n\right) \tag{7.3}$$

Some remarks. We give a result about the probability that a given node is a false positive or negative, but no result about the global (0-)thresholding error over all nodes. The result cannot be directly expanded with a simple union bound over all nodes. This would require to be on the event "there is no node in the set $\mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)$" but the probability of this event does not tend to zero. Let us prove this. On the one hand the order of magnitude of the measure of $\mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)$ is ε_n (the geometric Proposition 5.6 provides bounds) and that of the average number of points being in this set is then $n\varepsilon_n$. The probability that none of the nodes is in this set is:

$$P\left(\bigcap_{i \in [n]} \{Z_i \notin \mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)\}\right) = (1 - K_5 \varepsilon_n)^n = \exp\left(-K_5 n \varepsilon_n + o(n\varepsilon_n)\right)$$

On the other hand, the first term of the bound of Theorem 7.4.(3) requires that we assume $\varepsilon_n^2 n h_n^d \to +\infty$, which implies that $n\varepsilon_n \to +\infty$, because from n large enough, $n\varepsilon_n \geq n\varepsilon_n^2 h_n^d$.

Nevertheless we give a result about the global ε_n-thresholding error and we can assume that $\varepsilon_n \to 0$ to approximate the probability that there is no thresholding error. Thus (ε_n) is allowed to vanish, but slowly enough so that the probability of having at least one ε_n-false positive node tends to zero. More precisely, $\varepsilon_n^2 n h_n^d$ should still tend to infinity.

However we did not study the limit of the classification. In particular, it has been not proved here that the probability to have an ε_n-thresholding error

converges to the probability to have a 0-thresholding error. Finally, note that under Assumption 7.4, h_n is clearly vanishing when ε_n tends to zero, which seems natural from a density estimation point of view: the maximal accepted bias ε_n cannot vanish while the bandwidth h_n of the kernel estimator f_n does not.

As a conclusion, there is thus a compromise between the expected quality of the thresholding and the rate of convergence of the probability to get this quality: if one wants higher quality, one will wait longer to reach it. This comes directly from the density estimation problem: if one want less bias, one has to take a smaller bandwidth.

Probability of degree deviation

To prove Theorem 7.3, we need to control the probability that the normalized degree of a node i exceeds t, when the node is not located in $\mathcal{L}(t-\varepsilon_n)$, i.e. the density at the node position is less than $t-\varepsilon_n$. The proof of the following proposition is given in Appendix 7.E.

Proposition 7.13. *Under Assumption 7.4, for n large enough:*

$$P^{Z_i \notin \mathcal{L}(t-\varepsilon_n)}(T_i \geq t) \leq \frac{1}{1-\mu_f(t)} \exp\left(-\frac{\mu^2}{2}\frac{nh_n^d\varepsilon_n^2}{\|f\|_\infty + \varepsilon_n}\right)$$
$$P^{Z_i \in \mathcal{L}(t+\varepsilon_n)}(T_i < t) \leq \frac{1}{\mu_f(t+\varepsilon_0)} \exp\left(-\frac{\mu^2}{2}\frac{nh_n^d\varepsilon_n^2}{\|f\|_\infty + \varepsilon_n}\right)$$

Remark 1. Note that for any non-singleton borelian set A, conditionally on $\{Z_i \in A\}$, variables $(X_{ij})_{j\in[n]}$ are not independent variables, and then T_i is not a binomial distributed variable. In the proof, we have first conditioned on Z_i instead of $\{Z_i \notin \mathcal{L}(t-\varepsilon_n)\}$ so as to make possible to use a concentration inequality, which mostly requires independance.

Remark 2. A consequence of the proof is that the limit when n tends to infinity of the conditional density $P(T_i \geq t \mid f(Z_i) = u)$ and denoted by $\pi_t(u)$ has a phase transition at t. Indeed it can be derived that $\pi_t(u) = 0$ for $u < t$ and $\pi_t(u) = 1$ for $u > t$.

Remark 3. A bound similar to that found in Theorem 7.1, for the control of the underestimation by $\widehat{Q}_n(t)$, can be retrieved with the first inequality of Proposition 7.13. Indeed, there is no underestimation whenever there is no node in $\{f \leq t-\varepsilon\}$ (for both $\varepsilon > 0$ and h small enough, see Assumption 7.3.(b)) such that $T_i \geq t$ (see Proposition 7.3), that is there is no ε-false positive for the thresholding error.

Proof of Theorem 7.3

Proof.

1. Using an union bound and the bounds of degree deviation from Proposition 7.13:

$$\begin{aligned}
P\left(\overline{M_n^{\varepsilon_n}(t)}\right) &= P\left(\bigcup_{i\in[n]} \left(\{T_i < t\} \cup \{Z_i \notin \mathcal{L}(t-\varepsilon_n)\}\right) \cap \left(\{T_i \geq t\} \cup \{Z_i \in \mathcal{L}(t+\varepsilon_n)\}\right)\right) \\
&= P\left(\bigcup_{i\in[n]} \left(\{T_i < t\} \cap \{Z_i \in \mathcal{L}(t+\varepsilon_n)\}\right) \cup \left(\{T_i \geq t\} \cap \{Z_i \notin \mathcal{L}(t-\varepsilon_n)\}\right)\right) \\
&\leq \sum_{i\in[n]} P^{Z_i\in\mathcal{L}(t+\varepsilon_n)}(T_i < t) P(Z_i \in \mathcal{L}(t+\varepsilon_n)) \\
&\qquad + P^{Z_i\notin\mathcal{L}(t-\varepsilon_n)}(T_i \geq t) P(Z_i \notin \mathcal{L}(t-\varepsilon_n)) \\
&\leq \frac{2}{(1-\mu_f(t))\bigvee \mu_f(t+\varepsilon_0)} n \exp\left(-\frac{\mu^2}{2}\frac{nh_n^d\varepsilon_n^2}{\|f\|_\infty + \varepsilon_n}\right).
\end{aligned}$$

2. For all $i \in [n]$,

$$\begin{aligned}
P\left(\overline{M_n^{i,0}(t)}\right) &= P\left(\overline{M_n^{i,0}(t)} \cap \{f(Z_i) \geq t+\varepsilon_n\}\right) \\
&\quad + P\left(\overline{M_n^{i,0}(t)} \cap \{t-\varepsilon_n \leq f(Z_i) < t+\varepsilon_n\}\right) + P\left(\overline{M_n^{i,0}(t)} \cap \{f(Z_i) < t-\varepsilon_n\}\right)
\end{aligned}$$

Recall that (see above the proof of the first part of the theorem):

$$\overline{M_n^{i,0}(t)} = \left(\{T_i < t\} \cap \{Z_i \in \mathcal{L}(t)\}\right) \cup \left(\{T_i \geq t\} \cap \{Z_i \notin \mathcal{L}(t)\}\right).$$

Hence:

$$\begin{aligned}
P\left(\overline{M_n^{i,0}(t)}\right) &= P\left(\{T_i < t\} \cap \{f(Z_i) \geq t+\varepsilon_n\}\right) \\
&\quad + P\left(\{T_i \geq t\} \cap \{t-\varepsilon_n \leq f(Z_i) < t+\varepsilon_n\}\right) + P\left(\{T_i \geq t\} \cap \{f(Z_i) < t-\varepsilon_n\}\right) \\
&\leq P^{Z_i\in\mathcal{L}(t+\varepsilon_n)}(T_i < t) P(Z_i \in \mathcal{L}(t+\varepsilon_n)) + P^{Z_i\notin\mathcal{L}(t-\varepsilon_n)}(T_i \geq t) P(Z_i \notin \mathcal{L}(t-\varepsilon_n)) \\
&\quad + P\left(Z_i \in \mathcal{L}(t-\varepsilon_n) \setminus \mathcal{L}(t+\varepsilon_n)\right) \\
&\leq \frac{2}{(1-\mu_f(t))\bigvee \mu_f(t+\varepsilon_0)} \exp\left(-\frac{\mu^2}{2}\frac{nh_n^d\varepsilon_n^2}{\|f\|_\infty + \varepsilon_n}\right) + K_4\varepsilon_n
\end{aligned}$$

At the last line for the last term, we are using the geometric Proposition 5.6, bounding the volume of the set $\mathcal{L}(t-\varepsilon_n) \setminus \mathcal{L}(t+\varepsilon_n)$. □

7.4.3 Cluster classification error

False positive

Definition 7.4. *Let $\varepsilon \geq 0$. A node pair $\{i,j\} \in [n]$ is said to be a ε-false positive for the cluster classification error if node i is in the same component of $\widehat{X}(t)$ as node j and Z_i is not in the same $(t-\varepsilon)$-cluster as Z_j.*

The first result of this section (Theorem 7.4) states that, on the one hand, for n large enough, nodes selected by the algorithm may be not located in $\mathcal{L}(t)$, but nevertheless in $\mathcal{L}(t-\varepsilon_n)$ with high probability. On the other hand, it states that there is no ε_n-false positive in $\widehat{J}_n(t)$ with high probability. Let us define for all $i,j \in [n]$ and $\varepsilon \geq 0$ the event that the node pair $\{i,j\}$ is not a ε_n-false positive and denoted by $F_{ij}^{+,\varepsilon}(t)$. It means that either i and/or j is removed by the algorithm or that they are in the same component graph $\widehat{X}(t)$ and Z_i is in the same $(t-\varepsilon_n)$-cluster as Z_j. We define the corresponding events:

$$\begin{aligned} F_{ij}^{+,\varepsilon}(t) &= \{i \mathrel{\widehat{\sim}_t} j \Rightarrow Z_i \sim_{t-\varepsilon} Z_j\} \\ &= \{i \notin \widehat{J}_n(t)\} \cup \{j \notin \widehat{J}_n(t)\} \cup \{i \mathrel{\widehat{\sim}_t} j,\ Z_i \sim_{t-\varepsilon} Z_j\} \\ \text{and} \quad F^{+,\varepsilon}(t) &= \bigcap_{1 \leq i < j \leq n} F_{ij}^{+,\varepsilon}(t). \end{aligned}$$

The second result is just a consequence of the first one. It means that there is obviously no false positive at level t among nodes from $\mathcal{L}(t)$. Define these events:

$$\widetilde{F}_{ij}^{+,\varepsilon}(t) = \{(\{i,j\} \subset J_n(t) \text{ and } i \mathrel{\widehat{\sim}_t} j) \Rightarrow Z_i \sim_t Z_j\} \text{ and } \widetilde{F}^{+,\varepsilon}(t) = \bigcap_{1 \leq i < j \leq n} \widetilde{F}_{ij}^{+,\varepsilon}(t)$$

Unlike the first two results, the last one deals with each pair of nodes separately. For n large enough, if a given pair of nodes is selected by the algorithm, then both nodes are in $\mathcal{L}(t)$ and the pair is not a false positive at level t with high probability.

Theorem 7.4. *Under Assumptions 7.2 and 7.3:*

1. $P(F^{+,\varepsilon_n}(t)) \geq 1 - 6n \exp\left(-\frac{1}{32}\frac{\varepsilon_n^2 n h_n^d}{\|f\|_\infty + \varepsilon_n}\right)$
2. $P(\widetilde{F}^{+,\varepsilon_n}(t)) \geq 1 - 6n \exp\left(-\frac{1}{32}\frac{\varepsilon_n^2 n h_n^d}{\|f\|_\infty + \varepsilon_n}\right)$
3. *For some positive constant K_4, and for all $i,j \in [n]$,*

$$P(F_{ij}^{+,0}) \geq 1 - \left(6 \exp\left(-\frac{1}{32}\frac{\varepsilon_n^2 n h_n^d}{\|f\|_\infty + \varepsilon_n}\right) + K_4 \varepsilon_n\right)$$

Some remarks. Like in the paragraph 7.4.2 about the thresholding error, we cannot directly infer any bound for the probability of occurrence of at most one (0-)false positive with a simple union bound, for the same reason: $n\varepsilon_n$ does not vanish at all.

Nevertheless we give a result about the probability of occurrence of at most one ε_n-false positive. ε_n is then assumed to be vanishing. Firstly possible ε_n are restricted because we are using results of Section 7.2 and we need Assumptions 7.3 and 7.2 for that. In particular, $h_n \leq \eta_{\varepsilon_n}$ is true if for example f is Lipschitz continuous with Lipschitz constant L and $\varepsilon_n \geq 2Lh_n$. Secondly the first two results of Theorem 7.4 hold if $\varepsilon_n \to 0$, whenever $\varepsilon_n^2 n h_n^d \to +\infty$.

Once more, like in 7.4.2, note that h_n must vanish whenever ε_n is assumed so. Furthermore the assumption $h_n \leq \delta_{t-\varepsilon_n}$ is not a problem because h_n and ε_n tend to zero and $\varepsilon \mapsto \delta_{t-\varepsilon}$ is an increasing function.

The same conclusion than for the thresholding error is possible: there is a compromise between the expected quality of classification and the rate of convergence of the probability to reach it. The higher it is expected, the longer it is to get it.

Proof of Theorem 7.4

1. is a consequence of Proposition 7.2, of (7.8) in the proof of Proposition 7.3, and of the upper bound of Proposition 7.6.(2).

Indeed $\{U_n \leq \varepsilon\}$ is included in two events, and therefore it is also included in their intersection; and in fact this intersection is $F^{+,\varepsilon}(t)$:

$$\{U_n \leq \varepsilon_n\} \subset \left\{\widehat{J}_n(t) \subset J_n(t-\varepsilon_n)\right\} \cap \left(\bigcap_{i,j \in [n]} \left\{ i,j \in \widehat{J}_n(t) \text{ and } i \,\widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon_n} Z_j \right\} \right)$$
$$= \left\{ \forall i,j \in \widehat{J}_n(t),\ Z_i, Z_j \in \mathcal{L}(t-\varepsilon_n) \text{ and } i \,\widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon_n} Z_j \right\}$$

Then the probability of the event of the proposition is larger than:

$$P(U_n \leq \varepsilon_n) = 1 - P(U_n > \varepsilon_n) \geq 1 - 6n \exp\left(-\frac{1}{32} \frac{\varepsilon_n^2 n h_n^d}{\|f\|_\infty + \varepsilon_n} \right)$$

2. is a consequence of 1. For all $i,j \in [n]$, if $\varepsilon_n \leq \varepsilon_0$:

$$Z_i, Z_j \in \mathcal{L}(t) \text{ and } Z_i \sim_{t-\varepsilon_n} Z_j \Rightarrow Z_i \sim_t Z_j$$

Since $Q(t) = Q(t-\varepsilon_n)$ and each geometric $(t-\varepsilon_n)$-cluster has a non-empty intersection with $\mathcal{L}(t)$, there is no more than one geometric t-cluster in each $(t-\varepsilon_n)$-cluster. As a conclusion, if $Z_i, Z_j \in \mathcal{L}(t)$ are in the same $(t-\varepsilon_n)$-cluster, then they are in the same t-cluster as well.

3. is a consequence of 1, of the upper bound of Proposition 7.6.(2) and of the geometric Proposition 5.6.

It follows from the total probability formula that:

$$\begin{aligned} P(i \widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon_n} Z_j) &= P\left(\{i \widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon_n} Z_j\} \cap \{Z_i, Z_j \in \mathcal{L}(t)\}\right) \\ &\quad + P\left(\{i \widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon_n} Z_j\} \cap (\{Z_i \notin \mathcal{L}(t)\} \cup \{Z_j \notin \mathcal{L}(t)\})\right) \\ &\leq P(i \widehat{\sim}_t j \Rightarrow Z_i \sim_t Z_j) + 2P(Z_1 \in \mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)) \end{aligned}$$

Moreover, for all $i, j \in [n]$, we first get from this inequality 1:

$$P(i \widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon_n} Z_j) \geq P(U_n \leq \varepsilon_n) = 1 - P(U_n > \varepsilon_n)$$

By combining the two previous inequalities, we have:

$$P(i \widehat{\sim}_t j \Rightarrow Z_i \sim_t Z_j) \geq 1 - P(U_n > \varepsilon_n) - 2P(Z_1 \in \mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t))$$

First term of the right hand side is directly bounded in Proposition 7.6.(2). The second one requires the geometric Proposition 5.6 which bounds the volume of the set $\mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)$ and therefore the probability of one of the nodes to be located in it:

$$\begin{aligned} P(Z_1 \in \mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)) &= \int_{\mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)} f(z) dz \\ &\leq t \, \mathrm{Vol}(\mathcal{L}(t-\varepsilon_n) \backslash \mathcal{L}(t)) \end{aligned}$$

Finally, the right hand side can be bounded as follows:

$$P(i \widehat{\sim}_t j \Rightarrow Z_i \sim_t Z_j) \geq 1 - 6 \exp\left(-\frac{1}{32} \frac{\varepsilon_n^2 n h_n^d}{\|f\|_\infty + \varepsilon_n}\right) - K_4 \varepsilon_n$$

False negative

Theorem 7.4 now deals with false negative of the classification given by the t-components of $X_{\mathcal{L}(t)}$, not directly those of $\widehat{X}(t)$. Let $F_n^-(t)$ the event that there is no false negative in the classification based on these connected components of $X_{\mathcal{L}(t)}$ and with respect to the t-clusters, that is:

$$F_n^-(t) = \{\forall i, j \in J_n(t) \; i \widehat{\nsim}_t j \Rightarrow Z_i \nsim_t Z_j\}.$$

Theorem 7.5.

$$P(F_n^-(t)) \geq 1 - K_0 n \exp(-K_1 n h_n^d)$$

The proof is given by the remark after Proposition 7.8, and the bound is the same as that of Theorem 7.2.

