UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Étude morpho-statistique des réseaux sociaux. Application aux collaborations inter-organisationnelles.

THÈSE

présentée et soutenue publiquement le 12 janvier 2021

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Quentin Laporte

Composition du jury

Président :	Radu Stoica	Professeur, Université de Lorraine
Rapporteures :	Myriam Maumy-Bertrand	Maître de Conférences, Université de Technologie de Troyes
	Florence Sedes	Professeur, Université Toulouse 3
Examinateur :	Arnaud Martin	Professeur, Université Rennes 1
Invitée :	Marianne Clausel	Professeur, Université de Lorraine
Encadrants :	François Charoy	Professeur, Université de Lorraine
	Gérald Oster	Maître de Conférences, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Je tiens en premier lieu à remercier mes directeurs de thèse : François Charoy et Gérald Oster de m'avoir accordé leur confiance tout au long de ces années. Ils m'ont apporté de précieux conseils et ont été un réel soutien humain et amical.

Je remercie les membres du jury pour l'intérêt qu'ils ont porté à ce travail de thèse, le temps qu'ils ont consacré à son évaluation et la bienveillance dont ils ont fait preuve tout au long de ce processus. Les remarques qui ont été soulevées m'ont beaucoup apporté scientifiquement, et m'ont aussi humainement et personnellement touché.

Je remercie Radu Stoica d'avoir accepté d'être président du jury. En réalité, il a été bien plus que cela. Avec Marianne Clausel, ils m'ont aidé à décortiquer et trouver du sens dans les mathématiques que j'ai utilisées. Merci pour votre enthousiasme, votre énergie et le temps que vous avez consacré à m'accompagner.

Je tiens également à remercier mes amis, mes collègues et ma famille. Ils ont tous aidé d'une façon ou d'une autre à conserver ce précieux équilibre qui m'a permis d'avancer.

À celles et ceux qui ont rythmé mon quotidien de thésard par leur bonne humeur, leur gentillesse, les discussions et les débats parfois passionnés, mais toujours passionnants et sans compter les pauses café, source d'inspiration indéniable pour une thèse. Je pense à mes collègues de bureau : Matthieu et Victorien, et plus généralement aux membres de l'équipe COAST : Abir, Claudia, Linda, Amina, Clélie, Béatrice, Riyadh, Hoai Le, Anis, Cédric, Olivier, Alexandre ... Je pense aussi à mes amis de TÉLÉCOM Nancy qui se sont également lancés dans l'aventure d'une thèse : Sylvain, Kevin et Jean-Philippe tous animés par la même passion, celle du film Vaiana, allez savoir pourquoi ... Merci à tous les autres collègues doctorant dont j'ai fait la connaissance au Loria : Margaux, Athénaïs, Justine, Charles, l'équipe du PhD Coffee Time et celle du PhD Welcome week-end. Merci aux membres de l'administration du Loria de m'avoir aidé à affronter la complexité administrative à laquelle les doctorants doivent faire face, merci pour votre soutien et vos précieux conseils, merci à Annick J., Anne C., Delphine H., Sophie D., Josiane R., Nathalie B.. Au sein du laboratoire, il y a celles et ceux qui ont contribué à ce travail par leur sourire, leurs attentions et leur énergie toujours positive qu'ils ont su me communiquer quotidiennement, merci à Isabelle, Caroline, Tareq et toute l'équipe d'API.

À celles et ceux qui sont toujours restés à mes côtés, peu importe la distance. J'ai une pensée particulière pour la "team Muscadet" (Thibaud, PLS, Alex, Pierre, Karl, Léo, Baptiste, PJ ...), merci pour votre amitié de longue date chère à mes yeux. À mes amis de Nancy, celles et ceux que j'ai rencontrés à TÉLÉCOM Nancy : Lolita, Miranda, Vincent vous avez été d'un réel soutien quand cela était nécessaire. La thèse c'est aussi des rencontres, dans les Vosges, pendant une formation pédagogique, un moment précieux où un petit groupe d'amis s'est formé, "l'équipe du Ventron" : Claire, Joséphine, Margaux, Cécile et Sylvain, merci pour ces soirées, ces moments d'amitié qui ne font que commencer. Enfin dans toutes ces aventures, il y en a certaines qui dépassent le cadre de la thèse, qui donnent du sens aux voyages et nous font redécouvrir l'amitié comme jamais, merci aux Amis de Poa de m'avoir accordé votre amitié et de m'avoir fait vivre toutes ces belles aventures de solidarité et de partage.

À ma famille, sans qui je ne serai pas là. À mes parents, ma belle aimée Claire, mes sœurs, mon frère, Bran, ma tante, ma cousine, mes grands-parents Ama et Michel, ma

grand-mère Line, ma marraine, Stéphane et Valérie. Merci pour votre indéfectible soutien, vous avez cru en moi quand, il faut se l'avouer, je n'y croyais plus trop. Merci pour votre amour, ingrédient essentiel pour avancer dans la vie. En écrivant ces lignes, j'ai une pensée émue pour celles et ceux qui ne sont plus là, mais qui y croyaient aussi, je les embrasse fort.

Sommaire

Table des figures

Table des abréviations

ABC	*Approximate Bayesian Computation*
CFP	*Coupling From the Past*
DOSNs	*Distributed Online Social Networks*
ERGM	*Exponential Random Graph Model*
F2F	*friend-to-friend*
GIEC	Groupe d'experts intergouvernemental sur l'évolution du climat
GWD	*Geometrically Weighted Degree*
ICIJ	*International Consortium of Investigative Journalists*
i.i.d	indépendantes et identiquement distribuées
K-S	Kolmogorov-Smirnov
Mad	*Median absolute deviation*
MAP	*Maximum a posteriori*
MCMC	*Markov Chain Monte-Carlo*
MCMLE	*Monte-Carlo Likelihood Estimation*
MCNR	*Monte-Carlo Newton Raphson*
MRF	*Markov Random Field*
SNA	*Social network analysis*
TNT	"*tie no tie*"

Chapitre 1

Introduction

1.1 Contexte de la thèse

Les applications collaboratives facilitent l'échange d'informations et le travail entre différentes parties. Au cours du temps, les usages ont évolué avec les progrès techniques et le développement d'internet. Ainsi, les utilisateurs peuvent éditer localement un document, mais aussi collaborer en temps réel avec d'autres utilisateurs. Des suites d'applications web collaboratives comme Google Suite ou Office 365 (de Microsoft) permettent d'éditer simultanément un document. Les utilisateurs peuvent ainsi apporter des modifications au document et percevoir en temps réel celles des autres collaborateurs. Ces applications démocratisent l'usage des outils de collaboration dématérialisés et permettent de s'affranchir des contraintes géographiques et temporelles (collaborations asynchrones).

La majeure partie de ces applications reposent sur un système centralisé. L'ensemble des fonctionnalités telles que la communication entre les collaborateurs ou encore la persistance des documents sont assurées par un seul et même acteur central. Cette configuration soulève plusieurs problèmes, le premier porte sur la capacité du système à passer à l'échelle, c'est-à-dire à supporter un grand nombre d'utilisateurs simultanément [21]. Cette situation peut se produire lorsqu'un grand nombre de personnes souhaite éditer un document dans le même laps de temps. La limite ne réside pas tant dans la capacité de l'infrastructure à supporter une telle charge, mais dans la maîtrise de son coût. Déployer une infrastructure technique puissante pour des fournisseurs de services comme Google ou Microsoft n'est pas un problème, mais le faire pour un grand nombre de documents reviendrait à faire exploser les coûts de production. À cette première problématique s'ajoute une seconde, qui, porte sur des considérations liées à la sécurité des données et à la préservation de la vie privée. L'utilisateur n'a aucun contrôle sur la façon dont les données qu'il fournit sont stockées, sécurisées ni même qui peut y avoir accès. D'autant que les géants du web (les GAFAM) sont régulièrement la cible d'attaques informatiques conduisant parfois à la fuite de données personnelles. De plus, certaines plateformes sont accusées de collusion avec des gouvernements étrangers et sont suspectées d'ouvrir leurs bases de données à des agences gouvernementales. Ces deux derniers constats menacent à la fois les données personnelles des individus, mais aussi

la souveraineté des entreprises et des administrations qui n'ont plus la maîtrise de leurs données.