7.5 Simulation study

The goals of this study are comparing the empirical frequencies of under- and overestimation with theoretical bounds of the main theorems of the chapter, as well as illustrating and analyzing what the algorithm does.

7.5.1 Designs

Note first that in the whole study, we will use the Epanechnikov kernel on $\mathbb{R}^2$ as connection function k:

$$k(z) = \frac{2}{\pi}(1 - \|z\|^2)\mathbb{1}_{B(0,1)}(z)$$

In fact we have not detected any major effect of the kernel choice on the statistics presented in this study.

First design. Let $\phi_{a,b}$ the density of the Beta distribution with parameters (a, b), i.e. $\phi_{a,b}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\mathbb{1}_{[0,1]}(x)$ where B is the Beta function. Let us define the density f_1 on $\mathbb{R}^2$ as follows:

$$f_1(x, y) = \frac{1}{4}\phi_{a,b}(|x|)\phi_{a,b}(|y|)$$

This first design is simple, but a bit more difficult for the density estimation than a gaussian mixture, because of the singularity along the coordinates axis. The density level chosen is $t = 1.2$, and there are four clusters, whose boundary is the red density isoline on Figure 7.1. The size of the graphs is going from $n = 1000$ nodes up to $n = 40000$ nodes. The connection radius is $h = 0.15$, and satisfies the assumptions of the theorems given in the chapter. We will compare the convergence rates in the simulation and the theoretical bounds. Note that two other values of h have been tested; they will be useful to comment the robustness of the algorithm in the next section. We will also show some plots, called *clustering profiles*, deciphering the latent structure of the graph in some sense explained in 7.5.4.

Design with complex-shaped cluster. One of the advantages of the algorithm presented in this chapter is to be able to process complex-shaped clusters as well as clusters of a parametric mixture, with no additional cost. We chose in this design a density f_2 any gaussian mixture could hardly fit.

The sample $(Z_i)_{i\in[n]}$ is drawn as follows: $Z_i = (R_i\cos(\theta_i), R_i\sin(\theta_i))$ such that R_i is a gaussian mixture as below and θ_i is uniform over $[0, 2\pi]$:

$$R_i \sim 0.1\mathcal{N}(0, 1) + 0.9\mathcal{N}(5, 1) \text{ and } \theta_i \sim \mathcal{U}([0, 2\pi])$$

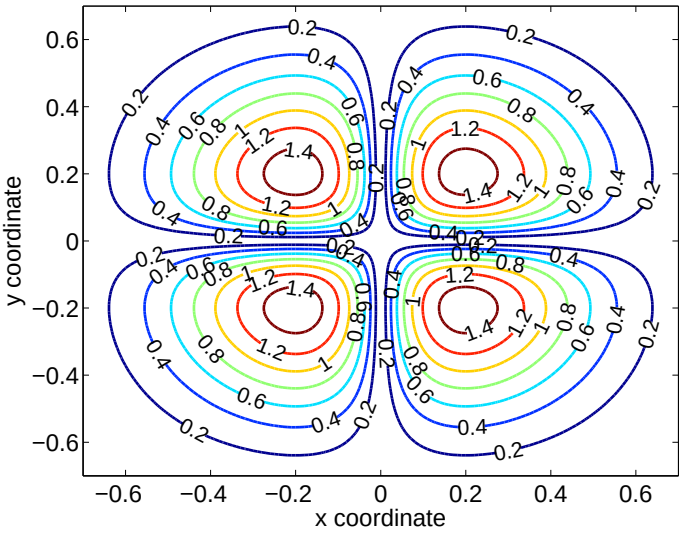


Figure 7.1: Density contour levels of f_1

Two values of h have been tested: $h = 0.9$ and $h = 2$. For this design we will just show the clustering profile.

7.5.2 Comparison with theoretical bounds in the first design

Figure 7.3 first shows the average of the estimator $\widehat{Q}_n(t)$ over 200 graphs. The convergence is reached at $n = 11000$ nodes. According to Figure 7.4, the bounds of the underestimation probability given in Theorem 7.1 are rather pessimistic.

The situation turns out to be much more complex for the overestimation. Recall that the bound of Theorem 7.2 (dashed lines on Figure 7.5) deals just with the probability of overestimation of the pseudo-estimator $\widetilde{Q}_n(t)$ (solid lines and circles). This bound is a little bit pessimistic in comparison with the frequency of overestimation of this pseudo-estimator. On the opposite, it is not a good bound at all for the probability of non-overestimation of $\widehat{Q}_n(t)$ (solid lines and plus signs). The frequency of overestimation vanishes after the bound. From the connectedness point of view, the consequence is that $\widehat{X}(t)$ and $X_{\mathcal{L}(t)}$ thus seem to be very different.

On Figure 7.6, the separate rates of false positive nodes and false negative nodes of the thresholding error are shown. The bounds are also displayed for false positive rates[2]. Note that these rates correspond to 0-false positive (respectively

2. Note that these average rates and the theoretical bounds cannot be directly compared except when they both equal zero. Indeed when the average rate equals zero, it means that the frequency of the event "there is at least one false positive node " is zero, which does not hold

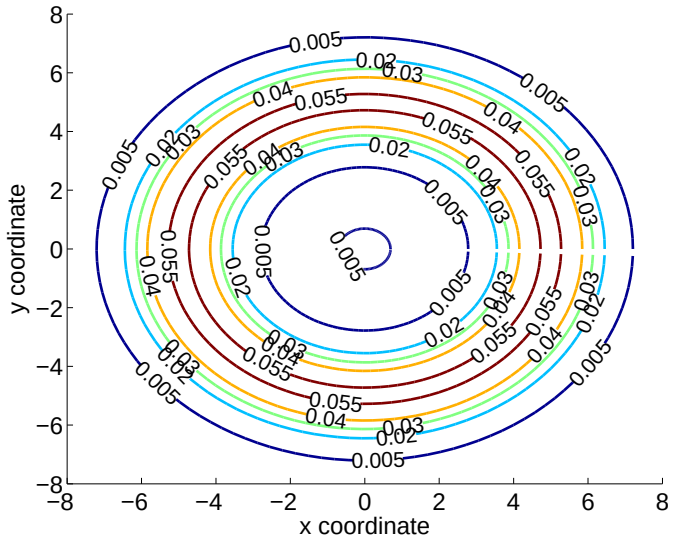


Figure 7.2: Density contour levels of f_2

negative) defined in 7.4.2, although the bounds holds for ε-false positive (respectively negative). Despite of this, it is even satisfied for 0-false positive. However, as we will explain later in the analysis, the 0-false negative rate does not seem to vanish especially for large values of h. The theoretical bounds are clearly not satisfied.

On Figure 7.7, it can be observed that the 0-cluster classification error rates are very small. Here again, the bounds for ε_n-false positive hold even for 0-false positive rates. It is also true for false negative for values of h below 0.2.

7.5.3 Algorithm robustness

Here we analyze and identify some issues experienced in the study in both regimes of low and high graph density — small and large h — in order to finally propose a more robust algorithm and perspectives for future work.

Underestimation with large h In Section 7.2, it is proved that underestimation comes either from the fact that the graph $\widehat{X}(t)$ does not cover one of the clusters or from connections between distinct clusters. Figure 7.9 shows the average number of clusters covered by $\widehat{X}(t)$, and it can be seen that in the case $h = 0.22$, there are very often uncovered clusters. In Section 7.2, it also explained that the graph $\widehat{X}(t)$ does not cover all clusters, either because no node is fallen

otherwise.

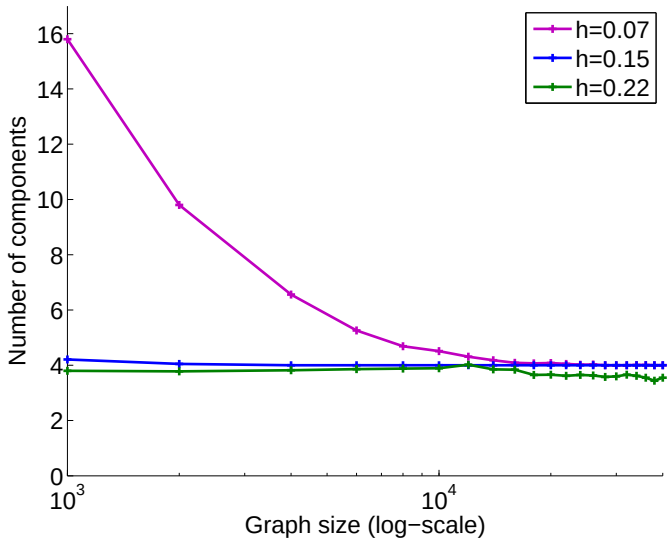


Figure 7.3: Average estimator $\widehat{Q}_n(t)$ as function of the graph size n

in one of the clusters, or because the normalized degree has underestimated the density, i.e. all nodes of the cluster are false negative of the thresholding error (see Proposition 7.4). The former is very unlikely for the range of n considered here, whereas the latter has a high probability when h is large, because of the large bias of the associated kernel estimator f_n. Remind that the bias does not vanish asymptotically when h is constant. Even though it is not necessary to prevent from underestimation — see Assumptions 7.2 and 7.3 made to prove Theorem 7.1 — h especially needs to be smaller than η_ε) so that the kernel estimator sufficiently renders variations of f_1. Figure 7.8 shows that it may be not true for example with $h = 0.22$. The kernel estimator f_n based on a sample of $n = 7000$ points with $h = 0.22$ and the density f_1 are both plotted cut by a vertical hyperplane; in particular, note that the kernel estimator is under the threshold on the inner boundary of $\mathcal{L}(t)$ and cannot even reach $t = 1.2$ in one of the clusters. As a consequence, the estimated level set $\{f_n \geq t\}$, denoted by $\widehat{\mathcal{L}}_n(t)$ is much smaller than $\mathcal{L}(t)$. Figure 7.8 also shows a realization of the graph under this model, and the isoline of f_n at level $t = 1.2$. The level set f_n has actually only one connected component. Moreover there is one cluster which is not covered by $\widehat{X}(t)$, i.e. where the nodes are all false negative.

On the one hand, in this model large values of h entail high false negative rates of the thresholding error which may not vanish when n tends to infinity for very large h. On the other hand, there are very few false positive nodes, since $\widehat{\mathcal{L}}_n(t)$ is very small, see Figure 7.6. Therefore in this model, there cannot be false positive nodes in low density regions which may create links between distinct clusters (see

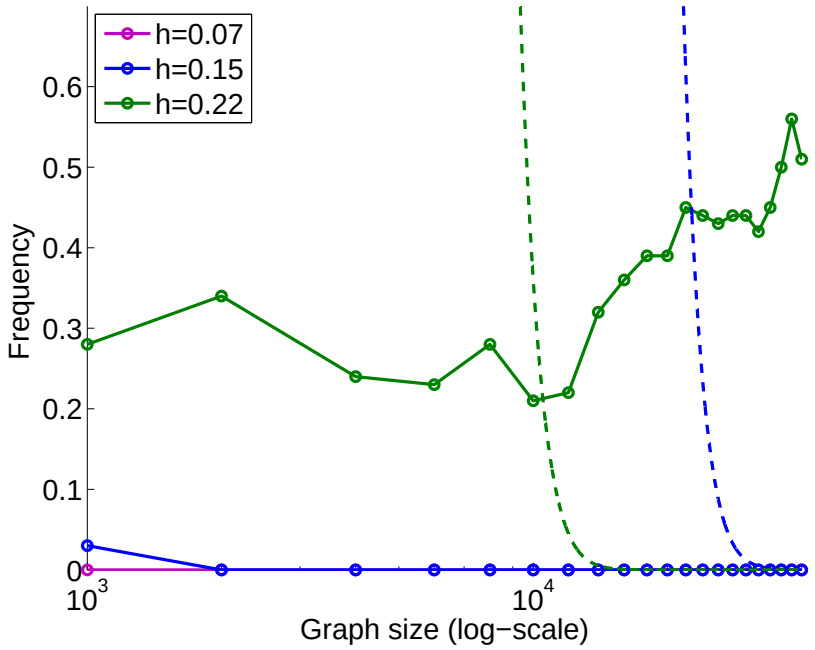


Figure 7.4: Underestimation frequency of $\widehat{Q}_n(t)$ (—) and bound (- - -) of Theorem 7.1 as function of n

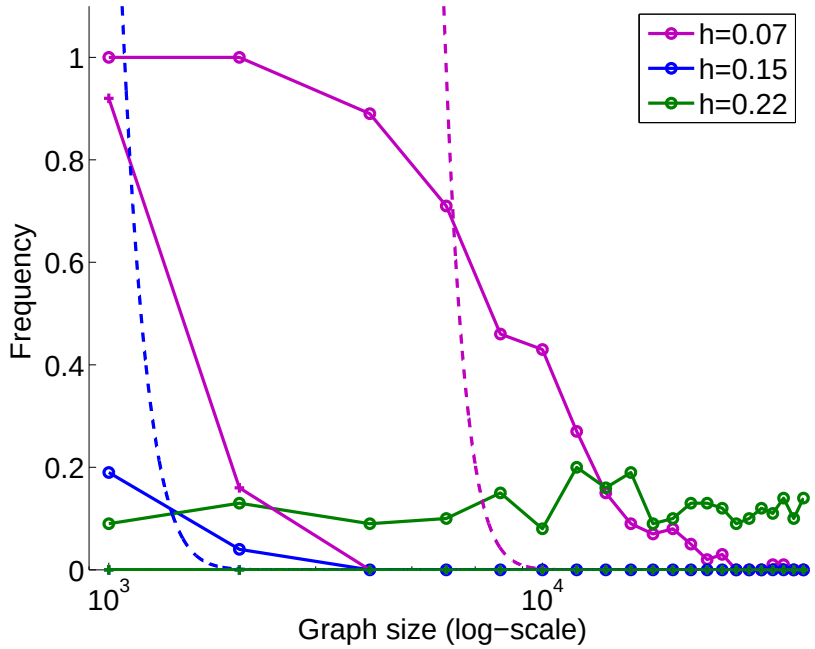
Proposition 7.2).

Note also that in this model, $\widehat{\mathcal{L}}_n(t)$ shrinks when h increases, and is especially smaller than $\mathcal{L}(t)$. Hence the gap between the estimated clusters is larger than δ_t, which actually compensates the increase of h. It can be deduced from Figure 7.9 (right), that if $\widehat{\mathcal{L}}_n(t)$ was closer to $\mathcal{L}(t)$, there could be direct edges between distinct clusters (see Proposition 7.3).

Overestimation with small h In this paragraph, we are illustrating the fact that overestimation essentially comes from very small components, especially when h is small.

Figure 7.6 first shows that there are a lot of false positive nodes if h is too small for the graph size n. As a consequence, these false positive nodes, as it can be observed on Figure 7.10, form small components outside the t-level set of f_1. In fact they belong to the t-level set of the density estimator f_n: one of the main problem thus comes from density estimation, i.e. from the selection of the nodes of $\widehat{X}(t)$ by thresholding. Few nodes in a small region of the latent space can create a peak in the density estimator. On the opposite, an empty region can create a hole even in a t-cluster, as it happens on Figure 7.10 in the second cluster on the right; it can disconnect components which are in opposite sides of a cluster. The

Figure 7.5: Overestimation frequency of $\widehat{Q}_n(t)$ and of $\widetilde{Q}_n(t)$ as functions of n, and bound of Theorem 7.2



$\circ$: $\widehat{Q}_n(t)$ $+$: $\widetilde{Q}_n(t)$ - - -: bound

algorithm is very sensitive to such small components, which artificially put up the estimator of the number of clusters.

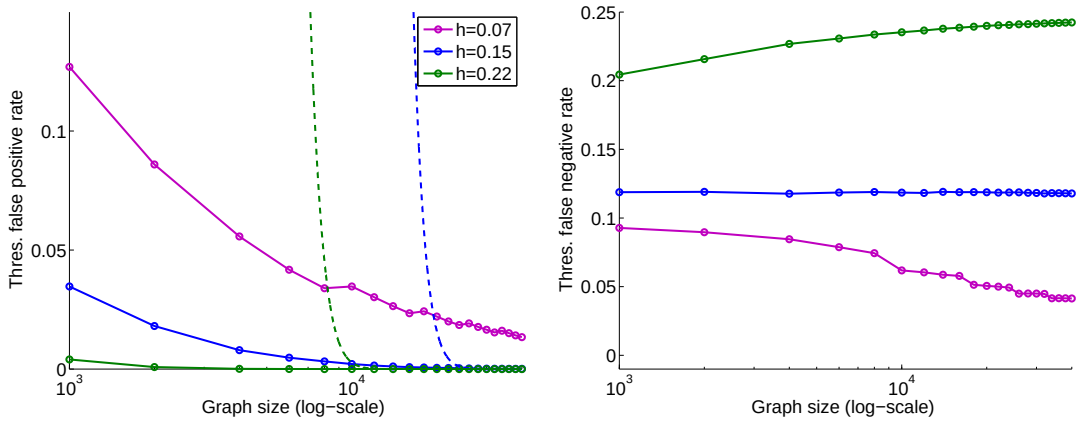
The histograms on Figure 7.11 show that a gap appears in the distribution of components sizes of both $\widehat{X}(t)$ and $X_{\mathcal{L}(t)}$ when n increases, and much slower in the distribution of $\widehat{X}(t)$ than in $X_{\mathcal{L}(t)}$. As a consequence, there are very often few large components which likely correspond to real clusters, and a lot of extra small components, which can be regarded as noise.