Les problèmes énoncés sont intrinsèquement liés au caractère centralisé des applications collaboratives que nous venons de citer. L'adoption d'un autre type d'architecture décentralisé s'avère être une solution qui permet de ne plus se reposer sur ces silos d'information. Une telle solution implique que tous les utilisateurs sont directement interconnectés, ils peuvent alors échanger directement et librement avec les autres collaborateurs sans avoir à faire appel à un service tiers dont ils n'ont pas la maîtrise [75]. Des technologies comme l'API WebRTC [26] permettent par exemple à deux applications web de communiquer directement à travers un lien pair-à-pair. Les données peuvent alors être stockées localement ce qui favorise la préservation de la vie privée. Lorsque le nombre de collaborateurs devient très important, il est possible de reposer sur d'autres topologies qui visent à réduire le nombre de connexions entre les pairs tout en assurant la délivrance des messages à terme [73].

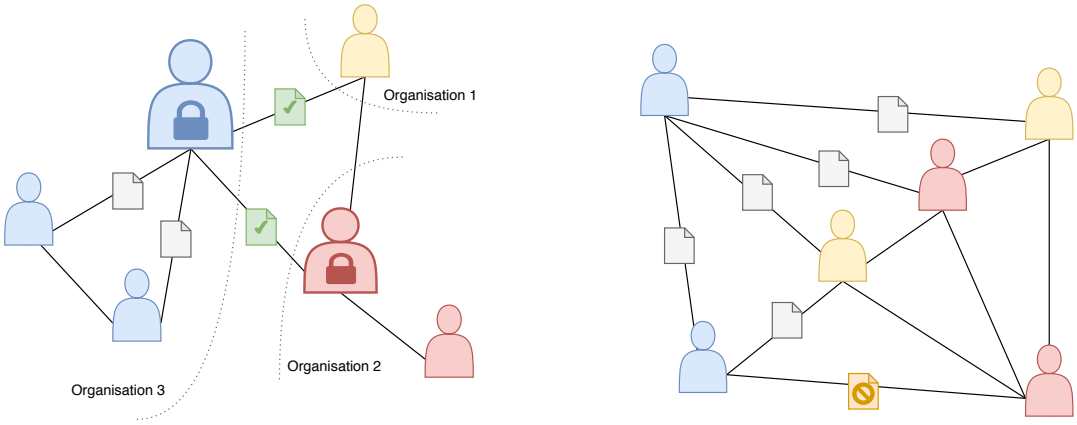
Les applications collaboratives pair-à-pair ouvrent donc la voie vers de nouveaux usages et surtout de nouvelles façons de communiquer plus proches des liens sociaux tissés par les individus.

Le cas d'usage le plus commun implique des collaborateurs appartenant à la même organisation, ou du moins au même cercle de confiance. Dans ce cas de figure, nul besoin d'appliquer une quelconque forme de contrôle sur les échanges. Les collaborateurs sont libres d'établir des connexions comme ils l'entendent et d'échanger à leur guise. Qu'en est-il lorsque plusieurs organisations collaborent entre elles ?

Dans certaines situations, il arrive que les membres de plusieurs organisations soient amenés à travailler ensemble pour répondre à des problématiques complexes ou de grande envergure. Prenons comme premier exemple, le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) [20] qui réunit des experts et chercheurs du monde entier (195 pays) pour établir un consensus scientifique sur l'impacte du réchauffement climatique dû aux activités humaines. À une échelle un peu moins importante, différents acteurs s'investissent pour mettre en place des initiatives de conservation de l'environnement pour des zones géographiques vastes telles que "The Yellowstone to Yukon Conservation Initiative (Y2Y)" ou la réserve Fitz-Stirling [39] (au sud-ouest de l'Australie). D'autres initiatives réunissant des institutions publiques et des acteurs économiques cherchent quant à elles à accroître le développement économique d'une région [62]. L'affaire des "Panama Papers" [52] qui a éclaté au grand jour en avril 2016 est le fruit du travail conjoint de nombreux journalistes à travers le monde. Plus d'une centaine de médias appartenant au consortium international des journalistes d'investigation se sont coordonnés pour traiter le grand volume de données qui a fuité (2,6 Téraoctets). Au-delà de cet effort de coordination, les journalistes impliqués ont dû prendre beaucoup de précautions pour ne pas compromettre leur source ni éveiller les soupçons des personnes incriminées. Enfin, dans un tout autre registre, les entreprises construisent souvent des partenariats avec d'autres acteurs ayant des compétences ou expertises complémentaires pour répondre à des projets de grande ampleur. En somme, dans tous ces exemples, des individus appartenant à différentes organisations collaborent pour à terme, atteindre un objectif commun. Nous emploierons ici le terme de *collaboration inter-organisationnelle* pour désigner ces différentes situations.

1.2 Motivations

Les collaborations inter-organisationnelles ont été le point de départ de ce travail de thèse. Ce terme “collaboration inter-organisationnelle” désigne très largement une situation où plusieurs groupes d'individus échangent et coproduisent des informations et des données. À l'origine, nous avions envisagé comme cas d'usage : la fédération d'organisations. C'est-à-dire l'union d'individus appartenant à différentes organisations et qui cherchent à atteindre un objectif commun. C'est un cas d'usage qui s'applique très aisément aux entreprises et administrations et qui s'illustre parfaitement à travers des projets interentreprises et interadministrations. L'objectif est de développer une approche pair-à-pair en mesure de supporter ce type de collaboration.

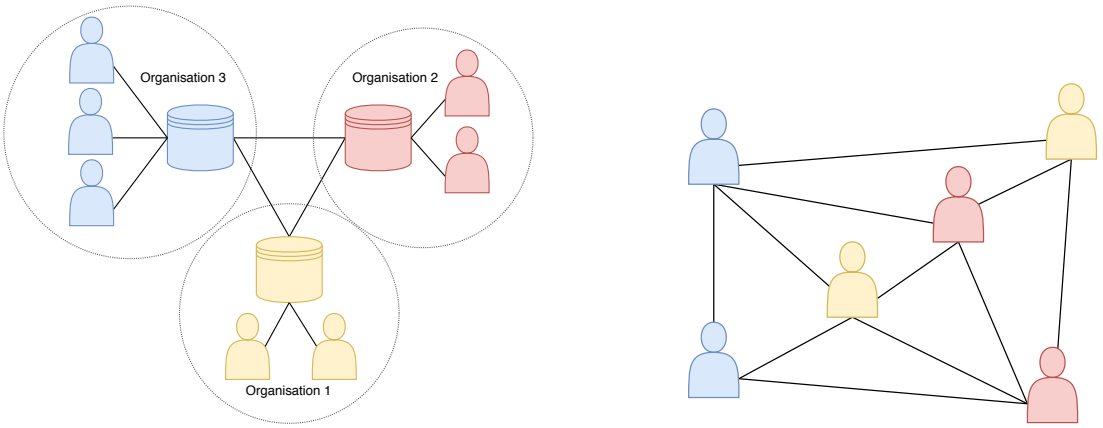


(a) (b)