7.5.4 Clustering Profile

Discussion. In a data set, none of the two values t or h_n is given. Moreover, in the model they both depend on the latent space, and are not invariant up to similarity transformations. However, the product th_n^d is invariant. Then we could choose only levels of n-normalized degree, that is:

$$T'_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

Figure 7.6: False positive rates (solid lines) with the theoretical bounds (dashed lines) (left) and false negative rates (solid lines, right) of the thresholding error as functions of n



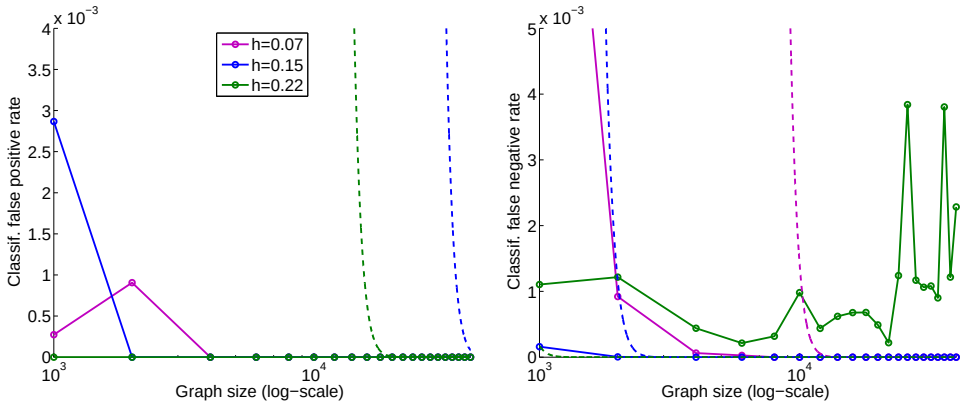
instead of the normalized degree T_i. But the question is still: how to choose a proper threshold ? In practice, we propose to explore the clustering structure of the latent space by trying all threshold values between $\min_{i\in[n]} T'_i$ and $\max_{i\in[n]} T'_i$. $(T'_{(i)})_i$ denotes the sequence of the n-normalized degrees sorted in ascending order. If $T'_{(i)} < T'_{(i+1)}$, the algorithm will give the same result for all threshold between these two values. This extra algorithm is still linear with respect to the number of edges. The plot $\widehat{Q}(u)$ for all $u \in \left[\min_{i\in[n]} T'_i, \max_{i\in[n]} T'_i\right]$ is called clustering profile of the graph X.

On Figure 7.13, some clustering profiles are given in this design for some values of h. On each plot, 30 cluster profiles are displayed.

For all $0 < t < \|f\|_\infty$, $Q(t) = 4$. In all profiles, at very low levels of normalized degrees —here normalized degrees means $\sum_{j\in[n]} X_{ij}/n$ — there are many components, because of isolated nodes. Then except for small h, there is a long range of levels with only one huge component. A phase transition occurs, which is even clearer when n is large, and the number of components suddenly grows from one to four. Under the level where the transition occurs, the cluster structure is thus not detectable. Moreover it does not seem to depend on the graph size. At very high levels, there is first a small increase, because there are few nodes remaining in the graph $\widehat{X}(t)$ which become isolated, and the number of components finally vanishes, until $\widehat{X}(t)$ is empty.

It is very interesting to see that the problem of underestimation when h is large does not appear anymore here. It was actually expected: from Figure 7.8, we could think that if we had reduced the threshold from t to $t-\varepsilon$, we would have compensated the bias, and would have found 4 clusters.

Figure 7.7: False positive (left) and false negative (right) rates of the cluster classification error (solid line) as functions of n both with the theoretical bounds (dashed lines)



Number of nodes n Number of nodes n

Design with complex-shaped clusters Unlike f_1, $Q(t)$ is not constant for all values of $t > 0$: $Q(t) = 1$ for very small t, then $Q(t) = 2$, and again $Q(t) = 1$. For small thresholds, nothing can be deduced: either there could be only one cluster, or it could be like in the first design, no clustered structure is detectable. There is a transition to two clusters, and then there is a new phenomenon here, when $Q(t)$ goes from two to one: $\widehat{Q}_n(t)$ suddenly increases, before falling to one. In fact, this sharp increase is not completely new: it can be interpreted the same way as the end of the clustering profile of f_1: it is related to the fact that the clusters are getting empty and there are few nodes remaining in the small cluster, which become isolated.

7.5.5 Conclusions and perspectives

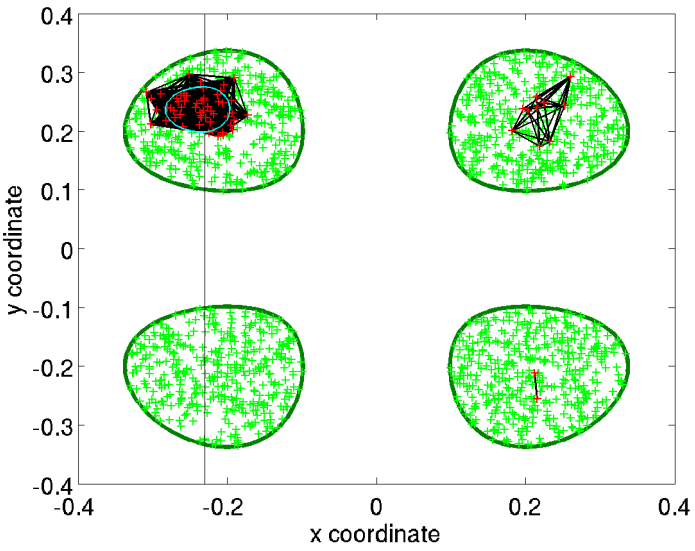
This algorithm seems to have good convergence properties, although it was not completely proved that it was consistent in this chapter. But it is not very robust: in the low graph density regime (small h), the model yields a lot of small components and the algorithm may dramatically overestimate the number of clusters. Thus the algorithm could be improved by filtering small components, for example by detecting the largest gap in the distribution of the component sizes. Theoretically, it could be first attempted to prove that overestimation asymptotically comes from isolated nodes, as in random geometric graphs (Gupta and Kumar, 1998). It could also be then attempted to construct a test to know from which size a component is significant, i.e. represents a cluster. This could then yield a

robust clustering test.

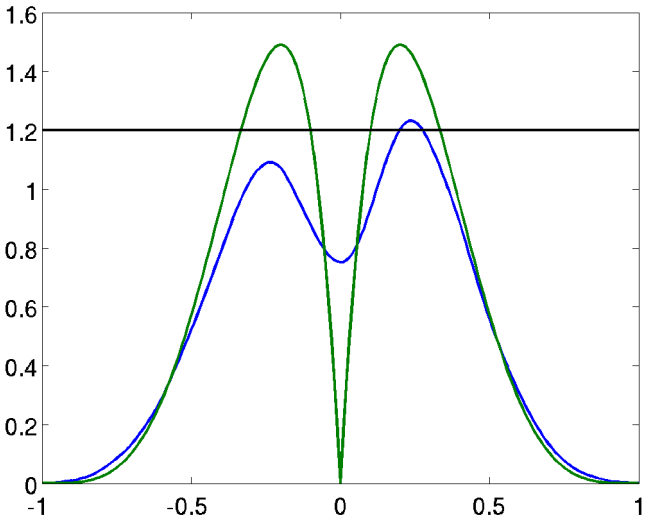
Conversely, the problem of the underestimation can be addressed by using the notion of centrality or *betweenness*, see Subsection 4.2.1. Indeed edges connecting distinct clusters should have a large betweenness because clusters are densely connected inside, and poorly connected between them. By adding an extra step to the algorithm, consisting in removing edges with large betweenness, more general kernels k could be considered as well, like the gaussian kernel. Such a kernel would allow long distance edges, and then intercluster connections, but these could be removed by the additional step. However there is probably no hope to use a heavy-tailed kernel.

We also showed the use of a practical algorithm to decipher the latent structure of a graph called clustering profile. The fact that it needs to choose neither h_n nor t is a major advantage. It can be also useful in the high graph density regime (large h), because it removes the problem of the bias of the density estimators. In this chapter we have developed theoretical arguments for a fixed t, and the bounds depend on t. These could be expanded to prove an uniform bound over $[0, \|f\|_\infty]$, removing a neighborhood of the t where $Q(t)$ changes.

Figure 7.8: Realization of $\widehat{X}(t)$ and isolines of f_1 and f_n (top); Kernel estimator f_n and density f_1 in the hyperplane $x = -0.17$ (bottom), with $n = 7000$ and $h = 0.22$

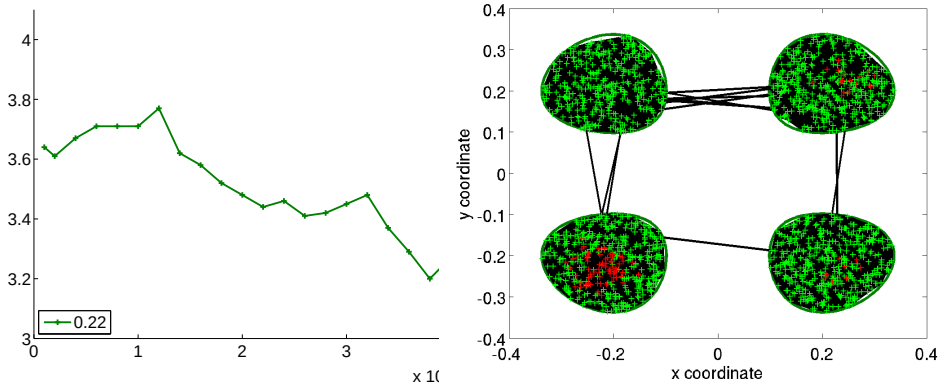


— : t-Isoline of f_n, — : t-Isoline of f_1
+ : false positive, + : nodes of $\widehat{X}(t)$



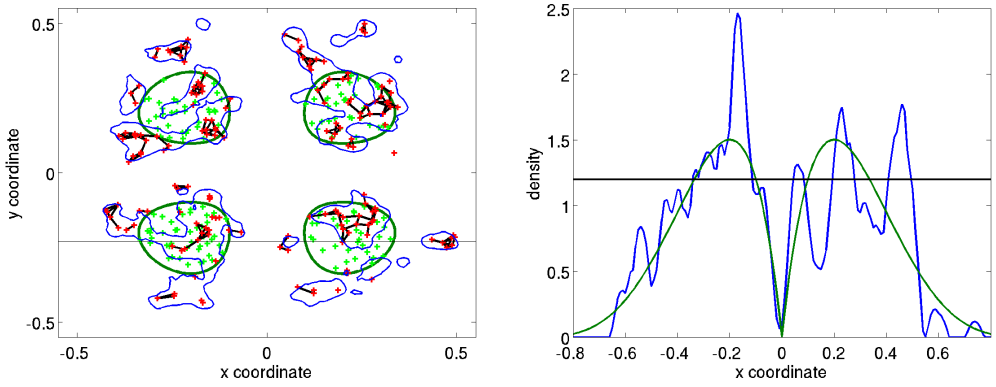
— : Kernel estimator f_n,
— : Density f_1, — : $t = 1.2$

Figure 7.9: Average number of clusters covered by $\widehat{X}(t)$ (left); Realization of $X_{\mathcal{L}(t)}$ and isolines of f_1 (right), with $n = 7000$ and $h = 0.22$



$\widehat{X}(t)$ $X_{\mathcal{L}(t)}$

Figure 7.10: Realization of $\widehat{X}(t)$ and isolines of f_1 and f_n (left); Kernel estimator and density f_1 in the hyperplane $y = -0.23$ (right), with $n = 1000$ and $h = 0.05$



— : t-Isoline of f_n, — : t-Isoline of f_1
+ : false positive, + : nodes of $\widehat{X}(t)$

— : Kernel estimator f_n,
— : Density f_1, — : $t = 1.2$

Figure 7.11: Average distributions of the sizes of the connected components of $\widehat{X}(t)$ (black) and of $X_{\mathcal{L}(t)}$ (red) with $h = 0.05$

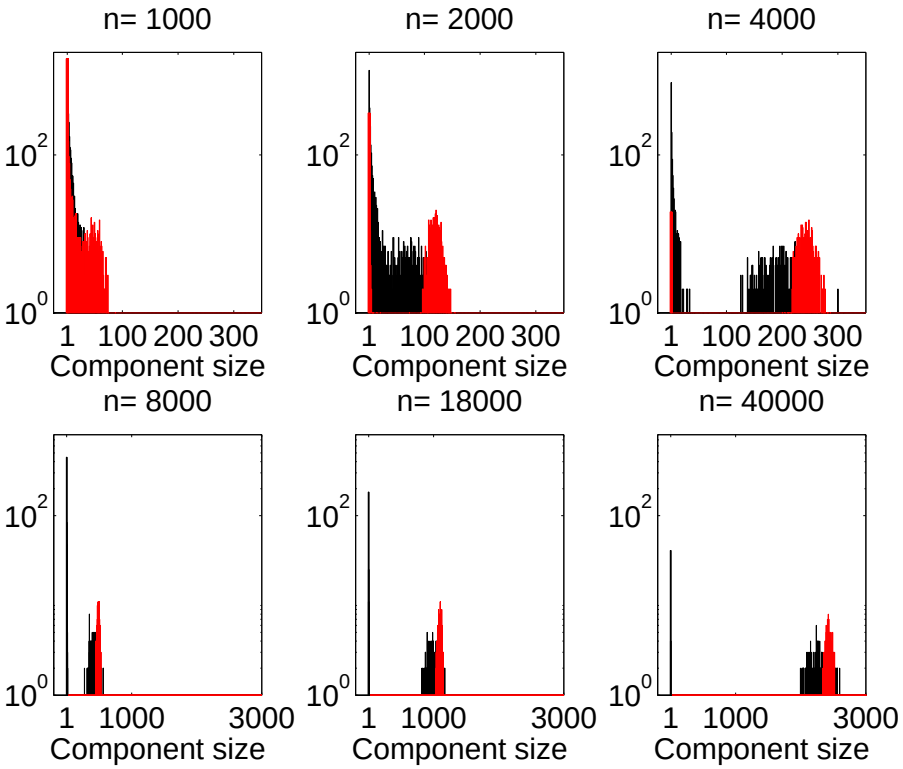


Figure 7.12: Average distributions of the sizes of the connected components of $\widehat{X}(t)$ and of $X_{\mathcal{L}(t)}$ with $h = 0.08$

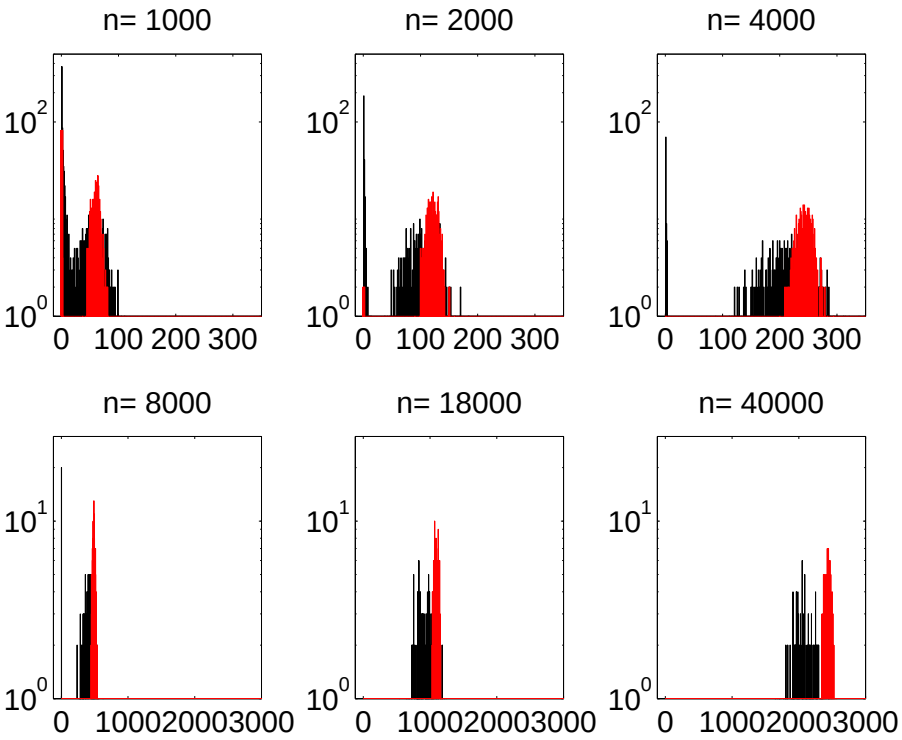


Figure 7.13: Clustering profiles of 30 graphs with $h = 0.15$ (left) and $h = 0.22$ (right)