FIGURE 1.1 – Scénarios de collaboration

Comme nous l'avons fait remarquer dans la présentation du contexte, les collaborations inter-organisationnelles couvrent un large spectre de collaborations. Ce sont finalement, autant de manières de collaborer, d'interagir et d'échanger de l'information. Nous pouvons facilement imaginer que les chercheurs du GIEC n'interagissent pas de la même manière avec leurs pairs que le font les journalistes de l'*International Consortium of Investigative Journalists* (ICIJ) avec leurs homologues en pleine investigation. Et ce pour de multiples raisons, les enjeux ne sont pas les mêmes, la nature du travail, la confiance et le niveau de confidentialité non plus. Ces facteurs ont un impact sur la structure sociale qui sous-tend de telles collaborations et qu'il faut prendre en compte. Les interactions entre les individus ont une place centrale dans la collaboration, puisqu'elles facilitent et conditionnent à la fois les échanges entre les parties prenantes [39, 62]. Cette pluralité des cas d'usage soulève des questions sur le partage des données et le contrôle appliqué à ces dernières. Nous pouvons envisager différents scénarios de collaboration inter-organisationnelle pour illustrer notre propos. Par exemple, dans les cas d'une collaboration interentreprises, nous supposons que chaque organisation (entreprise) doit contrôler les échanges sortants pour prévenir de toute divulgation d'informations sensibles. Ce premier exemple est illustré par la Figure 1.1(a). Nous remarquons ainsi que certains individus centralisent l'information avant de la distribuer aux autres acteurs extérieurs à l'organisation. Ces derniers appliquent donc une forme de contrôle sur les échanges inter-organisationnels. Il peuvent également être perçus

comme des points de défaillance. S'ils venaient à partir, ces individus mettraient en péril la collaboration. La collaboration inter-organisationnelle est ainsi ancrée dans une structure préétablie, reflétant les contours des organisations et pouvant aller jusqu'à s'articuler autour d'une structure hiérarchique. Ce premier scénario rejoint assez intuitivement la représentation que l'on peut se faire d'une fédération d'organisations. Inversement, nous pouvons envisager un scénario où les collaborateurs échangent directement avec leurs interlocuteurs, peu importe leur affiliation. Dans ce second cas de figure, le contrôle sur les échanges est amoindri à l'échelle des organisation et peut conduire à de la divulgation d'informations sensibles. En revanche, les collaborateurs détiennent la pleine maîtrise de leurs données et leurs échanges. Ce second scénario est illustré par la Figure 1.1(b). Il n'y a plus ici de frontières entre les organisations. Les connexions entre les parties prenantes illustrent la structure sociale établie par les collaborateurs.



(a) (b)

FIGURE 1.2 – Topologies pair-à-pair

L'agencement des interactions sociales est une source d'informations précieuse et pertinente pour déterminer quelle forme de topologie pair-à-pair serait en mesure de supporter les communications entre les collaborateurs. En fonction des configurations observées, nous pourrions même nous interroger sur la pertinence d'un modèle totalement distribué (*i.e.* pair-à-pair) pour les collaborations inter-organisationnelles. C'est par exemple le cas lorsque quelques individus jouent le rôle de régulateurs et centralisent l'information. Une première option envisageable pourrait être d'adopter une solution pair-à-pair hybride, composée de "*super-pairs*" [107]. Ces super-pairs sont des serveurs ou des pairs particuliers ayant de plus grandes capacités de calcul ou de stockage. Ils se positionnent ici à l'interface des organisations et agissent comme des passerelles vers les autres organisations. Cette topologie est présentée par la Figure 1.2(a). Le contrôle sur les échanges et la disponibilité des ressources sont délégués aux super-pairs (représentés pas les cylindres dans la figure). Tous les pairs sont directement connectés au super-pair et communiquent directement par le biais de ce dernier. Nous retrouvons ici de nombreux aspects des architectures centralisées, notamment leurs inconvénients : les super-pairs sont des points de défaillance, les utilisateurs perdent une partie de la maîtrise de leurs données ... Par ailleurs, ce type de topologie n'est pas adapté à des collaborations moins

structurées, des collaborations *ad hoc* par exemple. Une alternative est alors de considérer directement la structure sociale comme topologie. C'est une approche déjà utilisée par certaines applications pair-à-pair, notamment, certains réseaux sociaux distribués ou *Distributed Online Social Networks* (DOSNs) [23]. Les DOSNs sont des applications qui visent à reproduire les fonctionnalités des réseaux sociaux conventionnels (Facebook, Twitter ...) tout en reposant sur un paradigme de communication pair-à-pair. Dans ce contexte, la dimension sociale est également très importante. S'appuyer sur la structure sociale pour construire directement la topologie pair-à-pair permet de tirer pleinement parti de certaines propriétés comme la confiance entre les pairs [31]. En considérant que les individus sont exclusivement connectés à d'autres individus de confiance, nous pouvons aisément préserver la vie privée et limiter la fuite d'informations sensibles. Cette seconde proposition semble être plus alignée avec la philosophie pair-à-pair, qui cherche en substance à redonner le contrôle à l'utilisateur.

Nous avons finalement peu d'informations sur la manière dont les individus échangent réellement dans le cadre d'une collaboration inter-organisationnelle. Il semble difficile de faire un choix parmi les différentes hypothèses que nous avons émises, sans que celui-ci soit arbitraire. Ce qui nous conduit à considérer l'étude des intéractions sociales comme un préambule nécessaire à la mise en place d'une approche pair-à-pair pour les collaborations inter-organisationnelles.

1.3 Problématique

Nous cherchons à comprendre la manière dont les individus échangent et interagissent entre eux. Ce qui s'illustre plus formellement par la distribution des liens entre ces individus. L'objectif de ce travail de thèse est de développer une approche permettant d'analyser la structure des interactions. À partir d'exemples réels de collaborations, nous souhaitons caractériser la présence et l'agencement des liens entre les collaborateurs. Cette caractérisation doit être le point de départ qui renseigne sur la nature d'une topologie pair-à-pair à même de supporter les collaborations inter-organisationnelles.

Les exemples de collaboration que nous observons sont représentés sous la forme d'un graphe social. Les nœuds de ce graphe sont des individus et une arête relie deux individus qui ont collaboré. Nous voulons proposer une approche capable d'évaluer la propension d'occurrence des motifs d'interactions dans les graphes sociaux observés.

Cette problématique s'articule autour de trois aspects. Le premier aspect concerne la modélisation : que souhaitons-nous observer dans le graphe social ? Et comment ? L'extraction d'informations à partir d'un graphe nécessite avant toute chose de définir les motifs d'interactions que nous souhaitons mesurer, l'information elle-même en somme. Ces motifs doivent prendre en compte la dimension organisationnelle qui joue un rôle important dans le contexte d'étude. Il s'agit ici de représenter une structure organisationnelle "simple" où les individus sont tous affiliés à une et une seule organisation. Enfin, les motifs pris en compte doivent renseigner sur la nature de la topologie pair-à-pair. La problématique générale porte sur l'évaluation des propensions d'occurrence des motifs d'interactions dans les graphes et suggère l'usage d'une approche stochastique pour y répondre. De cette manière, l'occurrence d'un motif d'interaction, de même que le

graphe social observé, est vue comme la réalisation d'un processus stochastique. Il s'agit donc ici de développer un modèle stochastique intégrant la dimension organisationnelle et permettant d'évaluer la propension d'occurrence de motifs d'interaction utiles à la caractérisation des collaborations dans un contexte pair-à-pair.

Le second aspect porte sur l'estimation du modèle : à partir d'un graphe observé, comment estimer les propensions d'occurrence des motifs pris en compte par le modèle ? L'espace d'états des graphes explose à mesure que la taille du graphe augmente. Ce qui rend l'estimation directe d'un modèle stochastique (pour les graphes) difficile, voire impossible, dans de nombreux cas. Une solution envisageable est d'approcher cette estimation. L'objectif ici est de proposer une méthode d'inférence permettant d'approcher au mieux la vraie estimation et de déterminer l'incertitude autour de cette approximation.

En somme, ce travail vise à développer une approche stochastique permettant de déduire des propriétés morphologiques des graphes d'interactions.