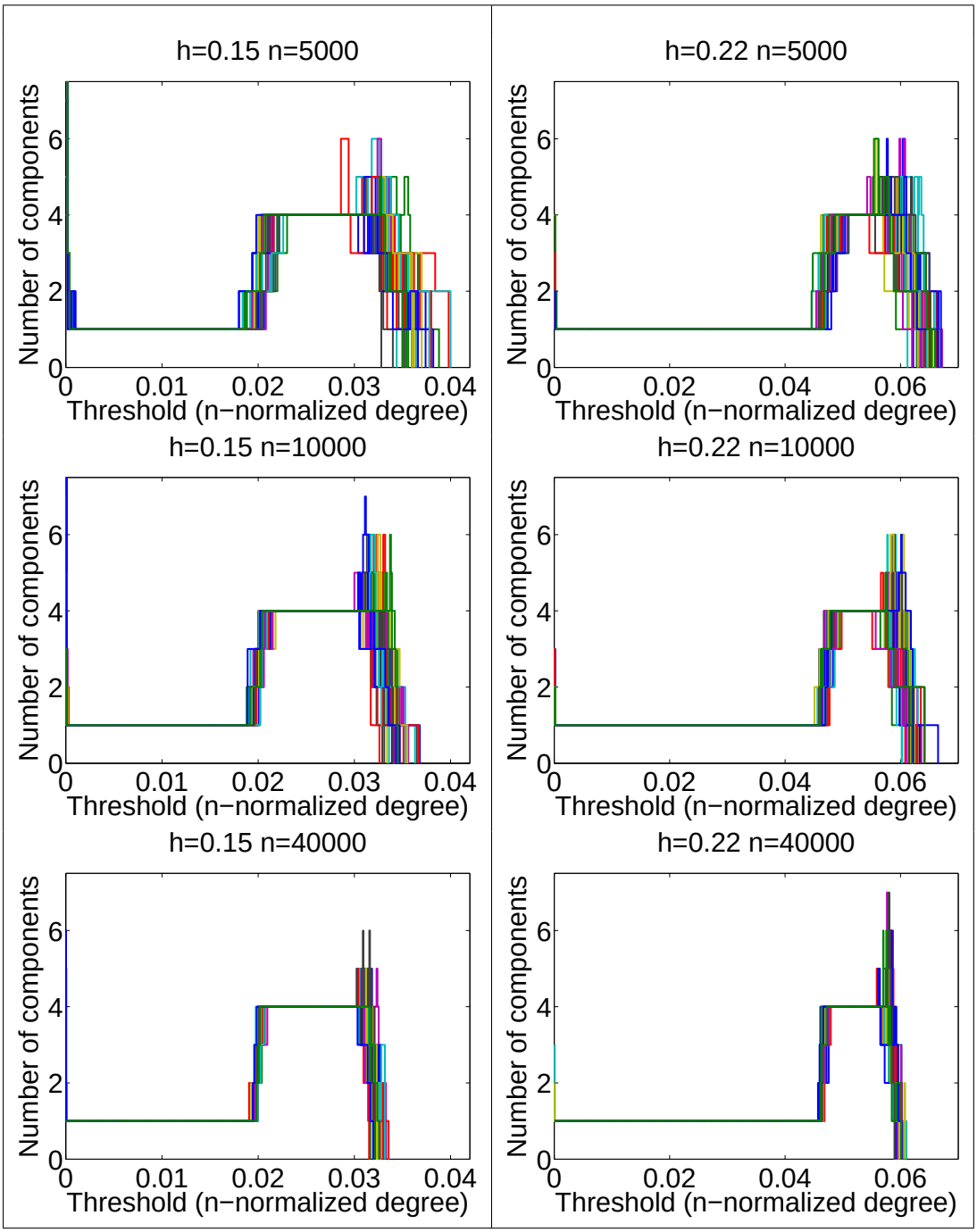
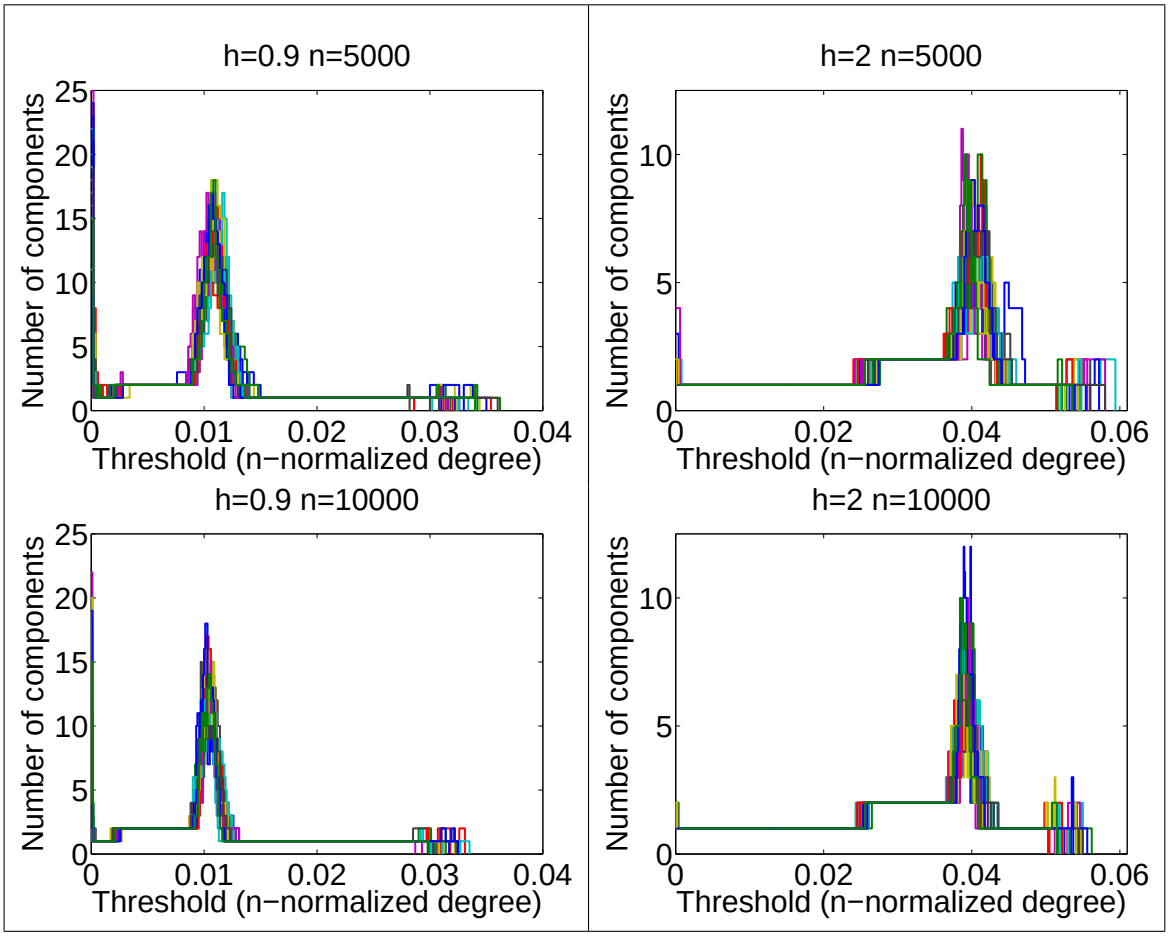


Figure 7.14: Clustering profiles of f_2 of 30 graphs with $h = 0.09$ (left), and $h = 2$ (right)



Appendix

7.A Proofs of Section 7.1

7.A.1 Proof of Proposition 7.1

$$\begin{aligned}
\zeta_n(h) &= \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \mathbb{E}\left(X_{ij}\right) &&= \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \mathbb{E}\left[\mathbb{E}^{Z_i,Z_j}(X_{ij})\right] \\
&= \frac{2}{n(n-1)} \sum_{1\leq i<j\leq n} \mathbb{E}\left[k\left(\frac{Z_i - Z_j}{h}\right)\right] &&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k\left(\frac{u-v}{h}\right) f(u)f(v)dudv \\
&= h^d \int_{\mathbb{R}^d} (k_h \star f)(u) \times f(u)du
\end{aligned}$$

Therefore:

$$\left|\zeta_n(h) - h^d \int_{\mathbb{R}^d} f^2\right| \leq h^d \int_{\mathbb{R}^d} |(k_h \star f - f)(u)| \times f(u)du \leq h^d \left\|k_h \star f - f\right\|_\infty$$

We further have $\left\|k_h \star f - f\right\|_\infty \leq s_h$, see Proposition 5.5 in Chapter 5 for a proof[3]. As f is uniformly continuous, $s_h \xrightarrow[h\to 0]{} 0$, and we thus have:

$$\left|\zeta_n(h) - h^d \int_{\mathbb{R}^d} f^2\right| = o(h^d)$$

which is exactly the proposition.

7.A.2 Compacity of $\mathcal{L}(t)$

An integrable continuous function is not necessarily vanishing at infinity, see for example the counterexample of the function made of peaks whose height is constant and base is vanishing quickly enough so that the function is integrable. On the opposite, if the function is also uniformly continuous, this cannot arise and the function vanishes at infinity. As a consequence, the t-level sets with $t > 0$ are necessarily bounded, which is shown in the following proposition.

3. This is a classical result, see e.g. Prakasa Rao (1983).

Proposition 7.14. *If $f : \mathbb{R}^d \to \mathbb{R}$ is positive, integrable and uniformly continuous, then for all $t > 0$, $\mathcal{L}(t)$ is compact.*

Proof. First $\mathcal{L}(t)$ is a closed set because f is continuous. Now we are going to show that under the assumptions of the proposition, f is vanishing at infinity [4]. From this it clearly follows that $\mathcal{L}(t)$ is bounded for $t > 0$. And a closed bounded set is compact in $\mathbb{R}^d$.

We are going to show that if f does not tend to zero at infinity and is uniformly continuously, then it is not integrable, which is equivalent to the result mentioned above. Since f is uniformly continuous, there exists $\eta > 0$ such that for all $z, z' \in \mathbb{R}^d$,

$$\|z - z'\| \leq \eta \Rightarrow |f(z) - f(z')| \leq t/2$$

Moreover as f does not tend to zero at infinity, there exists $t > 0$ and a sequence $(x_n)_{n\in\mathbb{N}}$ whose norm is tending to infinity, such that $\cap_{n\in\mathbb{N}}[x_n - \eta, x_n + \eta] = \varnothing$ and with $x_n \geq t$ for all $n \in \mathbb{N}$. As a result, for all $n \in \mathbb{N}$ and for all $z \in [x_n - \eta, x_n + \eta]$, $f(z) \geq t/2$. Since f is positive and $\cap_{n\in\mathbb{N}}[x_n - \eta, x_n + \eta] = \varnothing$, we have:

$$\int_{\mathbb{R}^d} f \geq \sum_{n\in\mathbb{N}} 2\eta \frac{t}{2} = +\infty$$

And f is not integrable. □

7.A.3 Proof of Lemma 7.1 (Geometrical lemma)

First, we prove that for $\varepsilon > 0$ small enough, the connected components of $\mathcal{L}(t + \varepsilon)$ are the intersection of the connected components of $\mathcal{L}(t)$ with $\mathcal{L}(t + \varepsilon)$. We will replace $\mathcal{L}(t)$ by $\mathcal{L}(t - \varepsilon)$ at the end of this section. The proofs of the intermediate propositions are postponed to Section 7.A.4.

Let $\varepsilon \geq 0$. Let us define for all $i \in [Q(t)]$

$$\mathcal{D}_i^\varepsilon(t) = \mathcal{C}_i(t) \cap \mathcal{L}(t + \varepsilon).$$

The following proposition describes the behavior of these sets.

Proposition 7.15. *For all $\varepsilon \geq 0$ and for all $i \in [Q(t)]$, only one of these assumptions is true:*

- *$\mathcal{D}_i^\varepsilon(t) = \varnothing$*
- *$\mathcal{D}_i^\varepsilon(t)$ is disconnected*
- *there exists $j \in [Q(t + \varepsilon)]$ such that $\mathcal{D}_i^\varepsilon(t) = \mathcal{C}_j(t + \varepsilon)$.*

4. This is a classical result we are proving here

From now on, to obtain the result, we just have to show that for ε small enough, $\mathcal{D}_i^\varepsilon(t)$ is neither empty (Proposition 7.16) nor disconnected (Proposition 7.17) for all $i \in [Q(t)]$.

Proposition 7.16. *Under Assumption 7.1, there exists $\varepsilon_1 > 0$ such that for all $\varepsilon \in [0, \varepsilon_1]$, $\mathcal{D}_i^\varepsilon(t)$ is not empty.*

Secondly, we prove that under Assumption 7.1, these sets are not disconnected for ε small enough as well.

Proposition 7.17. *Under Assumption 7.1, there exists $\varepsilon_2 > 0$ such that for all $\varepsilon \in [0, \varepsilon_2]$, $\mathcal{D}_i^\varepsilon(t)$ is connected.*

The lemma is thus proved as stated at the very beginning of this section.

To prove Lemma 7.1 as stated in Section 7.1.7, we need to show that the result is still true if we replace t by $t - \varepsilon$. In fact we are going to prove that for $\varepsilon > 0$ small enough, Assumption 7.1 is satisfied for the level $t - \varepsilon$ as well, i.e. that f is also of class $\mathcal{C}^1$ on a neighborhood of $\{f = t - \varepsilon\}$ and its gradient is non-zero on this contour line.

First, under Assumption 7.1, f is of class $\mathcal{C}^1$ on a neighborhood of $\{f = t\}$ and the gradient is non-zero on $\{f = t\}$. The continuity of the gradient ensures that it is also non-zero on a neighborhood of $\{f = t\}$. As a result, we just need to prove that for $\varepsilon > 0$ small enough, the contour line $\{f = t - \varepsilon\}$ is contained by this neighborhood. Because of the compacity of $\{f = t\}$, any neighborhood of $\{f = t\}$ contains a neighborhood of the form:

$$U_r = \{z \in \mathbb{R}^d;\ \mathrm{dist}(z, \{f = t\}) < r\},\ \text{for some } r > 0.$$

Proposition 7.18. *For all $r > 0$, there exists $\varepsilon_3 > 0$ such that for all $\varepsilon \in [0, \varepsilon_3]$,*

$$\{f = t - \varepsilon\} \subset U_r.$$

As a conclusion, the ε_0 of the lemma is provided by the minimum of all $\varepsilon_1, \varepsilon_2, \varepsilon_3$.

7.A.4 Proof of intermediate propositions

Proof of Proposition 7.15. Let us suppose that $\mathcal{D}_i^\varepsilon(t)$ is not empty and is connected. We have to prove the third statement. Since $\mathcal{D}_i^\varepsilon(t) \subset \mathcal{L}(t + \varepsilon)$ and is connected, there necessarily exists an unique $j \in [Q(t + \varepsilon)]$ such that

$$\mathcal{D}_i^\varepsilon(t) \subset \mathcal{C}_j(t + \varepsilon) \tag{7.4}$$

We now prove the opposite inclusion to conclude. Similarly, $\mathcal{C}_j(t + \varepsilon) \subset \mathcal{L}(t)$ and is connected, therefore there is an unique $i' \in [Q(t)]$ such that $\mathcal{C}_j(t+\varepsilon) \subset \mathcal{C}_{i'}(t)$.

Using 7.4, we obtain $\mathcal{D}_i^\varepsilon(t) \subset \mathcal{C}_{i'}(t)$. But according to its definition, $\mathcal{D}^\varepsilon(t)$ is also included in $\mathcal{C}_i(t)$ and is supposed to be not empty. As the connected components $(\mathcal{C}_i(t)_{i\in[Q(t)]}$ are disjoint, then i' actually equals i.

Therefore $\mathcal{C}_j(t+\varepsilon) \subset \mathcal{C}_i(t)$, and

$$\mathcal{C}_j(t+\varepsilon) \subset \mathcal{C}_i(t) \cap \mathcal{L}(t+\varepsilon) = \mathcal{D}_i^\varepsilon(t).$$

As a conclusion, $\mathcal{D}_i^\varepsilon(t) = \mathcal{C}_j(t+\varepsilon)$, which completes the proof of the alternative.

Proof of Proposition 7.16. Since t-clusters are compact and f is continuous, it has a maximum on each of these clusters. The maximum of f on $\mathcal{C}_i(t)$ is denoted by m_i. Note that for all i, $m_i > t$; otherwise, f would equal t over the whole cluster and t would be a local maximum of f in one of the clusters and the gradient would be zero at some point of the set $\{f = t\}$: this is not possible under Assumption 7.1. Let us define $\varepsilon_1 = \min_{i\in[Q(t)]} m_i - t$. As f is continuous,it takes all values between t and $t+\varepsilon_1$ on each of the $\mathcal{C}_i(t)$. As a result, for all $\varepsilon \in]0, \varepsilon_1]$, and all $i \in [Q(t)]$, $\mathcal{D}_i^\varepsilon(t)$ is not empty.

Proof of Proposition 7.17. We prove the contraposive statement again, that is: we assume that there exists a vanishing sequence of non-negative numbers $(epsi_n)$, such that for all n, $\mathcal{D}_i^{\varepsilon_n}(t)$ is disconnected and we prove that it implies that there exists some point of $\{f = t\}$ where the gradient of f is zero.

For all n, as the connected components of $\mathcal{D}_i^{\varepsilon_n}(t)$ are compact, there exist two sequences (u_n), (v_n) such that:

$$\|u_n - v_n\| = \min\{\|z - z'\|;\ z, z' \text{ are in distinct connected components of } \mathcal{D}_i^{\varepsilon_n}(t)\}.$$

Since $u_n, v_n \in \mathcal{C}_i(t)$, which is compact, we can assume that up to extractions, (u_n) and (v_n) both converge towards respectively u and v, both in $\mathcal{C}_i(t)$. Note that $\|u_n - v_n\| \xrightarrow[n\to\infty]{} 0$ because

$$\bigcup_{n\in\mathbb{N}} \mathcal{D}_i^{\varepsilon_n}(t) = (\mathcal{C}_i(t))^\circ$$

where the $\circ$ denotes the interior of the set. As a consequence, $u = v$. Moreover, u_n and v_n are in the boundary of $\mathcal{D}_i^\varepsilon(t)$, and this boundary is included in $\{f = t+\varepsilon_n\}$, using the continuity of f. Using again the continuity of f, $f(u_n) = t+\varepsilon_n \xrightarrow[n\to\infty]{} f(u)$ and hence $f(u) = t$, i.e. $u \in \{f = t\}$.

As f is of class $\mathcal{C}^1$ on the neighborhood of $\{f = t\}$, f is continuously differentiable at both u_n and v_n for n large enough. Furthermore, the gradient of f is non-zero on $\{f = t\}$, hence by continuity of the gradient on the neighborhood, the

gradient is also non-zero at u_n and v_n for n large enough. Using the implicit function theorem, there exist local parametrizations of $\{f = t + \varepsilon_n\}$ at these points, respectively

$$\phi_n : V \to \{f = t + \varepsilon_n\} \text{ and } \psi_n : W \to \{f = t + \varepsilon_n\}.$$

where V and W are open sets of $\mathbb{R}^{d-1}$ containing 0, with $\phi_n(0) = u_n$ and $\psi_n(0) = v_n$. $\|u_n - v_n\|^2$ is the minimum of the function

$$(r, s) \in \mathbb{R}^{d-1} \times \mathbb{R}^{d-1} \mapsto \|\phi_n(r) - \psi_n(s)\|^2$$

on $V \times W$, therefore its gradient is zero at 0. The gradient of this function at the point (r, s) is:

$$\left(2\left\langle \frac{\partial \phi_n}{\partial r_1}(r), \phi_n(r) - \psi_n(s)\right\rangle, \ldots, 2\left\langle \frac{\partial \phi_n}{\partial r_{d-1}}(r), \phi_n(r) - \psi_n(s)\right\rangle,\right.$$
$$\left. -2\left\langle \frac{\partial \psi_n}{\partial s_1}(s), \phi_n(r) - \psi_n(s)\right\rangle, \ldots, -2\left\langle \frac{\partial \psi_n}{\partial s_{d-1}}(s), \phi_n(r) - \psi_n(s)\right\rangle\right).$$

As a result, at the point 0, we obtain that all derivatives of ϕ_n and ψ_n are orthogonal to $\phi_n(0) - \psi_n(0) = u_n - v_n$. But the $d-1$ partial derivatives of ϕ_n (respectively of ψ_n) form a basis of the tangent hyperplane to $\{f = t + \varepsilon_n\}$ at u_n (respectively v_n). Therefore $u_n - v_n$ is orthogonal to both tangent hyperplanes at u_n and v_n.

Since the gradient of f is also orthogonal to the tangent hyperplane, $\overrightarrow{grad}\, f(u_n)$, $\overrightarrow{grad}\, f(v_n)$ and $u_n - v_n$ are actually collinear. Since the gradient points towards the direction where f is growing, $\overrightarrow{grad}\, f(u_n)$ and $\overrightarrow{grad}\, f(v_n)$ both point towards the opposite direction of the segment $[u_n, v_n]$ (on this segment, f is less than $t+\varepsilon_n$ whereas it is more in the opposite direction). Hence the vectors $\overrightarrow{grad}\, f(u_n)$ and $\overrightarrow{grad}\, f(v_n)$ are collinear and in opposite directions, and

$$\left\langle \overrightarrow{grad}\, f(u_n), \overrightarrow{grad}\, f(v_n)\right\rangle \leq 0. \tag{7.5}$$

But by continuity of $\overrightarrow{grad}\, f$ and of the scalar product,

$$\left\langle \overrightarrow{grad}\, f(u_n), \overrightarrow{grad}\, f(v_n)\right\rangle \xrightarrow[n\to\infty]{} \left\langle \overrightarrow{grad}\, f(u), \overrightarrow{grad}\, f(u)\right\rangle = \|\overrightarrow{grad}\, f(u)\|^2 \tag{7.6}$$

Combining 7.5 and 7.6, we obtain that $\|\overrightarrow{grad}\, f(u)\|^2$ is negative and therefore necessarily equals zero, which actually means that $\overrightarrow{grad}\, f(u) = 0$.

Proof of Proposition 7.18. We actually prove this by contradiction. We assume that there exists $r > 0$, a vanishing sequence of non-negative numbers (ε_n), and a sequence (z_n) of points of $\mathbb{R}^d$ such that for all n,

$$z_n \in \{f = t - \varepsilon_n\} \text{ and } \operatorname{dist}(z_n, \{f = t\}) > r.$$

Let $m = \max_{n\in\mathbb{N}} \varepsilon_n$. We have $m < t$, then $\mathcal{L}(t-m)$ is compact. The sequence (z_n) is in this compact, because $\{f = t - \varepsilon_n\} \subset \mathcal{L}(t-m)$. Then it converges towards a point z of $\mathcal{L}(t-m)$ up to an extraction. Since both f and the function $z \mapsto \operatorname{dist}(z, \{f = t\})$ are continuous, we have:

$$f(z_n) \xrightarrow[n\to\infty]{} f(z) \text{ and } \operatorname{dist}(z_n, \{f = t\}) \xrightarrow[n\to\infty]{} \operatorname{dist}(z, \{f = t\})$$

On the one hand $z \in \{f = t\}$ because $f(z_n) = t - \varepsilon_n \xrightarrow[n\to\infty]{} t$ and therefore $f(z) = t$, and on the other hand, $\operatorname{dist}(z, \{f = t\}) \geq r$, because for all n, $\operatorname{dist}(z, \{f = t\}) > r$. This yields the contradiction.

7.B Proofs of Section 7.2

7.B.1 Proof of Proposition 7.2

On the event $\{U_n \leq \varepsilon\}$, if $i \in \widehat{J}_n(t)$, i.e. $T_i \geq t$, we have:

$$\begin{aligned} f(Z_i) &= T_i + f(Z_i) - T_i \\ &\geq t - \sup_{i\in[n]} |T_i - f(Z_i)| = t - U_n \\ &\geq t - \varepsilon \end{aligned}$$

In fact, on the event $\{U_n \leq \varepsilon\}$, if the normalized degree T_i is higher than t, then $Z_i \in \mathcal{L}(t-\varepsilon)$, or equivalently, $i \in \widehat{J}_n(t) \Rightarrow i \in J_n(t-\varepsilon)$. This proves the inclusion of the proposition.

7.B.2 Proof of Proposition 7.3

According to Assumption 7.3.(a), two nodes cannot be connected if their distance is more than h_n. Moreover, according to Assumption 7.3.(b), we assumed that for h_n small enough, $h_n < \delta_{t-\varepsilon}$. Therefore these imply:

$$X_{ij} = 1 \Rightarrow \|Z_i - Z_j\| \leq h_n < \delta_{t-\varepsilon}.$$

On the event $\{\widehat{J}_n(t) \subset J_n(t-\varepsilon)\}$, if $i \in \widehat{J}_n(t)$ and $j \in \widehat{J}_n(t)$ are connected, then i and j are in $J_n(t-\varepsilon)$ and $\|Z_i - Z_j\| < \delta_{t-\varepsilon}$. Z_i and Z_j cannot be in distinct

$(t-\varepsilon)$-clusters, by definition of $\delta_{t-\varepsilon}$. As a result, on the event $\{\widehat{J}_n(t) \subset J_n(t-\varepsilon)\}$, for all $i, j \in [n]$:

$$X_{ij} = 1 \Rightarrow Z_i \sim_{t-\varepsilon} Z_j. \tag{7.7}$$

It also follows from this that if i and j are in the same component of $\widehat{X}_n(t)$, i.e. there is a path in $\widehat{X}_n(t)$ between i and j, then all nodes of the path, including i and j, are in the same geometric $(t-\varepsilon)$-cluster: we are thus just using the assertion (7.7) above to each edge of the path and the transitivity of $\sim_{t-\varepsilon}$. Therefore, on $\{U_n \leq \varepsilon\}$:

$$i \,\widehat{\sim}_t j \Rightarrow Z_i \sim_{t-\varepsilon} Z_j. \tag{7.8}$$

It can be now deduced from this that on $\{U_n \leq \varepsilon\}$, all nodes of any connected component of $\widehat{X}_n(t)$ are located in the same $(t-\varepsilon)$-cluster.

7.B.3 Proof of Proposition 7.4

We first show the following inclusion:

$$\{U_n \leq \varepsilon\} \subset \left\{J_n(t+\varepsilon) \subset \widehat{J}_n(t)\right\} \tag{7.9}$$

On the event $\{U_n \leq \varepsilon\}$, if moreover $i \in J_n(t+\varepsilon)$, i.e. $f(Z_i) \geq t+\varepsilon$, then:

$$\begin{aligned} T_i &= f(Z_i) + T_i - f(Z_i) \\ &\geq t + \varepsilon - \sup_{i \in [n]} (f(Z_i) - T_i) \geq t + \varepsilon - U_n \geq t \end{aligned}$$

As a summary, on the event $\{U_n \leq \varepsilon\}$, $f(Z_i) \geq t+\varepsilon \Rightarrow T_i \geq t$, which amounts to the inclusion claimed above.

Moreover for all $(t+\varepsilon)$-cluster $\mathcal{C}_i(t+\varepsilon)$, $i \in [Q(t+\varepsilon)]$, $V_{\mathcal{C}_i(t+\varepsilon)} \subset J_n(t+\varepsilon)$. Therefore on the previous event, we also have $V_{\mathcal{C}_i(t+\varepsilon)} \subset \widehat{J}_n(t)$, and $V_{\mathcal{C}_i(t+\varepsilon)} \neq \varnothing \Rightarrow V_{\mathcal{C}_i(t+\varepsilon)} \cap \widehat{J}_n(t) \neq \varnothing$. This proves the proposition.

7.B.4 Proof of Proposition 7.5

According to Lemma 7.1, each $(t-\varepsilon_0)$-cluster has non-empty intersection with the $(t+\varepsilon_0)$-level set, and $Q(t-\varepsilon_0) = Q(t) = Q(t+\varepsilon_0)$. Therefore each $(t-\varepsilon_0)$-cluster contains exactly one $(t+\varepsilon_0)$-cluster. On the event $B_n^{\varepsilon_0}(t)$, there is at least one node i_q of the graph $\widehat{X}(t)$ in each $(t+\varepsilon_0)$-cluster $\mathcal{C}_Q(t+\varepsilon_0)$, with $k \in [Q(t)]$ (recall that $Q(t+\varepsilon_0) = Q(t)$). Moreover on the event $E_n^{\varepsilon_0}(t)$, the connected component of the node i_q is completely included in the $(t-\varepsilon_0)$-cluster containing the $(t+\varepsilon_0)$-cluster $\mathcal{C}_Q(t+\varepsilon_0)$. As a conclusion, there is at least one connected component of $\widehat{X}(t)$ in each $(t-\varepsilon_0)$-cluster, and we finally have $\widehat{Q}_n(t) \geq Q(t-\varepsilon_0) = Q(t)$.

7.B.5 Proof of Proposition 7.6.(1)

Using an union bound, we get:

$$P(\overline{A_n(t)}) = P\left(\bigcup_{i=1}^{Q(t)}\{V_{\mathcal{C}_i(t)} = \varnothing\}\right) \leq \sum_{q=1}^{Q(t)} P\left(\bigcap_{i\in[n]}\{V_{\mathcal{C}_i(t)} = \varnothing\}\right).$$

$\{V_{\mathcal{C}_i(t)} = \varnothing\}$ means that none of the nodes is in $\mathcal{C}_i(t)$. Recall also that $(Z_i)_{i\in[n]}$ are mutually independent. Hence:

$$\begin{aligned} P(\overline{A_n(t)}) &\leq \sum_{q=1}^{Q(t)} P\left(\bigcap_{i\in[n]}\{Z_i \notin \mathcal{C}_i(t)\}\right) \\ &\leq \sum_{k=1}^{Q(t)}(1-\alpha_Q(t))^n \leq Q(t)(1-\alpha_0(t))^n. \end{aligned}$$

7.B.6 Proof of Proposition 7.6.(2)

The proof relies on the following lemma, which bounds the probability that the estimation error by the underlying kernel estimator f_n exceeds some fixed ε. This kind of result can also be found in Prakasa Rao (1983), a proof adapted to our setting is given in Chapter 5, see Proposition 5.1, we are giving again now:

Lemma 7.4. *Let $\varepsilon > 0$. Under Assumption 7.2.*(a)*, for all $z \in \mathbb{R}^d$:*

$$P\left(|f_n(z) - f(z)| > \varepsilon\right) \leq 2\exp\left(-\frac{1}{8}\frac{\varepsilon^2 n h_n^d}{(\|f\|_\infty + \frac{2\varepsilon}{3})}\right)$$

The proof of Proposition 7.6.(2) follows. We have

$$\begin{aligned} P(U_n > \varepsilon_0) &\leq P\left(\bigcup_{i=1}^{n}\{|T_i - f(Z_i)| > \varepsilon_0\}\right) \\ &\leq \sum_{i=1}^{n} P(|T_i - f(Z_i)| > \varepsilon_0) \\ &\leq \sum_{i=1}^{n} P(|T_i - f_n(Z_i)| > \varepsilon_0/2) + P(f_n(Z_i) - f(Z_i) > \varepsilon_0/2) \\ &\leq \sum_{i=1}^{n} \mathbb{E}_Z\left[P^Z(|T_i - f_n(Z_i)| > \varepsilon_0/2)\right] + \mathbb{E}_{Z_i}\left[P^{Z_i}(|f_n(Z_i) - f(Z_i)| > \varepsilon_0/2)\right] \end{aligned}$$

Using directly Bernstein's inequality for the first term:

$$\begin{aligned}
P^Z(|T_i - f_n(Z_i)| > \varepsilon_0/2) &= P^Z\left(\left|D_i - \sum_{j=1}^n k\left(\frac{Z_i - Z_j}{h_n}\right)\right| \geq \frac{\varepsilon_0 n h_n^d}{2}\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{(\varepsilon_0 n h_n^d)^2}{\sum_{j=1}^n k\left(\frac{Z_i-Z_j}{h_n}\right)\left(1 - k\left(\frac{Z_i-Z_j}{h_n}\right)\right) + \frac{\varepsilon_0 n h_n^d}{3}}\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{\frac{1}{nh_n^d}\sum_{j=1}^n k\left(\frac{Z_i-Z_j}{h_n}\right) + \frac{\varepsilon_0}{3}}\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{f_n(Z_i) + \frac{\varepsilon_0}{3}}\right)
\end{aligned}$$

Then, taking the expectation, and splitting it between the following events:

$$\{f_n(Z_i) \leq f(Z_i) + \varepsilon_0/2\} \text{ and } \{f_n(Z_i) > f(Z_i) + \varepsilon_0/2\}.$$

$$\begin{aligned}
\mathbb{E}_Z\left[P^Z(|T_i - f_n(Z_i)| > \varepsilon_0/2)\right] &\leq \mathbb{E}_Z\left(2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{f_n(Z_i) + \frac{\varepsilon_0}{3}}\right)\mathbb{1}_{\{f_n(Z_i) \leq f(Z_i) + \varepsilon_0/2\}}\right) \\
&\quad + \mathbb{E}_Z\left(\mathbb{1}_{\{f_n(Z_i > f(Z_i) + \varepsilon_0/2\}}\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right) P(f_n(Z_i) \leq f(Z_i) + \varepsilon_0/2) \\
&\quad + P\left(f_n(Z_i) - f(Z_i) > \varepsilon_0/2\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right) + \mathbb{E}_{Z_i}\left[P^{Z_i}(f_n(Z_i) - f(Z_i) > \varepsilon_0/2)\right]
\end{aligned}$$

Thus we obtain this intermediate result:

$$P(U_n > \varepsilon_0) \leq \sum_{i=1}^n \left[2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right) + 2\mathbb{E}_{Z_i}\left[P^{Z_i}(|f_n(Z_i) - f(Z_i)| > \varepsilon_0/2)\right]\right].$$

Now we deal with $P^{Z_i}\left(|f_n(Z_i) - f(Z_i)| > \varepsilon_0/2\right)$, appearing twice. It follows from Lemma 5.1 with $\varepsilon = \varepsilon_0/2$:

$$P^{Z_i}\left(|f_n(Z_i) - f(Z_i)| > \varepsilon_0/2\right) \leq 2\exp\left(-\frac{1}{32}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \frac{\varepsilon_0}{3}}\right)$$

It can now be concluded that:

$$\begin{aligned}
P(U_n > \varepsilon_0) &\leq \sum_{i=1}^n \left[2\exp\left(-\frac{1}{2}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right) + 4\exp\left(-\frac{1}{32}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \frac{\varepsilon_0}{3}}\right)\right] \\
&\leq 6n\exp\left(-\frac{1}{32}\frac{\varepsilon_0^2 n h_n^d}{\|f\|_\infty + \varepsilon_0}\right)
\end{aligned}$$

7.C Proofs of Section 7.3

7.C.1 Proof of Lemma 7.2

First, we define for C, C' two parts of $\mathbb{R}^d$:

$$C - C' = \{z - z';\ z \in C,\ z' \in C'\}$$

Note that $h_n^{-1}(C - C) \subset B(0, \gamma)$, as for all $z, z' \in C$:

$$\left\| \frac{z - z'}{h_n} \right\| \leq \frac{\text{diam}(C)}{h_n} \leq \gamma.$$

Therefore for all nodes $i, j \in V_C$, $Z_i, Z_j \in C$, hence:

$$\frac{Z_i - Z_j}{h} \in h_n^{-1}(C - C) \subset B(0, \gamma)$$

Define for all $i, j \in V_C$:

$$p'_{ij} = P_{n_C,p}(X_{ij} = 1) = p \text{ and } p_{ij} = P^Z(X_{ij} = 1) = k\left(\frac{Z_i - Z_j}{h_n}\right).$$

As a consequence:

$$0 < p'_{ij} = p = \min_{z \in h^{-1}(C-C)} k(z) \leq k\left(\frac{Z_i - Z_j}{h}\right) = p_{ij}.$$

Applying Lemma 7.5, we obtain the inequality of the proposition.