La finalité de ce travail est d'appliquer l'approche de modélisation et d'inférence que nous avons développée dans les deux premiers points sur un exemple réel. Il existe d'ores et déjà des exemples de collaborations dites auto-organisées dans le monde réel. C'est-à-dire des collaborations implicites qui se sont organisées par la force des choses, dans un contexte donné. L'objectif recherché est d'exploiter ces exemples de collaboration à travers l'approche de modélisation et d'inférence. Ainsi, nous pouvons mieux comprendre la manière dont les individus ont échangé dans ce contexte et déduire des propriétés de la topologie. Nous envisageons aussi d'observer l'évolution des motifs d'intéractions au cours du temps en nous appuyant sur des données longitudinales.

1.4 Organisation du document

Le manuscrit est organisé de la manière suivante. Le Chapitre 2 introduit les problématiques et les enjeux liés à la modélisation structurelle de graphes sociaux. Nous y présentons une famille de modèles stochastiques : les *Exponential Random Graph Model*s (ERGMs), très largement utilisée pour ce type de problèmes. L'estimation de ces modèles pose néanmoins certaines difficultés, nous présentons certaines stratégies permettant de les contourner.

Dans le Chapitre 3, nous développons l'approche de modélisation et d'inférence que nous proposons pour répondre à la problématique. Cette approche s'inscrit dans la famille des ERGMs, mais s'inspire aussi de la modélisation de processus spatiaux. Notre modèle intègre la dimension organisationnelle en distinguant différents types de liens : les liens inter-organisationnels et les liens intra-organisationnels. Nous utilisons un algorithme d'inférence bayésienne pour répondre aux difficultés relatives à l'estimation des paramètres. Cette approche est démontrée sur des modèles simples pour valider la procédure d'estimation des paramètres.

Dans le Chapitre 4, nous mettons en œuvre l'approche sur un exemple concret de collaboration inter-organisationnelle : les collaborations initiées par les chercheurs d'un laboratoire. À partir des données bibliographiques disponibles sur l'archive ouverte HAL, nous avons pu tracer le graphe de collaborations des chercheurs qui sous-tend l'activité de publications scientifiques (Figure 4.2). Notre approche permet notamment d'identifier

la faible propension d'occurrences des liens inter-organisationnels dans ces collaborations. Ce qui peut à la fois témoigner d'une faible tendance générale à collaborer avec d'autres laboratoires, mais peut aussi révéler certaines faiblesses dans ces collaborations.

Nous montrons par ailleurs dans ce travail que notre approche, initialement dédiée à l'étude des collaborations inter-organisationnelles, peut aisément s'appliquer à d'autres contextes. Nous prenons comme second exemple dans le Chapitre 5, les interactions entre les élèves de différentes classes d'une école primaire. L'approche permet de mettre en évidence la manière dont les élèves interagissent au sein de leur classe et les élèves des autres classes. Ce qui présente notamment un intérêt pour l'étude de la propagation d'un virus en milieu scolaire. Dans ce chapitre, nous proposons également une stratégie de parallélisation de l'algorithme de simulation (l'échantillonneur de Gibbs). Cette stratégie vise à étendre la capacité de l'algorithme d'estimation des paramètres du modèle, à traiter des graphes de taille importante [2]. Nous présentons une évaluation empirique de cette stratégie de parallélisation.

Nous faisons le bilan et la synthèse des contributions proposées par ce travail de thèse dans le Chapitre 6. Nous présentons les limites de notre approche et les perspectives envisageables.

2. 5000 individus au sein d'une organisation et 5000 individus extérieurs à l'organisation

Chapitre 2

Fondements techniques

Sommaire

2.1 Modélisation structurelle des graphes et ses enjeux

Les graphes ou réseaux sont des objets mathématiques très souvent utilisés pour représenter des relations qu'elles soient entre des individus, des molécules, des astres ou tout autre agent interagissant avec d'autres agents [10]. C'est de ce fait une représentation abstraite très flexible qui s'adapte à de nombreux objets du monde réel. Depuis quelques années, l'étude structurelle de réseaux a fait l'objet de nombreux travaux de recherche notamment dans le domaine de la bio-informatique, de la cosmologie ou encore des sciences sociales et de la théorie du management [3].

Nous nous intéressons plus ici aux réseaux représentant des structures sociales que l'on nomme également réseaux sociaux (à ne pas confondre avec les applications comme Facebook) et considérons ainsi un ensemble d'individus potentiellement reliés les uns aux autres par des liens relationnels.

Nous pouvons prendre comme exemple des réseaux sociaux représentant des collaborations entre des individus. Collaborer et mettre en commun ses connaissances et ses forces s'avère être très souvent une nécessité pour répondre à des enjeux majeurs. Que ce soit pour résoudre des problématiques environnementales [12, 20, 39, 38] ou économiques [62] les parties impliquées s'allient, se structurent et s'interconnectent pour à terme mettre en œuvre des solutions. L'étude des collaborations permet entre autres choses de mieux comprendre les mécanismes favorisant les échanges entre les parties, mais aussi d'identifier les faiblesses et les dépendances qui pourraient mettre en péril de telles entreprises. Étendre la compréhension de ces réseaux de collaboration, c'est se donner la possibilité de les améliorer et de les rendre plus efficaces et résilients [77, 79].

Pour étudier ces réseaux, nous faisons appel à la théorie des graphes [103]. Nous considérons ici que les graphes constituent une représentation formelle des réseaux au sens général du terme. Les nœuds du graphe symbolisent les individus ou des groupes d'individus, et les arêtes sont des relations binaires, représentant les liens sociaux sociaux tissés entre ces derniers.

On désigne généralement un graphe $\mathcal{G}$ par la notation :

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

où $\mathcal{V}$ et $\mathcal{E}$ sont respectivement l'ensemble des nœuds et l'ensemble des arêtes. La taille du graphe correspond au nombre de nœuds qui le composent soit $N = |\mathcal{V}|$, de la même manière, nous définissons le nombre d'arêtes par $E = |\mathcal{E}|$.

En complément de cette notation, la matrice d'adjacence est très souvent utilisée pour représenter plus en détail la structure relationnelle décrite par le graphe. La matrice d'adjacence est une matrice carrée, binaire, de taille $N \times N$:

$$y = \Big[y_{i,j} \in \{0, 1\} \Big]_{0 \leq i,j < N},$$

où $y_{i,j}$ désigne l'arête entre les nœuds i et j de telle sorte que :

$$\begin{cases} y_{i,j} = 1, & \text{désigne l'existence de l'arête entre les nœuds } i \text{ et } j. \\ y_{i,j} = 0, & \text{désigne l'absence de l'arête entre les nœuds } i \text{ et } j. \end{cases}$$

Cela revient à considérer pour un graphe dirigé, l'ensemble des arêtes activées dans la matrice d'adjacence : $\mathcal{E} = \{(i,j), y_{i,j} = 1\}_{0 \leq i,j < N}$. Notons par ailleurs que pour un graphe non-dirigé la matrice d'adjacence est symétrique telle que : $y_{i,j} = y_{j,i}$. Que ce soit pour des graphes dirigés ou non dirigés, nous faisons l'hypothèse que les extrémités d'une arête sont deux nœuds distincts, nous considérons donc que la diagonale de la matrice d'adjacence est nulle :

$$y = \begin{pmatrix} 0 & y_{1,2} & \cdots & y_{1,N-1} \\ y_{2,1} & 0 & \cdots & y_{2,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N-1,1} & y_{N-1,2} & \cdots & 0 \end{pmatrix}.$$

Les graphes illustrant des comportements sociaux réels peuvent être des sources d'informations riches pour mieux comprendre les tendances comportementales individuelles ou collectives qui sous-tendent la création des liens. Des approches descriptives consistant à mesurer certaines métriques comme les mesures de *centralité*, la *densité*, le *rayon*, le *degré moyen* etc. permettent de quantifier certains phénomènes sociaux [18, 103]. Elles sont cependant limitées lorsqu'il s'agit d'identifier des processus complexes à l'origine même de la formation du graphe social. La création d'un lien n'est généralement pas due au pur hasard, mais reflète des tendances comportementales complexes liées notamment à la présence d'autres liens sociaux dans le graphe. De plus, les tendances comportementales ne sont pas invariantes. Si nous observons par exemple chez un grand nombre de sujets une propension à tisser des liens réciproques avec autrui, cette tendance ne se vérifiera pas pour tous les sujets. Cette composante "chaotique" dans le processus de formation des liens [105] motive l'adoption d'une autre approche considérant le graphe observé comme une réalisation d'un processus stochastique. Le graphe observé est un *graphe aléatoire* qui représente donc une configuration particulière d'arêtes parmi l'ensemble des graphes possibles. Chaque arête possible du graphe est une variable aléatoire binaire, notée $Y_{i,j}$ dont la réalisation $Y_{i,j} = y_{i,j}$ résulte d'un processus stochastique.

Les processus conduisant à la formation de liens dans le graphe peuvent être caractérisés par des structures locales. Prenons par exemple le cas illustré par la Figure 2.1. Deux individus, Alice et Charlie entretiennent tous les deux un lien d'amitié (matérialisé par les deux traits pleins) avec Bob. C'est une configuration particulière qui du fait que Alice et Charlie partagent tous les deux une connaissance commune (Bob), peut favoriser la création d'un lien d'amitié entre ces deux individus (matérialisé par un trait en pointillé). Cette tendance illustre en d'autres termes l'adage : "les amis de mes amis sont mes amis" [92]. L'apparition d'un lien entre Alice et Bob fait alors émerger une structure en "triangle". La prépondérance de structures en triangle dans un graphe social pourrait alors renseigner sur la généralisation de cette tendance comportementale à une population donnée.

Bien que dans les faits ce type de comportement puisse faire émerger des structures triangulaires, l'approche adoptée reste toutefois peu réaliste pour l'étude globale d'un graphe. D'une part, parce que la présence de triangles peut être due à d'autres facteurs comme une simple inclination à créer un grand nombre de liens avec d'autres individus. D'autre part parce que les processus sociaux influençant l'émergence des liens, sont très souvent couplés entre eux et imbriqués. Il n'est donc pas réaliste de les

étudier indépendamment. La modélisation structurelle des graphes apparaît donc comme complexe.

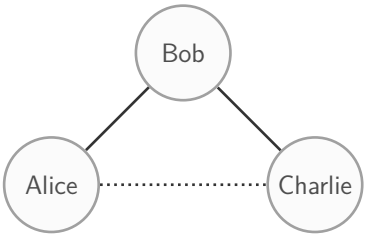


FIGURE 2.1 – Graphe représentant des liens d'amitié entre Alice, Bob et Charlie. *(Les traits pleins représentent les liens existants, là où les pointillés symbolisent les liens hypothétiques)*

2.2 Les Exponential Random Graph Models

2.2.1 Forme générale du modèle

Les *Exponential Random Graph Model*s sont une famille de modèles très souvent utilisés pour la modélisation structurelle des graphes. C'est une classe de modèles pertinente au regard des remarques précédemment énoncées. Les ERGMs s'inscrivent en effet dans le cadre conceptuel selon lequel le graphe observé est le produit de l'émergence de structures locales comme les triangles vus précédemment. Comme souligné par WASSERMAN et PATTISON [104], cette famille peut inclure une variété importante de structures locales et permet ainsi de tester un grand nombre d'hypothèses sur les processus sociaux sous-jacents. De plus, autour de cette famille de modèles s'est développé un ensemble d'outils [43, 50, 64] popularisant les ERGMs au-delà des communautés de mathématiques appliquées et de statistiques spatiales.

La forme générale de ce type de modèle s'exprime sous la forme d'une probabilité d'existence d'une réalisation d'un graphe aléatoire y conditionnellement à l'ensemble des paramètres θ régissant ce modèle :

$$p(Y = y|\theta) = \frac{\exp(U(y|\theta))}{\kappa(\theta)} = \frac{\exp(\langle\theta, t(y)\rangle)}{\kappa(\theta)} = \frac{\exp(\sum \theta_i t_i(y))}{\kappa(\theta)}, \tag{2.1}$$

où la fonction U se nomme la fonction d'énergie et $\langle\cdot,\cdot\rangle$ est le produit scalaire. Le graphe observé est décrit au travers du vecteur des statistiques suffisantes $t(y)$. À chaque configuration (ou structure locale) considérée est associée une statistique $t_i(y)$ qui donne le nombre de fois où elle apparaît dans le graphe. Chaque statistique est pondérée par un paramètre $\theta_i \in \theta$. L'interprétation des résultats peut s'apparenter à celle d'une régression linéaire. Les statistiques peuvent être vues comme des variables explicatives et les paramètres qui les pondèrent renseignent sur leur présence dans le graphe observé. Si le paramètre θ_i est positif, cela implique que la structure locale comptabilisée par la statistique $t_i(y)$ a plus de chance de se réaliser que par pur hasard. À l'inverse, un paramètre négatif informe sur le fait que la structure locale associée à moins de chance

d'exister que par pur hasard. En substance, le modèle (et ses paramètres) renseigne donc sur la présence de structures locales dans un graphe et donc de l'impact des processus sociaux sous-jacents qui ont engendré ce graphe. Il reste à présent à définir quelles sont les structures locales que nous pouvons prendre en compte dans le modèle et ce qu'elles représentent du point de vue la modélisation des interactions. Ces points sont développés dans la Section 2.2.2.

Une constante de normalisation $\kappa(\theta)$ est placée au dénominateur (2.1) pour garantir que la valeur de la probabilité $P(Y = y|\theta)$ est inférieure à 1. Cette constante de normalisation représente toutes les configurations possibles du graphe pour une taille fixée et peut s'exprimer ainsi :

$$\kappa(\theta) = \sum_{y \in \mathcal{Y}} \exp(U(y|\theta)), \tag{2.2}$$

où $\mathcal{Y}$ décrit l'espace des configurations. Pour un graphe y de taille définie, l'espace des configurations $\mathcal{Y}$ comprend tous les graphes de même taille, qu'il est possible d'engendrer : du graphe sans aucune arête au graphe complet et toutes les configurations intermédiaires possibles. La constante $\kappa(\theta)$ est numériquement difficile voire impossible à calculer du fait de la taille de $\mathcal{Y}$, et ce, même pour des graphes de taille raisonnable. Le nombre de configurations d'arêtes pour un graphe non dirigé de taille n est $2^{\binom{n}{2}}$ ce qui représente pour un graphe de 10 nœuds (un graphe de taille très modeste) plus de 10^{13} configurations possibles, un ordre de grandeur comparable au nombre de galaxies recensées dans l'univers (entre 1000 et 3000 milliards [19]).

2.2.2 Des hypothèses de dépendance à la configuration du modèle

Comme nous l'avons souligné précédemment, la création d'un lien entre deux individus n'est généralement pas due au hasard et peut dépendre de la présence d'autres liens dans le graphe. Ces dépendances reflètent les tendances comportementales que nous souhaitons modéliser. Une fois les hypothèses de dépendances formulées, nous pouvons définir les structures locales qui composent notre modèle. Les structures locales sont en effet les configurations d'arêtes qui sont interdépendantes au regard des hypothèses considérées.

En termes généraux, l'apparition d'un lien dans un graphe est par conséquent conditionnée par toutes les arêtes du reste du graphe. D'une manière plus formelle nous pouvons exprimer cette probabilité de réalisation d'une arête (i, j) ainsi :

$$P(Y_{i,j} = y_{i,j}|Y_{-i,j} = y_{-i,j}, \theta), \tag{2.3}$$

où $y_{-i,j}$ représente l'ensemble des autres arêtes observées. Il n'est bien sûr pas réaliste de conditionner l'apparition d'une arête au reste du graphe, cela reviendrait à prendre en compte un trop grand nombre de structures locales. Nous ferions face aux mêmes limites combinatoires que celles observées pour le calcul de la constante de normalisation. Nous devons donc formuler des hypothèses de dépendances cohérentes et réalistes au regard des processus sociaux étudiés, mais aussi assez raisonnables pour que le nombre de configurations à considérer ne soit pas trop important.