We then prove Lemma 7.5 used in the previous proof and related to the monotony of connectedness:

Lemma 7.5. *Let* $\mathbf{p} = (p_{ij})_{i,j\in[n]}, \mathbf{p}' = (p'_{ij})_{i,j\in[n]} \in [0,1]^{n\times n}$. *If for all* $i, j \in [n]$, $p'_{ij} \leq p_{ij}$, *then:*

$$P_{\mathbf{p}'}(\textit{connectedness}) \leq P_{\mathbf{p}}(\textit{ connectedness})$$

Proof. Let $0 < p \leq p' \leq 1$ and $X' \sim \mathcal{B}(p')$. We now show that there exists a random variable $X \sim \mathcal{B}(p)$ such that $X' \geq X$. X is defined the following way:

$$X = 0 \text{ on } \{X' = 0\} \text{ et } X|\ X' = 1 \quad \sim \mathcal{B}\left(\frac{p}{p'}\right).$$

By construction, X is a Bernoulli-distributed random variable and its parameter is p:

$$P(X = 1) = P(X = 1|\ X' = 0)P(X' = 0) + P(X = 1|\ X' = 1)P(X' = 1) = p.$$

Let $X' \sim \mathcal{ER}(n, \mathbf{p}')$. For all $i, j \in [n]$, $X'_{ij} \sim \mathcal{B}(p'_{ij})$. Let $X_{ij} \sim \mathcal{B}(p_{ij})$ such that $X'_{ij} \geq X_{ij}$, as in the previous paragraph. As variables X'_{ij} are mutually independent, and the distribution of X_{ij} is defined conditionally on X'_{ij}, variables $(X_{ij})_{i,j\in[n]}$ can also be chosen as being mutually independent. Therefore the distribution of the random graph X is $\mathcal{ER}(n, \mathbf{p})$. Moreover $X \subset X'$, because $X'_{ij} \geq X_{ij}$ for all $i, j \in [n]$. Thus, if X is connected, then so is X'. As a conclusion:

$$P(X' \text{ connected}) \leq P(X \text{ connected}).$$

□

7.C.2 Tilings and covers: first attempt with cubist style

The first attempt described here to construct a proper tiling of $\mathcal{L}(t)$ which could be used in Proposition 7.8, consists in taking cubic tiles. This attempt will turn out to be too rough, but is a preliminary step to the final construction.

Definition 7.5. *If $\rho > 0$, $\mathcal{T}$ is said to be ρ-cubic if all elements are closed hypercubes of side ρ.*

Notation. By abusing the notation, if $\mathcal{T}$ is a cover of any subset of $\mathbb{R}^d$, we will also denote $\mathcal{T}$ the set covered by $\mathcal{T}$, that is the union of the tiles of the cover $\mathcal{T}$, if confusion is not possible. For example, $X_{\mathcal{T}}$ will denote the subgraph induced by the set covered by $\mathcal{T}$.

Note that we will consider only hypercubes the sides of which are parallel to the axes associated to the canonical basis of $\mathbb{R}^d$.

One property which is needed from cubic tilings such that each tile shares at least one face with another tile will be needed. It is first proved that for ρ small enough, such a tiling exists.

Lemma 7.6. *For all subset of $\mathbb{R}^d$, there exists $\rho > 0$ and a ρ-cubic tiling $\mathcal{T}$ of this subset such that each tile of $\mathcal{T}$ shares at least one face with another tile of $\mathcal{T}$.*

The proof of this lemma is omitted in this manuscript.

Let $\mathcal{T}$ be a ρ-cubic tiling of a compact subset A of $\mathbb{R}^d$. From $\mathcal{T}$, we construct:

1. the $\rho/2$-cubic tiling of A made by splitting each ρ-cubic tile of $\mathcal{T}$ in 2^d distinct hypercubes of side $\rho/2$, denoted by $\mathcal{T}^{\boxplus}$. Elements of this set are called $\mathcal{T}$-*subtiles*, or just subtiles if the context is clear.
2. the cover of A defined as the set of all tiles which can be made by assembling 2^d subtiles sharing the same corner, and denoted by $\mathcal{T}^{\square}$.

Note that $\mathcal{T}^{\square}$ is made of the tiles C included in the set covered by $\mathcal{T}$ and such that $C = C' + \sum_{i=1}^{d} \sigma_i(\rho/2)e_i$ with $C' \in \mathcal{T}$, $(e_1, \ldots, e_d)$ the canonical basis of $\mathbb{R}^d$ and $(\sigma_1, \ldots, \sigma_d) \in \{-1, 0, 1\}^d$. Note also that:

$$(\mathcal{T}^{\square})^{\boxplus} = \mathcal{T}^{\boxplus}.$$

Furthermore $\mathcal{T}^{\square}$ is a cover of A and its most useful property is that for all adjacent tiles C, C', it also contains another tile O overlapping partially both C and C'. More precisely, $\mathcal{T}^{\square}$ is a $\mathcal{T}^{\boxplus}$-autocover according to the following definition:

Definition 7.6. *Let $\mathcal{T}$,$\mathcal{T}'$ be covers of a compact set. $\mathcal{T}$ is said to be a $\mathcal{T}'$-*autocover *if for all adjacent tiles $C, C' \in \mathcal{T}$, there exists $O \in \mathcal{T}$ and $c, c' \in \mathcal{T}'$ such that:*

$$c \subset C \cap O \text{ and } c' \subset C' \cap O$$

Proposition 7.19. *$\mathcal{T}^{\square}$ is a $\mathcal{T}^{\boxplus}$-autocover.*

Note that we proceed this way, that is, by first defining a ρ-cubic tiling, then taking the associated subtiling, and finally constructing the auto-cover, instead of directly constructing this autocover from a given $\rho/2$-cubic tiling (i.e. without the first step), in order to prevent from removing too many subtiles in the final autocover: it could be not a cover, because some parts of the covered set could be thereby skipped.

7.C.3 Connected components of the subgraph induced by a cubic tiling

The goal of this subsection is to show that the subgraph $X_{\mathcal{T}}$ induced by some cubic tiling $\mathcal{T}$ of $\mathcal{L}(t)$ has at most $Q(t)$ connected components on some event of high probability for n large enough.

Let $\gamma \in [0, 1[$. Let $\mathcal{T}_n$ a ρ_n-cubic tiling of $\mathcal{L}(t)$, with:

$$\rho_n = \frac{2\gamma h_n}{3\sqrt{d}}$$

We give a first version of Proposition 7.8 but for $X_{\mathcal{T}_n}$ instead of $X_{\mathcal{L}(t)}$.

Proposition 7.20. *On the intersection of the following events, there is at most one component of $X_{\mathcal{T}_n}$ in each connected component of $\mathcal{T}_n$, and the number of connected components of $X_{\mathcal{T}_n}$ is larger than $Q(t)$:*

– there is at least one of the $(Z_i)_{i\in[n]}$ in each subtile:

$$\forall c \in \mathcal{T}_n^{\boxplus},\ c \cap \{Z_i, i \in [n]\} \neq \varnothing \text{ (or equivalently, } V_c \neq \varnothing)$$

– *X_C is connected for every tile $C \in \mathcal{T}_n^{\square}$*

Proof. We first assume that the number of connected components of $\mathcal{T}_n$ is one, and we show that $X_{\mathcal{T}_n}$ is connected. Then the proof can be applied to each connected component of $\mathcal{T}_n$ to derive the general case.

It is also assumed each tile of $\mathcal{T}_n$ has at least one common face to another tile. Thus there is no tile connected to the rest only by a corner. This is possible because f is continuously differentiable on a neighborhood of $\{f = t\}$.

Let $i, j \in V_{\mathcal{T}_n}$, and $C, C' \in \mathcal{T}_n$ such that $Z_i \in C$ and $Z_j \in C'$. If $C = C'$, then there exists a path in $X_{\mathcal{T}_n}$ between i and j because X_C is connected and $X_C \subset X_{\mathcal{T}_n}$. If $C \neq C'$, since $\mathcal{T}_n$ is a connected domain and each tile is adjacent to at least one other tile, there exists $m-1$ tiles of $\mathcal{T}_n$: $C_1, \ldots, C_{m-1}$ such that for all $k \in [m]$, C_{k-1} and C_k are adjacent (with $C_0 = C$ and $C_m = C'$). Then we prove the existence of a path in $X_{\mathcal{T}_n}$ between i and j by induction with respect to m:

If $m = 0$ **:** C and C' are adjacent. As $\mathcal{T}_n^{\square}$ is a $\mathcal{T}_n^{\boxplus}$-autocover, there exists $O \in \mathcal{T}_n^{\square}$ and $c, c' \in \mathcal{T}_n^{\boxplus}$ such that $c \subset C \cap O$ and $c' \subset C' \cap O$. Moreover:

$$c \cap \{Z_i; i \in [n]\} \neq \varnothing \text{ and } c' \cap \{Z_i; i \in [n]\} \neq \varnothing$$

Therefore we can find nodes $k \in V_c$ and $l \in V_{c'}$. Since $X_C, X_O, X_{C'}$ are connected and $X_C, X_O, X_{C'} \subset X_{\mathcal{T}_n}$, there are paths in $X_{\mathcal{T}_n}$ between i and k, k and l, l and j. By concatenating these three paths, we have got a path in $X_{\mathcal{T}_n}$ between i and j.

If $m \geq 1$ **:** Let $k \in V_{C_{m-1}}$. According to the induction hypothesis, there is a path in $X_{\mathcal{T}_n}$ between i and k. Then C_{m-1} and $C_m = C'$ are adjacent, therefore there exists a path in $X_{\mathcal{T}_n}$ between k and j according to the previous case. The concatenation of these two paths gives a path in $X_{\mathcal{T}_n}$ between i and j.

□

7.C.4 Problem of the excess volume covered by a cubic tiling

Now we would like to conclude the same but with $X_{\mathcal{L}(t)}$, i.e. derive Proposition 7.8 from Proposition 7.20. As $\rho_n \to 0$, the sequence of ρ_n-cubic tilings $(\mathcal{T}_n)$ covers $\mathcal{L}(t)$ better and better when n tends to infinity: the excess volume between the set effectively covered by the tiles — also denoted $\mathcal{T}_n$ by abusing the notation — and $\mathcal{L}(t)$, tends to zero:

$$\mathrm{Vol}\left(\mathcal{T}_n \backslash \mathcal{L}(t)\right) = \mathcal{O}\left(\rho_n^d\right) \xrightarrow[n\to\infty]{} 0$$

Thus we could think that $X_{\mathcal{L}(t)}$ and $X_{\mathcal{T}_n}$ are very similar. But it is actually wrong: though the excess volume is very small, it is too big to be ignored because

the average number of nodes falling in this volume does not generally tend to zero. If (unluckily) this volume is also lower bounded by $K\rho_n^d$, where $K > 0$, then:

$$E(n_{\mathcal{T}_n \setminus \mathcal{L}(t)}) = n \int_{\mathcal{T}_n \setminus \mathcal{L}(t)} f(z)dz \geq t^- K n \rho_n^d \xrightarrow[n\to\infty]{} +\infty$$

where $t^- = \min_{z \in \mathcal{T}_n} f(z)$, which is non-negative for n large enough, by using the continuity of f in the neighborhood of $\{f = t\}$. Thus $X_{\mathcal{L}(t)}$ and $X_{\mathcal{T}_n}$ could differ of a number of nodes tending to infinity.

7.C.5 Refinement of the tilings and covers: proofs of Section 7.3.2

Now we refine the simple cubic tiling by removing the extra volume, and playing at puzzles.

Definition 7.7. *A tiling of a compact set A is said to be perfect if the set covered by the tiling exactly equals A.*

In order to construct just a perfect tiling, it would be sufficient to take the intersection of all tiles of any ρ_n-cubic tiling with the set we want to cover $\mathcal{L}(t)$. However this process may yield tiles arbitrarily small along the boundary of the set. This is not any problem if h_n does not vanish, because the side ρ_n of the tiles is then not vanishing and the fixed ρ_n-tiling does not encounter any problem. Otherwise, if $h_n \to 0$, that is in the sparse regime, then the side ρ_n vanishes too (it is proportional to h_n). For each n, there might be a tile $C_n \in \mathcal{T}_n$, such that α_{C_n} is vanishing and $0 < \underline{\lim}\, n\alpha_{C_n} < +\infty$. The bound of Proposition 7.7 would thus not ensure anymore that X_{C_n} is connected with probability tending to one. In addition to that, subtiles made from C_n would not be ensured to be not empty with probability tending to one.

In the following lemma, we add an n as subscript to recall that it takes into account the possible case $h = h_n$.

Lemma 7.7. *"Puzzle". For ρ small enough, there exists a perfect tiling of $\mathcal{L}(t)$, denoted by $\mathcal{T}_n^{\oplus}$, and a perfect $\mathcal{T}_n^{\oplus}$-autocover $\mathcal{T}_n^{\circ}$ such that:*

- *For all $c \in \mathcal{T}_n^{\oplus}$, $\mathrm{Vol}(c) \geq (\rho_n/2)^d$*
- *For all $C \in \mathcal{T}_n^{\circ}$, $\rho_n^d \leq \mathrm{Vol}(C) \leq (3\rho_n/2)^d$*

The proof basically consists in sticking the small pieces of subtiles — cut off by the intersection with $\mathcal{L}(t)$ — to complete neighbouring subtiles.

Proof. Let $\mathcal{T}_n$ a ρ_n-cubic tiling of $\mathcal{L}(t)$. We first take the intersection of each $\mathcal{T}_n$-subtile (each element of $\mathcal{T}_n^{\boxplus}$) with $\mathcal{L}(t)$ to yield a perfect tiling. As a result, the subtiles crossed by the boundary of $\mathcal{L}(t)$ are either reduced or removed:

$$\mathcal{T}^{\boxplus} \mapsto \left\{c \cap \mathcal{L}(t);\ c \in \mathcal{T}^{\boxplus}\right\}$$

This new set is a perfect tiling. Pieces of subtiles which are cut off by this operation are going to be sticked to neighbouring subtiles to prevent from having too small subtiles.

ρ_n is assumed to be small enough so that each subtile reduced by the intersection shares at least one corner with another subtile which is not cut. It is then possible to map each $c \in \mathcal{T}_n^{\boxplus}$ such that $c \cap \mathcal{L}(t) \subsetneq c$, to $a(c) \in \mathcal{T}^{\boxplus}$, where $a(c)$ is any tile c' from $\mathcal{T}^{\boxplus}$, sharing a common corner with c and $c' \subset \mathcal{L}(t)$.

In other words, we are able to stick the reduced subtiles — those such that $c \cap \mathcal{L}(t) \subsetneq c$ — to an entire neighbouring subtile, $a(c)$, so as to shape a new subtile, bigger than those from $\mathcal{T}^{\boxplus}$. Note that several pieces of subtiles can be sticked to the same entire subtile: for a given entire subtile — such that $c \subset \mathcal{L}(t)$ — the set of the cut subtiles which are sticked to c is $a^{-1}(c)$. Thus the new tiling is the following:

$$\mathcal{T}_n^{\oplus} = \left\{c \cup a^{-1}(\{c\});\ c \in \mathcal{T}^{\boxplus} \text{ and } c \subset \mathcal{L}(t)\right\}$$

$\mathcal{T}_n^{\oplus}$ is a perfect tiling, and unlike the previous one above, each subtile has a volume greater than a non-positive constant, because it contains at least one $\mathcal{T}_n$-subtile, whose side is $\rho_n/2$. Hence for all $c \in \mathcal{T}_n^{\oplus}$, $\mathrm{Vol}(c) \geq (\rho_n/2)^d$. In this manuscript, we admit that each tile of $\mathcal{T}_n^{\oplus}$ shares at least one face with another tile of $\mathcal{T}_n^{\oplus}$.