Plus il y a de dépendances entre les liens, plus la probabilité de réalisation d'un graphe sera difficile à calculer. Cela se vérifie même pour deux arêtes. Si deux arêtes distinctes sont indépendantes, cela signifie que l'apparition de l'une ne dépend aucunement de l'état de l'autre arête. En d'autres termes, la probabilité que ces deux arêtes existent (probabilité jointe) se réduit au produit des deux probabilités d'existence. Dans l'autre cas, la probabilité jointe et plus difficile à calculer et dépend des hypothèses de dépendances formulées.

Le choix des hypothèses de dépendances est une étape préliminaire cruciale puisqu'elle permet de faire le lien entre le réel, les comportements et phénomènes sociaux étudiés et la modélisation, les structures locales comptabilisées par les statistiques suffisantes (2.1). Nous présentons ici les hypothèses de dépendance les plus connues et les configurations de modèles associées.

Hypothèse de Bernoulli

Chronologiquement la première hypothèse de dépendance est l'hypothèse de Bernoulli. C'est en réalité une hypothèse d'indépendance, puisque cette hypothèse considère les arêtes comme des variables aléatoires indépendantes et identiquement distribuées dans le graphe aléatoire. À la manière d'un lancé de pièce, l'existence d'une arête dépend uniquement d'une probabilité $p(Y_{i,j} = 1|\theta)$. Les modèles adoptant cette hypothèse se rapprochent d'ailleurs du modèle de Erdős–Rényi [25] qui, de manière identique, construit le graphe en sélectionnant uniformément et avec une probabilité p définie l'ensemble des arêtes existantes.

La forme générale du modèle de Bernoulli revient à considérer uniquement comme structure locale les arêtes elles-mêmes et peut s'écrire de cette manière :

$$p(Y = y|\theta) = \frac{\exp(\theta_L L(y))}{\kappa(\theta)}. \tag{2.4}$$

L'unique statistique suffisante $L(y)$ donne le nombre d'arêtes dans le graphe et s'exprime pour un graphe non dirigé comme :

$$L(y) = \sum_{i<j \in \mathcal{V}} y_{i,j}. \tag{2.5}$$

Dans le cas d'un graphe dirigé cela revient à considérer toutes les variables aléatoires $y_{i,j}$, sans la condition $i < j$. Le paramètre associé θ_L pondère la présence d'arête dans le graphe de telle sorte que la probabilité d'existence d'une arête respecte la relation :

$$P(Y_{i,j} = 1|Y_{-i,j} = y_{-i,j}, \theta_l) = P(Y_{i,j} = 1|\theta_L) = \frac{e^{\theta_L}}{e^{\theta_L} + 1}.$$

Il est à noter que la probabilité conditionnelle d'émergence d'une arête (i, j) décrite par (2.3) est simplifiée de manière conséquente puisqu'elle ne dépend aucunement des autres arêtes du graphe. Ce qui fait que le modèle de Bernoulli est la configuration la plus simple des ERGMs, bien que peu réaliste du fait de l'absence de dépendance.

Hypothèse "Dyad-independent"

L'hypothèse "Dyad-independent" est une extension de l'hypothèse de Bernoulli et considère pour des graphes dirigés que l'émergence d'une arrête (i, j) est dépendante de l'arête (j, i). Ce type d'interactions que l'on nomme aussi dyade (ou interaction dyadique) représente des relations réciproques entre deux individus (i et j en l'occurrence). La dépendance se limite en revanche à la dyade, deux dyades distinctes (i, j) et (k, l) sont donc indépendantes l'une de l'autre.

Le *modèle "Dyad-independent"* proposé par Holland et Leinhardt [46] est le premier qui repose réellement sur une hypothèse de dépendance. Il s'écrit sous la forme :

$$p(Y = y|\theta) = \frac{\exp(\theta_L L(y) + \theta_M M(y))}{\kappa(\theta)}. \tag{2.6}$$

S'ajoute alors à (2.4) la statistique suffisante $M(y)$ qui compte le nombre de dyades présentes dans le graphe de telle sorte que :

$$M(y) = \sum_{i<j} y_{i,j} y_{j,i}.$$

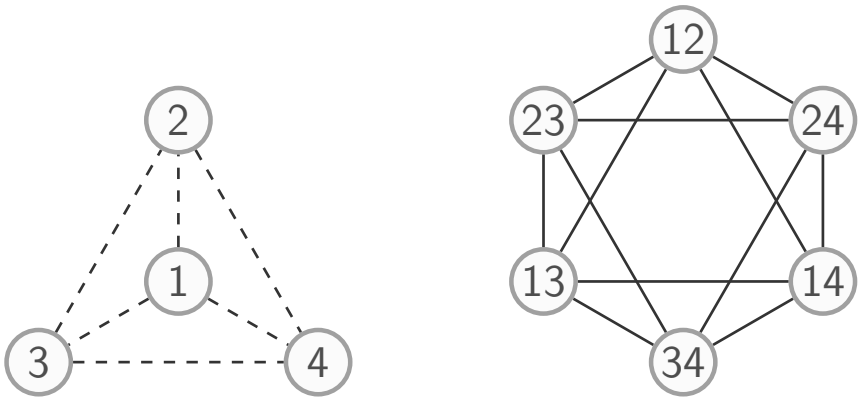
Ce modèle que l'on nomme également *modèle p1* permet à la fois de quantifier la propension pour un individu d'être émetteur ou récepteur, mais aussi la tendance pour deux individus à tisser des liens réciproques. En outre, d'autres modèles reposent sur cette hypothèse d'indépendance entre les dyades, comme *le modèle p2* [24] qui considère de surcroît les attributs des nœuds comme des covariables. Cela permet ainsi d'introduire des facteurs exogènes (comme l'âge des individus, leur genre ...) dans le processus de création des liens.

Hypothèse de Markov

Jusqu'à présent, les hypothèses de dépendances présentées sont limitées et ne permettent pas de modéliser des comportements comme celui présenté par la figure 2.1. Pour rappel, cet exemple illustre une tendance comportementale selon laquelle le fait de partager une connaissance commune (Bob) peut favoriser la création d'un lien entre deux personnes (Alice et Charlie). C'est un motif comportemental assez récurrent qui nécessite cependant des hypothèses de dépendance plus fortes.

Dans ce cas de figure, la réalisation de l'arête (Alice, Charlie) dépend des arêtes (Alice, Bob) et (Bob, Charlie). Nous entrevoyons ici une relation de dépendance selon laquelle les arêtes qui sont "reliées" par un même nœud sont dépendantes les unes des autres. Toutes les arêtes qui sont connectées par un même nœud sont définies comme voisines les unes des autres. Frank et Strauss [28] proposent l'hypothèse de dépendance de Markov qui étend la dépendance des réalisations des arêtes à leur *voisinage*.

Les notions de voisinage et de dépendance dans ce voisinage peuvent être généralisées de la manière suivante. La réalisation de deux arêtes (i, j) et (i, k), qui partagent un nœud commun i sont dépendantes l'une de l'autre. Inversement, deux arêtes (i, j) et (k, l) telles que i, j, k et l sont des nœuds distincts, sont toutes les deux indépendantes. Nous définissons le voisinage tel que deux arêtes sont dites voisines l'une de l'autre si elles partagent un nœud commun.