Define now $\mathcal{T}^{\bigcirc}$ the set of the tiles formed by taking the union of 2^d subtiles of $\mathcal{T}_n^{\oplus}$ sharing one common corner. It is a perfect cover.

Moreover each subtile of $\mathcal{T}_n^{\oplus}$ contains one regular cubic subtile, therefore each tile $C \in \mathcal{T}_n^{\bigcirc}$ contains one regular ρ_n-cubic tile, and the volume of C is bounded from below by ρ_n^d. Furthermore if ρ_n is small enough, there is no more than a stripe of subtiles bordering half of the ρ_n-cubic tile which can be sticked to it. Hence the final tiles — made of a regular ρ_n-cubic tile plus the sticked subtiles — are not bigger than an hypercube made of 3^d subtiles. Thus an upper bound of their volume is $(3\rho_n/2)^d$.

Finally, it is also a $\mathcal{T}_n^{\oplus}$-autocover. Indeed if $C, C' \in \mathcal{T}_n^{\bigcirc}$ are adjacent, their common face borders 2^{d-1} subtiles of C and 2^{d-1} others of C', the union of which form one tile $O \in \mathcal{T}_n^{\bigcirc}$ as described in the definition of the autocover.

□

7.C.6 Proofs of the bounds of Section 7.3.3

Proof of Proposition 7.9

According to Lemma 7.7 Puzzle, the volume of a subtile of $\mathcal{T}_n^{\oplus}$ is bounded from below by $(\rho_n/2)^d$. Hence, for all $c \in \mathcal{T}_n^{\oplus}$:

$$P(V_c = \varnothing) = \left(1 - \int_c f(z)dz\right)^n \leq \left(1 - t(\rho_n/2)^d\right)^n \leq \exp\left(-\frac{t}{2^d} n\rho_n^d\right)$$

Proof of Proposition 7.10

First note that the diameter all tiles of $\mathcal{T}_n^{\bigcirc}$ is smaller than γh: using an argument of the proof of the Lemma 7.7, all tiles $C \in \mathcal{T}_n^{\bigcirc}$ are at most included in an hypercube made of 3^d cubic subtiles of side $\rho_n/2$, therefore the diameter of C is at most $3\rho_n\sqrt{d}/2 = \gamma h$.

The bound of the proposition comes from Proposition 7.3. We just need to bound α_C for all $C \in \mathcal{T}_n^{\bigcirc}$. Indeed α_C is related to the volume of C and the Lemma Puzzle 7.7 is helpful for bounding the volume:

$$t\rho_n^d \leq t\,\mathrm{Vol}(C) \leq \alpha_C = \int_C f(z)dz \leq \|f\|_\infty \mathrm{Vol}(C) \leq \|f\|_\infty \left(\frac{3\rho_n}{2}\right)^d$$

Number of subtiles in $\mathcal{T}_n^{\oplus}$: proof of Proposition 7.11

Firstly there are less subtiles in $\mathcal{T}_n^{\oplus}$ than in $\mathcal{T}_n^{\boxplus}$ because of the cutting off and the sticking.

Since f is continuously differentiable in the neighborhood of $\{f = t\}$, the box-counting dimension of $\mathcal{L}(t)$ is d. Therefore the number of ρ_n-cubic tiles required to cover it is $\mathcal{O}\left((\rho_n/2)^{-d}\right)$; there exists a non-negative constant K_2 — independent of n and ρ_n — such that for n large enough:

$$\mathrm{Card}(\mathcal{T}_n^{\boxplus}) \leq K_2 \left(\frac{\rho_n}{2}\right)^{-d} \tag{7.10}$$

Number of subtiles in $\mathcal{T}_n^{\bigcirc}$: proof of Proposition 7.12

It will be useful to first bound the number of tiles in $\mathcal{T}^{\square}$ and the number of subtiles in $\mathcal{T}^{\boxplus}$. In fact they are related to each other by the following inequality:

Proposition 7.21.

$$\mathrm{Card}(\mathcal{T}_n^{\square}) \leq \mathrm{Card}(\mathcal{T}_n^{\boxplus})$$

Proof. Since the sides of the (sub)-tiles are parallel to the axes, all corners of each (sub)-tile can be denoted by one vector of $\mathbb{R}^d$ from $\{-1,1\}^d$, because there is one and only one corner in the direction of each of these vectors from the center of the (sub)-tile. One of these directions is chosen, and we consider the set S of all corners which are in this direction from the center of one of the subtiles in $\mathcal{T}_n^{\boxplus}$. In the two-dimensional case, if the axes were called North-South and East-West, and the direction chosen was $(-1,1)$ for instance, we would consider the set of all corners which are located South-East of a subtile from $\mathcal{T}_n^{\boxplus}$.

Firstly each tile of $\mathcal{T}_n^{\square}$ is made by taking the union of subtiles sharing the same corner, therefore the set S has more elements than $\mathcal{T}_n^{\square}$, because each corner which is shared by 2^d subtiles is necessarily in the chosen direction for one of the 2^d subtiles.

Secondly, each corner which is in the chosen direction from the center of a subtile can be injectively mapped to this subtile, then:

$$\mathrm{Card}(\mathcal{T}_n^{\square}) \leq \mathrm{Card}(S) \leq \mathrm{Card}(\mathcal{T}_n^{\boxplus})$$

□

We conclude with this lemma to prove Proposition 7.12.

Let us notice that there are less tiles in the perfect auto-cover $\mathcal{T}_n^{\bigcirc}$ than in the perfect cover $\mathcal{T}_n^{\square}$, just because the subtiling it has been built from contains less subtiles (see the previous construction in Lemma 7.7).

From the inequality (7.10) in the proof of Proposition 7.11, we then derive:

$$\mathrm{Card}(\mathcal{T}_n^{\bigcirc}) \leq \mathrm{Card}(\mathcal{T}_n^{\square}) \leq \mathrm{Card}(\mathcal{T}_n^{\boxplus}) \leq K_2 \left(\frac{\rho_n}{2}\right)^{-d}$$

7.D Supplementary materials for the non-overestimation of $Q(t)$

In Section 7.3 it was proved that $X_{\mathcal{L}(t)}$, which is thought of as a graph asymptotically close to $\widehat{X}(t)$, does not overestimate $Q(t)$ asymptotically. Despite this apparent similarity, the remaining gap between these graphs is actually still big. The goal of this section is to enlighten what we have to deal with, to reach the complete consistency of the algorithm.

7.D.1 Non-overestimation

Let $\widehat{X}^{\varepsilon}(t)$ the graph induced by the set $\widehat{J}(t) \cap J_n(t+\varepsilon)$, that is the graph whose nodes are selected by the algorithm and are in the $(t+\varepsilon)$-level set of f as well. Define also $\widehat{Q}_n^{\varepsilon}(t)$ the number of connected components of $\widehat{X}^{\varepsilon}(t)$.

Theorem 7.6. *Under Assumptions 7.2,*

$$P(\widehat{X}^{\varepsilon_n}(t) = X_{\mathcal{L}(t+\varepsilon_n)}) \leq K_6 n \exp\left(-K_7 n h_n^d \varepsilon_n^2\right)$$

and:

$$P(\widehat{Q}_n^{\varepsilon}(t) > Q(t)) \leq K_6 n \exp\left(-K_7 n h_n^d \varepsilon_n^2\right)$$

where K_6, K_7 are non-negative constant depending on d, f, k and t.

The proof is a consequence of Propositions 7.4 and of Proposition 7.6.(2). The main argument is similar to Section 7.2.1: here it is proved that if the density is not too much overestimated, then all nodes from $\mathcal{L}(t+\varepsilon)$ are all selected by the algorithm.

Proof. According to Proposition 7.4, for all $\varepsilon > 0$:

$$\{U_n \leq \varepsilon\} \subset \left\{J_n(t+\varepsilon) \subset \widehat{J}_n(t)\right\}$$

As a consequence, on the event $\{U_n \leq \varepsilon_n\}$, $J_n(t+\varepsilon_n) \subset \widehat{J}_n(t) \cap J_n(t+\varepsilon_n)$. The other inclusion is also true, and therefore we have:

$$\begin{aligned}
\{U_n \leq \varepsilon_n\} &\subset \left\{J_n(t+\varepsilon_n) = \widehat{J}_n(t) \cap J_n(t+\varepsilon_n)\right\} \\
&\subset \left\{\widehat{X}^{\varepsilon_n}(t) = X_{\mathcal{L}(t+\varepsilon_n)}\right\} \\
&\subset \left\{\widehat{Q}_n^{\varepsilon_n}(t) = \widetilde{Q}_n(t+\varepsilon_n)\right\}
\end{aligned}$$

The first line below comes from the assumption that for all $n \in \mathbb{N}^*$, $\varepsilon_n \leq \varepsilon_0$, so that $Q(t+\varepsilon_n) = Q(t)$ (see the comment of Assumption 7.1). Then the second line is the total probability formula. The fourth line follows from Propositions 7.2 and 7.6.(2).

$$\begin{aligned}
P\left(\widehat{Q}_n^{\varepsilon_n}(t) > Q(t)\right) &= P\left(\widehat{Q}_n^{\varepsilon_n}(t) > Q(t+\varepsilon_n)\right) \\
&= P\left(\widehat{Q}_n^{\varepsilon_n}(t) > Q(t+\varepsilon_n),\ U_n \leq \varepsilon_n\right) \\
&\quad + P^{U_n > \varepsilon_n}\left(\widehat{Q}_n^{\varepsilon_n}(t) > Q(t+\varepsilon_n)\right) P(U_n > \varepsilon_n) \\
&\leq P\left(\widehat{Q}_n^{\varepsilon_n}(t) > Q(t+\varepsilon_n), \widehat{Q}_n^{\varepsilon_n}(t) = \widetilde{Q}_n(t+\varepsilon_n)\right) + P(U_n > \varepsilon_n) \\
&\leq P\left(\widetilde{Q}_n^{\varepsilon_n}(t) > Q(t+\varepsilon_n)\right) + P(U_n > \varepsilon_n) \\
&\leq K_0 n \exp(-K_1 n h_n^d) + 6n \exp\left(-\frac{1}{32}\frac{\varepsilon_n^2 n h_n^d}{\|f\|_\infty + \varepsilon_n}\right) \\
&\leq K_6 n \exp\left(-K_7 n h_n^d \varepsilon_n^2\right)
\end{aligned}$$

where K_6, K_7 are non-negative constant depending on d, f, k and t. □

Comment. This theorem highlights that most of the problem to prove the non-overestimation of $Q(t)$ is controlling the nodes near the boundary of the t-level set. On the event $\{U_n \leq \varepsilon\}$, nodes being in $\mathcal{L}(t)\backslash\mathcal{L}(t+\varepsilon)$ are not necessarily in the graph $\widehat{X}(t)$ unlike nodes being in $\mathcal{L}(t+\varepsilon)$. Moreover some nodes from $\mathcal{L}(t-\varepsilon)\backslash\mathcal{L}(t)$ may be in the graph $\widehat{X}(t)$. As a consequence we have to deal with uncertainty about these nodes, which makes hard the control of the connectedness inside the clusters. Connectivity is a very delicate property of graphs and each of these nodes may be isolated for instance.

Biau et al. (2007) do not encounter this problem, because in their setting, the connection radius h'_n and the bandwidth h_n of the density estimator are not the same. They assume $\varepsilon_n = o(h'_n)$, and the boundary of $\mathcal{L}(t)$ is therefore a negligible region of the balls of the covering, which is used in their Proposition A.2.

7.D.2 Local connectedness in the neighborhood of the boundary of $\mathcal{L}(t)$

We follow the strategy of Section 7.3, and we would like to use the local connectedness as well to prove the non-overestimation by $\widehat{Q}_n(t)$. Let $\mathcal{T}_n$ be a ρ_n-cubic tiling of $\mathcal{L}(t-\varepsilon_n)$. For all tiles of $\mathcal{T}_n$ included in $\mathcal{L}(t+\varepsilon_n)$, we already know that the local connectedness holds using the previous theorem: the two subgraphs of $\widehat{X}(t)$ and $X_{\mathcal{L}(t)}$ induced by such a tile are actually equal. The problematic tiles are clearly those of the boundary.

A motivating idea would be to mimic what the algorithm does in each tile, but in an approximate and simplifying manner so that we can easily bound with standard methods the probability of local connectedness. Let C be a tile of $\mathcal{T}_n$. Let Γ_t the graph operator which removes nodes whose degree is smaller than t. Then we have:

$$\widehat{X}_C = \Gamma_t X_C$$

Nodes of X_C are removed by the algorithm using their degrees and therefore this procedure is not independent of X_C itself. However, if the degree is replaced by the outer degree, that is the number of neighbors of the node, which are not in C, the procedure becomes independent of X_C. The outer degree, denoted by D_i^{out} satisfies the following:

$$D_i^{\text{out}} = \sum_{j \in [n]\backslash V_C} X_{ij} \text{ and } D_i^{\text{out}} \leq D_i$$

Γ_t^C denotes the same operator as Γ_t, but based on the outer degree. Then we can write:

$$\begin{aligned} P(\Gamma_t X_C \text{ is disconnected}) &= P(\Gamma_t X_C \text{ is disconnected, } \Gamma_t^C \text{ is disconnected}) \\ &\quad + P(\Gamma_t X_C \text{ is disconnected, } \Gamma_t^C \text{ is connected}) \\ &\leq P(\Gamma_t^C \text{ is disconnected}) \\ &\quad + P(\Gamma_t X_C \text{ is disconnected, } \Gamma_t^C \text{ is connected}) \end{aligned}$$

Further developments can deal with bounding these two terms. The bound of the first term can be inspired by the scheme of the proof of Proposition 7.7, in a situation where nodes are randomly removed from the underlying Erdős-Rényi random graph. Thus involves the number of nodes of the pruned graph, which has to be controlled.

The second term depends on how Γ_t^C differs from $\Gamma_t X_C$. The difference between the degree and the outer degree may be not negligible unless it is assumed that $\rho_n = o(h_n)$. Indeed the average number of neighbors a node of C has in C is then negligible in comparison with the average number of nodes i can be connected to, and therefore the difference between the degrees should not change the thresholding with high probability.