(a) Graphe aléatoire original G (b) Graphe de dépendance D

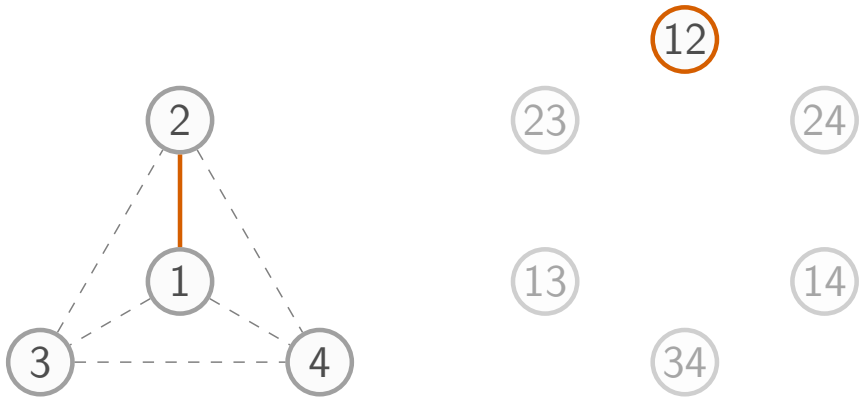
FIGURE 2.2 – Graphe original et son graphe de dépendance tels qu'illustrés dans [28]

On remarque ici que la probabilité conditionnelle de réalisation d'une arête (2.3) se réduit donc au voisinage de cette dernière :

$$P(Y_{i,j} = y_{i,j}|Y_{-i,j} = y_{-i,j}, \theta) = P(Y_{i,j} = y_{i,j}|Y_{V_{i,j}} = y_{V_{i,j}}, \theta),$$

où $V_{i,j}$ dénote ici le voisinage de l'arête (i, j).

Ces interdépendances entre les arêtes ne sont pas facilement identifiables dans un graphe. FRANK et STRAUSS [28] proposent une vue alternative du graphe, que l'on nomme ici *graphe de dépendances* et qui représente les interactions entre les arêtes du graphe. Ainsi, chaque nœud du graphe de dépendances correspond à une arrête du graphe observé. Les nœuds du graphe de dépendances sont connectés en fonction des hypothèses de dépendances formulées. Nous considérons pour les dépendances de Markov que deux arêtes sont connectées si l'une appartient au voisinage de l'autre. Prenons comme exemple la Figure Fig. 2.2 qui représente à gauche un graphe aléatoire, notre graphe observé (Figure 2.2(a)) et à droite le graphe de dépendance associé (Figure 2.2(b)). Le graphe G est un graphe non dirigé qui comprend quatre nœuds et dont l'ensemble des arêtes possibles sont représentées en pointillé. À droite, chaque nœud du graphe de dépendance D représente une variable aléatoire associée à une arête du graphe G. Il y a autant de nœuds dans le graphe de dépendances qu'il y a d'arêtes possibles dans le graphe G, soit $\binom{n}{2}$ nœuds pour un graphe non dirigé de taille n (et $n(n-1)$ nœuds pour un graphe dirigé). Le nœud 12 du graphe de dépendance est donc associé à l'arête $(1, 2)$ du graphe aléatoire observé. D'après la définition du voisinage que nous avons donnée, deux arêtes sont connectées si elles partagent un nœud commun. Remarquons par exemple dans le graphe de dépendance que les nœuds 12 et 24 sont connectés puisque les arêtes correspondantes dans le graphe aléatoire partagent toutes les deux le nœud 2. Inversement, les nœuds 12 et 34 du graphe de dépendance D ne sont pas connectés puisque les arêtes $(1, 2)$ et $(3, 4)$ ne partagent aucun nœud commun dans le graphe G. Nous constatons ainsi que le graphe de dépendance n'est pas un graphe complet, même si toutes les arêtes du graphe aléatoires sont considérées. Sans cette propriété importante, il n'y aurait aucun intérêt à nous reposer sur les hypothèses de Markov puisque cela impliquerait que la réalisation d'une arête est conditionnellement dépendante à toutes les autres.



(a) Graphe aléatoire original G

(b) Graphe de dépendance de Bernoulli D

FIGURE 2.3 – Graphe de dépendance de Bernoulli

Il est intéressant de remarquer que toutes les arêtes conditionnellement dépendantes les unes des autres dans le graphe G, sont toutes interconnectées dans le graphe de dépendances D. Elles font apparaître des sous-ensembles complets que l'on nomme cliques. Autrement dit, ces arêtes interdépendantes dans le graphe G forment des cliques dans le graphe de dépendances. Par ailleurs, les structures locales comptabilisées par les statistiques suffisantes sont composées uniquement d'arêtes dépendantes les unes des autres. C'est ce que nous avons observé dans une moindre mesure avec les hypothèses de Bernoulli et "Dyad independent". De ces deux constats, nous esquissons un lien entre les cliques présentes dans le graphe de dépendances et les formes des structures locales du modèles : les cliques définissent les structures locales considérées par le modèle. Prenons un exemple simple, de l'hypothèse de Bernoulli et le modèle associé. Dans le cadre des hypothèses de Bernoulli il n'y a aucune dépendance entre les arêtes, ce qui implique que le graphe de dépendance est uniquement composé de nœuds sans aucune connexion, comme le montre la Figure 2.3. Du fait de l'absence de liens dans le graphe D, les seules cliques présentes sont les nœuds eux-même, c'est-à-dire, les arêtes dans le graphe G. Pour illustrer nos propos, nous avons entouré une des cliques (le nœud 12) et l'arête correspondante dans le graphe G. Nous retrouvons ainsi la seule structure locale considérée par le modèle de Bernoulli, l'arête elle-même. Ce raisonnement peut-être étendu à l'hypothèse "Dyad-dependent" (et au modèle $p1$), il faudra alors considérer les arêtes et leurs réciproques (par exemple l'arête 12 et l'arrête 21). Les arêtes et leurs réciproques sont connectés dans le graphe de dépendances et forment donc des cliques qui correspondent au interactions dyadiques dans le graphe G.

Nous venons ici de mettre en exergue d'une manière intuitive le lien direct entre les hypothèses de dépendances et les structures locales par le biais du graphe de dépendances. Qu'en est-il pour les hypothèses de Markov et son modèle ? Cette intuition se généralise pour le modèle de Markov grâce au théorème d'Hammersley-Clifford introduit par BESAG [8] et repris par FRANK et STRAUSS [28] pour la modélisation des graphes. Ce théorème définit la probabilité d'un graphe de la manière suivante. Pour tout graphe $y \in Y$ respectant les propriétés de Markov, la probabilité d'un graphe p, telle que $p(y) > 0$,

s'écrit sous la forme d'une *distribution de Gibbs* :

$$\begin{aligned} p(y) &= \kappa^{-1} \exp\left(\sum_{A \subseteq J^*} \theta_A \Pi_{\{i,j\} \in A} y_{i,j}\right) \\ &= \kappa^{-1} \exp\left(\sum_{A \subseteq J^*} \theta_A t_a(y)\right), \end{aligned} \tag{2.7}$$