7.E Proofs of Section 7.4

7.E.1 Proof of Proposition 7.13

Conditionally on Z_i, T_i is a binomial distributed variable, because the Bernoulli variables $(X_{ij})_{j\in[n]}$ are independent and have the same parameter, denoted by $\theta_n(Z_i)$, where:

$$\theta_n(z) = P^{Z_i=z}(X_{ij} = 1) = \int_{\mathbb{R}^d} k\left(\frac{z-u}{h_n}\right) f(u)du = h_n^d k_{h_n} \star f(z).$$

For all $z \in \mathbb{R}^d$, the parameter $\theta_n(z)$ is bounded as follows, with the Proposition 5.5 in Chapter 5 (see also Prakasa Rao, 1983):

$$h_n^d(f(z) - s_{h_n}) \leq \theta_n(z) \leq h_n^d(f(z) + s_{h_n}) \tag{7.11}$$

First, the probability that T_i exceeds t conditionally on the position of the node i is bounded from above:

$$\begin{aligned} P^{Z_i}(T_i \geq t) &= P^{Z_i}\left(D_i - n\theta_n(Z_i) \geq nh_n^d t - n\theta_n(Z_i)\right) \\ &\leq P^{Z_i}\left(D_i - n\theta_n(Z_i) \geq nh_n^d t - n\theta_n(Z_i)\right) \\ &\leq P^{Z_i}\left(D_i - n\theta_n(Z_i) \geq nh_n^d(t - f(Z_i) - s_{h_n})\right) \end{aligned}$$

(using right hand side of (7.11)). Now we integrate over the event $\{Z_i \notin \mathcal{L}(t-\varepsilon_n)\}$:

$$P^{Z_i \notin \mathcal{L}(t-\varepsilon_n)}(T_i \geq t) P(Z_i \notin \mathcal{L}(t-\varepsilon_n)) = \int_{\mathcal{L}(t-\varepsilon_n)^c} P^{Z_i=z}(T_i \geq t) f(z) dz.$$

And $P(Z_i \notin \mathcal{L}(t-\varepsilon_n)) = 1 - P(Z_i \in \mathcal{L}(t-\varepsilon_n)) \geq 1 - P(Z_i \in \mathcal{L}(t)) = 1 - \mu_f(t)$. Hence:

$$\begin{aligned} P^{Z_i \notin \mathcal{L}(t-\varepsilon_n)}(T_i \geq t) &\leq \frac{1}{1-\mu_f(t)} \int_{\mathcal{L}(t-\varepsilon_n)^c} P^{Z_i=z}(D_i - n\theta_n(z) \geq nh_n^d(t - f(z) - s_{h_n}) f(z) dz \\ &\leq \frac{1}{1-\mu_f(t)} \int_{\mathcal{L}(t-\varepsilon_n)^c} P^{Z_i=z}(D_i - n\theta_n(z) \geq nh_n^d \varepsilon_n') f(z) dz \end{aligned} \tag{7.12}$$

because for all $z \in \mathcal{L}(t-\varepsilon_n)^c$, $t - f(z) - s_{h_n} \geq t - (t-\varepsilon_n) - s_{h_n} = \varepsilon_n - s_{h_n} = \varepsilon_n'$. Furthermore under Assumption 7.4 $\varepsilon_n' = \varepsilon_n - s_{h_n} = \mu\varepsilon_n \geq 0$, therefore we have for all $z \in \mathcal{L}(t-\varepsilon_n)^c$ by using Bernstein's inequality:

$$\begin{aligned} P^{Z_i=z}(D_i - n\theta_n(z) \geq nh_n^d \varepsilon_n') &\leq \exp\left(-\frac{1}{2} \frac{(nh_n^d \varepsilon_n')^2}{n\theta_n(z)(1-\theta_n(z)) + \frac{nh_n^d \varepsilon_n'}{3}} \right) \\ &\leq \exp\left(-\frac{1}{2} \frac{(nh_n^d \varepsilon_n')^2}{nh_n^d k_{h_n} \star f(z) + \frac{nh_n^d \varepsilon_n'}{3}} \right) \\ &= \exp\left(-\frac{1}{2} \frac{nh_n^d \varepsilon_n'^2}{k_{h_n} \star f(z) + \frac{\varepsilon_n'}{3}} \right) \\ &\leq \exp\left(-\frac{1}{2} \frac{nh_n^d \varepsilon_n'^2}{\|f\|_\infty + \varepsilon_n} \right) \end{aligned} \tag{7.13}$$

The last line (7.13) follows from this upper bound of the denominator (with $s_{h_n} \leq \varepsilon_n$ once more):

$$k_{h_n} \star f(z) + \frac{\varepsilon_n'}{3} \leq f(z) + s_{h_n} + \frac{\varepsilon_n - s_{h_n}}{3} = f(z) + \frac{\varepsilon_n + 2s_{h_n}}{3} \leq \|f\|_\infty + \varepsilon_n.$$

Now combining (7.13), which is an uniform bound with respect to z, and (7.12), we establish that:

$$P^{Z_i \notin \mathcal{L}(t-\varepsilon_n)}(T_i \geq t) \leq \frac{1}{1-\mu_f(t)} \exp\left(-\frac{1}{2} \frac{nh_n^d \varepsilon_n'^2}{\|f\|_\infty + \varepsilon_n} \right) \tag{7.14}$$

Using Assumption 7.4, $\varepsilon_n' = \varepsilon_n - s_{h_n} \geq \mu\varepsilon_n$, we finally have the inequality of the proposition.

The proof of the second inequality is definitely similar, and we just give some pieces of it. Using left hand side of (7.11), we have this time:

$$P^{Z_i}(T_i < t) \leq P^{Z_i}\left(D_i - n\theta_n(Z_i) \geq -nh_n^d(f(Z_i) - t - s_{h_n})\right).$$

For all $z \in \mathcal{L}(t+\varepsilon_n)$, $f(z) - t - s_{h_n} \geq (t+\varepsilon_n) - t - s_{h_n} = \varepsilon'_n$. By integration over $\{Z_i \in \mathcal{L}(t+\varepsilon_n)\}$, the analogous inequality of (7.12) is:

$$P^{Z_i \in \mathcal{L}(t+\varepsilon_n)}(T_i < t) P(Z_i \in \mathcal{L}(t+\varepsilon_n)) \leq \int_{\mathcal{L}(t+\varepsilon_n)} P^{Z_i = z}(D_i - n\theta_n(z) \leq -nh_n^d \varepsilon'_n) f(z) dz.$$

Since $\varepsilon \leq \varepsilon_0$, we have $P(Z_i \in \mathcal{L}(t+\varepsilon_n)) \geq \mu_f(t+\varepsilon_0)$. Moreover the right hand side of the inequality inside $P^{Z_i \in \mathcal{L}(t-\varepsilon_n)}$ is the same as (7.12) up to the sign, hence the Bernstein's inequality gives the same bound, and the computation goes on alike.

Bibliography

Jean-Jacques Daudin Laurent Pierre Alain Celisse et al. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.*, 6:1847–1899, 2012. ISSN 1935-7524. doi: 10.1214/12-EJS729.

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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011.

Ery Arias-Castro and Nicolas Verzelen. Community detection in random networks. *arXiv preprint arXiv:1302.7099*, 2013.

Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *The Annals of Probability*, 17 (1):9–25, 1989.

Adelchi Azzalini and Nicola Torelli. Clustering via nonparametric density estimation. *Statistics and Computing*, 17(1):71–80, 2007.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

Andrew Barbour and Denis Mollison. Epidemics and random graphs. *Stochastic processes in epidemic theory*, 86:86–89, 1990.

Andrew D Barbour, Lars Holst, and Svante Janson. *Poisson approximation*. Clarendon press Oxford, 1992.

Matthew James Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

Marcel Berger and Bernard Gostiaux. *Géométrie différentielle: variétés, courbes et surfaces*. Presses Universitaires de France-PUF, 1992.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

Julian Besag. Statistical analysis of non-lattice data. *The statistician*, pages 179–195, 1975.

Julian Besag. Markov chain Monte Carlo for statistical inference. *Center for Statistics and the Social Sciences*, 2001.

G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM: P&S*, 11:272–280, 2007.

P.J. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068, 2009.

P.J. Bickel, A. Chen, and E. Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):38–59, 2011.

Etienne Birmele. Detecting local network motifs. *Electronic Journal of Statistics*, 6:908–933, 2012.

Vincent Blondel, Gautier Krings, and Isabelle Thomas. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies*, 42(4), 2010.

B. Bollobás. *Random graphs*. Cambridge Univ Pr, 2001.

B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007. ISSN 1098-2418.

Béla Bollobás and Oliver Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks*, 1:34, 2003.

Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004a.

Béla Bollobás and Oliver Riordan. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1(1):1–35, 2004b.

Béla Bollobás, Oliver Riordan, Joel Spencer, Gábor Tusnády, et al. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.

Ronald L Breiger, Scott A Boorman, and Phipps Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3):328–383, 1975.

M Broniatowski, G Celeux, and J Diebolt. Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data analysis and informatics*, 3:359–373, 1983.

Pierce G Buckley and Deryk Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Mathematics*, 282(1):53–68, 2004.

Gilles Celeux and Jean Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quarterly*, 2(1):73–82, 1985.

Gilles Celeux and Jean Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics: An International Journal of Probability and Stochastic Processes*, 41(1-2):119–134, 1992.

Gilles Celeux and Gérard Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3): 315–332, 1992.

Antoine Channarond, Jean-Jacques Daudin, and Stéphane Robin. Classification and estimation in the Stochastic Blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601, 2012.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343–351, 2010.

Louis HY Chen. Poisson approximation for dependent trials. *The Annals of Probability*, pages 534–545, 1975.

D.S. Choi, P.J. Wolfe, and E.M. Airoldi. Stochastic blockmodels with growing number of classes. *Arxiv preprint arXiv:1011.4644*, 2010.

Fan Chung, Linyuan Lu, T Gregory Dewey, and David J Galas. Duplication models for biological networks. *Journal of Computational Biology*, 10(5):677–687, 2003.

A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Estimating the number of clusters. *Canadian Journal of Statistics*, 28(2):367–382, 2000.

Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. Cluster analysis: a further approach based on density estimation. *Computational Statistics & Data Analysis*, 36(4):441–459, 2001.

J.J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and computing*, 18(2):173–183, 2008.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Pierre A Devijver. Baum's forward-backward algorithm revisited. *Pattern Recognition Letters*, 3(6):369–373, 1985.

Jean Diebolt and Christian P Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 363–375, 1994.

Sander Dommers, Remco van der Hofstad, and Gerard Hooghiemstra. Diameters in preferential attachment models. *Journal of Statistical Physics*, 139(1):72–107, 2010.

William E Donath and Alan J Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

M Draief and L Massoulié. Epidemics and rumours in complex networks, volume 369 of London Mathematical Society Lecture Notes, 2010.

Emile Durkheim. *De la division du travail social: étude sur l'organisation des sociétés supérieures*. F. Alcan, 1893.

Richard Durrett. *Random graph dynamics*, volume 20. Cambridge university press, 2007.

P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

Paul Erdős and Alfred Renyi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist*, 38(4):343–347, 1961.

Leonhard Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

Shimon Even. *Graph algorithms*. Cambridge University Press, 2011.

K. Faust and S. Wasserman. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.

S.E. Fienberg and S.S. Wasserman. Categorical data analysis of single sociometric relations. *Sociological methodology*, 12:156–192, 1981. ISSN 0081-1750.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

Ove Frank and David Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

E.N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144, 1959. ISSN 0003-4851.

M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821, 2002.

Piyush Gupta and Panganamala R Kumar. Critical power for asymptotic connectivity in wireless networks. In *Stochastic analysis, control, optimization and applications*, pages 547–566. Springer, 1998.

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.

John Harris, Jeffry L Hirst, and Michael J Mossinghoff. *Combinatorics and graph theory*. Springer, 2008.

John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

John A Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

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.

P.W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373): 33–50, 1981. ISSN 0162-1459.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

David R Hunter and Mark S Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3): 565–583, 2006.

T.S. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, pages 129–159, 2000.

Svante Janson, Donald E Knuth, Tomasz Łuczak, and Boris Pittel. The birth of the giant component. *Random Structures & Algorithms*, 4(3):233–358, 1993.

Svante Janson, Tomasz Luczak, and VF Kolchin. *Random graphs*. Cambridge Univ Press, 2000.

Jon Kleinberg. The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

Jon M Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S Tomkins. The web as a graph: Measurements, models, and methods. In *Computing and combinatorics*, pages 1–17. Springer, 1999.

Eric D Kolaczyk. *Statistical analysis of network data*. Springer, 2009.

Jacques Lafontaine. *Introduction aux variétés différentielles*. SOFEDIS, 2012.

A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11:033015, 2009.

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.

Jean-Benoist Leger, Corinne Vacher, and Jean-Jacques Daudin. Detection of structurally homogeneous subsets in graphs. *Statistics and Computing*, pages 1–18, 2013.

Charles E Leiserson, Ronald L Rivest, Clifford Stein, and Thomas H Cormen. *Introduction to algorithms*. The MIT press, 2001.

Robert F Ling. A probability theory of cluster analysis. *Journal of the American Statistical Association*, 68(341):159–164, 1973.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 539. Wiley New York, 1987.

F. Lorrain and H.C. White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80, 1971.

Alfred James Lotka. The frequency distribution of scientific productivity. *Journal of Washington Academy Sciences*, 1926.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957, 2006.

David D McFarland and Daniel J Brown. Social Distance as Metric: A Systematic Introduction to Smallest Space Analysis. *EO Laumann. Bonds of Pluralism: The Form and Substance of Urban Social Networks. New York: John Wiley*, pages 213–252, 1973.

Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

Bojan Mohar and Y Alavi. The Laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, 2:871–898, 1991.

Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

Mark EJ Newman and Duncan J Watts. Renormalization group analysis of the small-world network model. *Physics Letters A*, 263(4):341–346, 1999a.

Mark EJ Newman and Duncan J Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60(6):7332, 1999b.

M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577, 2006.

M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

B. Pelletier and P. Pudlo. Operator norm convergence of spectral clustering on level sets. *The Journal of Machine Learning Research*, 12:385–416, 2011.

Mathew Penrose. *Random geometric graphs*, volume 5. Oxford University Press Oxford, 2003.

F. Picard, V. Miele, J.J. Daudin, L. Cottret, and S. Robin. Deciphering the connectivity structure of biological networks using MixNet. *BMC bioinformatics*, 10(Suppl 6):S17, 2009.

BLS Prakasa Rao. *Nonparametric functional estimation*. Academic Press, New York, 1983.

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (< i> p< sup>*</sup>) models for social networks. *Social networks*, 29(2):173–191, 2007.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional Stochastic Block Model. *Arxiv preprint arXiv:1007.1684*, 2010.

T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic block-models for graphs with latent block structure. *Journal of Classification*, 14(1): 75–100, 1997.

Tom AB Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

Tom AB Snijders, Philippa E Pattison, Garry L Robins, and Mark S Handcock. New specifications for exponential random graph models. *Sociological methodology*, 36(1):99–153, 2006.

David Strauss and Michael Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

Aleksandr B Tsybakov. *Introduction à l'estimation non paramétrique*, volume 41. Springer, 2003.

R. Van Der Hofstad. Random graphs and complex networks. *Available on http://www. win. tue. nl/rhofstad/NotesRGCN. pdf*, 2009.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17 (4):395–416, 2007.

U. Von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

Ulrike Von Luxburg and Shai Ben-David. Towards a statistical theory of clustering. In *Pascal workshop on statistics and optimization of clustering*, 2005.

M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2): 1–305, 2008.

B. Wang and D. Titterington. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, page 584. AUAI Press, 2004a.

Bo Wang and DM Titterington. Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20(3): 151–170, 2004b.

Bo Wang and DM Titterington. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

Stanley Wasserman and Garry L Robins. An introduction to random graphs, dependence graphs, and p*. *Models and methods in social network analysis*, 27: 148–161, 2005.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

Greg CG Wei and Martin A 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, 1990.

CF Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

G Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.

Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

Recherche de structure dans un graphe aléatoire: modèles à espace latent

Résumé: Cette thèse aborde le problème de la recherche d'une structure (ou clustering) dans les noeuds d'un graphe. Dans le cadre des modèles aléatoires à variables latentes, on attribue à chaque noeud i une variable aléatoire non observée (latente) Z_i, et la probabilité de connexion des noeuds i et j dépend conditionnellement de Z_i et Z_j. Contrairement au modèle d'Erdős-Rényi, les connexions ne sont pas indépendantes identiquement distribuées; les variables latentes régissent la loi des connexions des noeuds. Ces modèles sont donc hétérogènes, et leur structure est décrite par les variables latentes et leur loi; ce pourquoi on s'attache à en faire l'inférence à partir du graphe, seule variable observée.

La volonté commune des deux travaux originaux de cette thèse est de proposer des méthodes d'inférence de ces modèles, consistentes et de complexité algorithmique au plus linéaire en le nombre de noeuds ou d'arêtes, de sorte à pouvoir traiter de grands graphes en temps raisonnable. Ils sont aussi tous deux fondés sur une étude fine de la distribution des degrés, normalisés de façon convenable selon le modèle.

Le premier travail concerne le Stochastic Blockmodel. Nous y montrons la consistence d'un algorithme de classification non supervisée à l'aide d'inégalités de concentration. Nous en déduisons une méthode d'estimation des paramètres, de sélection de modèles pour le nombre de classes latentes, et un test de la présence d'une ou plusieurs classes latentes (absence ou présence de clustering), et nous montrons leur consistence.

Dans le deuxième travail, les variables latentes sont des positions dans l'espace $\mathbb{R}^d$, admettant une densité f, et la probabilité de connexion dépend de la distance entre les positions des noeuds. Les clusters sont définis comme les composantes connexes de l'ensemble de niveau $t > 0$ fixé de f, et l'objectif est d'en estimer le nombre à partir du graphe. Nous estimons la densité en les positions latentes des noeuds grâce à leur degré, ce qui permet d'établir une correspondance entre les clusters et les composantes connexes de certains sous-graphes du graphe observé, obtenus en retirant les noeuds de faible degré. En particulier, nous en déduisons un estimateur du nombre de clusters et montrons sa consistence en un certain sens.

Clustering in a random graph: models with latent space

Abstract: This thesis addresses the clustering of the nodes of a graph, in the framework of random models with latent variables. To each node i is allocated an unobserved (latent) variable Z_i and the probability of nodes i and j being connected depends conditionally on Z_i and Z_j. Unlike Erdős-Renyi's model, connections are not independent identically distributed; the latent variables rule the connection distribution of the nodes. These models are thus heterogeneous and their structure is fully described by the latent variables and their distribution. Hence we aim at infering them from the graph, which the only observed data.

In both original works of this thesis, we propose consistent inference methods with a computational cost no more than linear with respect to the number of nodes or edges, so that large graphs can be processed in a reasonable time. They both are based on a study of the distribution of the degrees, which are normalized in a convenient way for the model.

The first work deals with the Stochastic Blockmodel. We show the consistency of an unsupervised classification algorithm using concentration inequalities. We deduce from it a parametric estimation method, a model selection method for the number of latent classes, and a clustering test (testing whether there is one cluster or more), which are all proved to be consistent.

In the second work, the latent variables are positions in the $\mathbb{R}^d$ space, having a density f. The connection probability depends on the distance between the node positions. The clusters are defined as connected components of some level set of f. The goal is to estimate the number of such clusters from the observed graph only. We estimate the density at the latent positions of the nodes with their degree, which allows to establish a link between clusters and connected components of some subgraphs of the observed graph, obtained by removing low degree nodes. In particular, we thus derive an estimator of the cluster number and we also show the consistency in some sense.