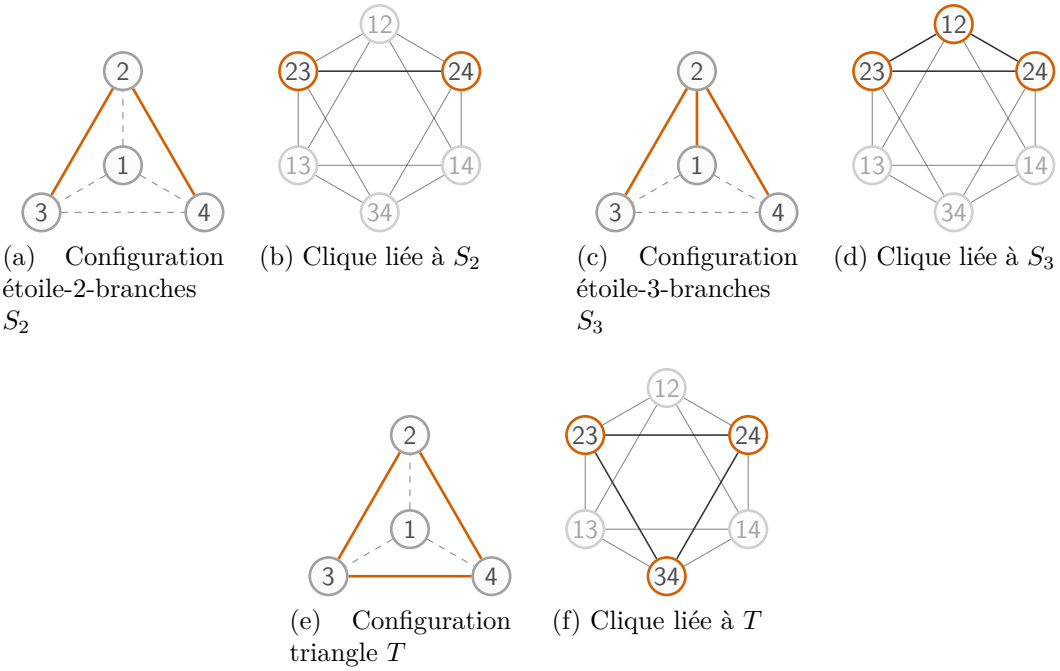
où J^* correspond à l'ensemble des cliques du graphe de dépendance, A est une clique (une configuration), θ_A est le paramètre associé à cette clique et t_A est une fonction indicatrice qui renvoie 1 si la configuration A existe et 0 sinon. La forme de cette probabilité est semblable à la définition générale des ERGMs décrite par l'Équation (2.1). Le différence principale réside dans le fait que la distribution de Gibbs introduite par le théorème d'Hammersley-Clifford [41] ne s'appuie pas sur les statistiques suffisantes mais prend en compte toutes les configurations possibles dans le graphe considéré.

Nous pouvons identifier les structures locales du modèle de Markov en reprenant la démarche qui nous a permis de reconnaître les cliques pour le modèle de Bernoulli. La Figure 2.4 présente les différentes configurations du modèles de Markov et les cliques associées dans le graphe de dépendances. Comme pour le modèle de Bernoulli, nous présentons ici les différents types de configuration possibles sans pour autant énumérer de manière exhaustive toutes les cliques du graphe de dépendances. D'autres cliques sont donc identifiables, mais les configurations correspondantes sont équivoques à celles décrites par la Figure 2.4. Par ailleurs, la configuration "singleton" du modèle de Bernoulli est aussi une configuration du modèle de Markov, un nœud étant quoi qu'il soit une clique à part entière, la plus petite clique possible (Figure 2.3).

Les premières configurations présentées par la Figure 2.4 sont les configurations dites en étoiles, ou "*k-star*" dans la littérature anglophone (avec k le nombre de branches de l'étoile). Elles illustrent un phénomène de centralisation où un acteur entretien plusieurs relations avec ses alter-egos. Dans notre exemple, seules les étoiles à deux et trois branches existent, elles sont représentées respectivement par les Figures 2.4(a) et 2.4(c). Cette limite s'explique par le fait qu'une *clique maximale* du graphe D correspond à une étoile à trois branches dans le graphe G (ou à une configuration triangulaire, nous y reviendrons par la suite). Pour un graphe dirigé de taille n, la configuration en étoile peut au maximum concentrer $n - 1$ connexions. Notons par ailleurs qu'une étoile à une branche ("1-star") correspond ni plus ni moins qu'à l'unique configuration du modèle de Bernoulli, l'arête elle-même.

La seconde configuration identifiée (Figure 2.4(e)) est la structure en triangle. Ce motif structurel a déjà été évoqué dans notre exemple introductif impliquant deux individus (Alice et Charlie) qui ont une connaissance commune (Bob). Ce type de configuration permet de modéliser des relations transitives ou plus généralement des phénomènes de regroupement (*clustering*). Tout comme la configuration *"3-star"*, la clique associée à la configuration triangle dans le graphe D est une clique maximale.

Les configurations d'ordre supérieur comme les triangles ou les "3-star" contiennent des configurations d'ordre inférieur (comme les "2-star" ou les arêtes). Des configurations triangle sont composées par exemple de configurations "2-star" et d'arêtes. Cette



(a) Configuration étoile-2-branches S_2

(b) Clique liée à S_2

(c) Configuration étoile-3-branches S_3

(d) Clique liée à S_3

(e) Configuration triangle T

(f) Clique liée à T

FIGURE 2.4 – Identification des cliques et des configurations locales associées

organisation hiérarchique des motifs structurels offre la possibilité d'évaluer la prépondérance d'un motif d'ordre supérieur par rapport à des interactions d'ordre inférieur. En considérant uniquement les arêtes et les triangles, nous pouvons par exemple quantifier la prédominance de triangles dans un graphe sachant le nombre d'arêtes. KOSKINEN et DARAGANOVA [56] montrent que pour un graphe non dirigé de 20 nœuds, le nombre de configurations possibles (sous les hypothèses de Markov) s'élève à plus de 10 millions. Soit aussi plus de 10 millions de paramètres à évaluer dans la distribution (2.7) introduite par le théorème d'Hamersley-Clifford. En réalité, il n'est pas nécessaire de considérer séparément des structures de même nature. Deux configurations triangle considérées par la distribution (2.7) ont finalement la même probabilité d'existence. Nous faisons ainsi l'hypothèse que les probabilités d'existence de structures de même nature sont *homogènes*. Sous cette hypothèse, un paramètre n'est plus associé à une configuration unique, mais à un ensemble de configurations analogues. Ce qui permet de réduire considérablement le nombre de paramètres de la distribution (2.7) et par la même occasion de retrouver la forme générale des ERGMs définie par (2.1).

Ce qui nous permet de définir le modèle de Markov ainsi :

$$p(Y = y|\theta) = \frac{\exp(\theta_L L(y) + \theta_{S_2} S_2(y) + \theta_{S_3} S_3(y) + \cdots + \theta_{S_k} S_k(y) + \theta_T T(y))}{\kappa(\theta)}, \tag{2.8}$$

où la statistique pour une configuration *"1-star"* est identique à la statistique $L(y)$ (décrite par (2.5)), pour la configuration *"2-star"* elle se définit ainsi :

$$S_2(y) = \sum_{i<j} \sum_{k \neq i,j} y_{ij} y_{ik}.$$

Frank et Strauss [28] mettent en évidence le lien entre la configuration en étoile et le degré du nœud au centre de l'étoile et donnent la définition d'une statistique "k-star" (pour un graphe non-dirigé) pour $k \geq 2$:

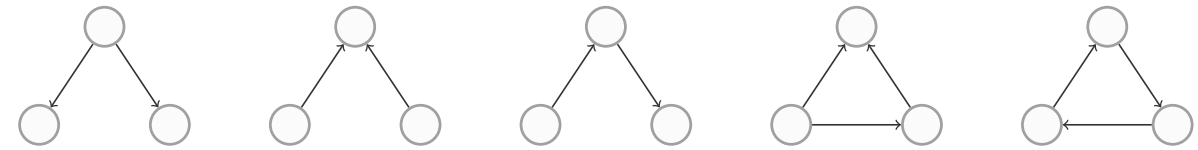
$$S_k(y) = \sum_{j \geq k} \binom{j}{k} d_j(y), \tag{2.9}$$

où $d_j(y)$ est une fonction qui donne le nombre de nœuds de degré j dans le graphe y.

Enfin, la statistique T s'exprime ainsi :

$$T(y) = \sum_{i<j<k} y_{ij} y_{kj} y_{ki}.$$

Pour un graphe dirigé de même taille le graphe de dépendance est plus conséquent puisque nous distinguons les arêtes (i, j) et les arêtes (j, i) de sens opposé. Le nombre de configurations possibles est donc plus important pour un même nombre de nœuds.



(a) "out-2-star" (b) "in-2-star" (c) "2-path" (d) Triangle transitif (e) 3-cycle

Figure 2.5 – Les configurations "2-star" et triangle pour un graphe non dirigé peuvent se décliner en plusieurs configurations pour un graphe dirigé.

Comme illustré par la Figure 2.5, les configurations "2-star" et triangle que nous avons définie pour des graphes dirigés se déclinent en plusieurs configurations d'arêtes dirigées. Ce qui nous permet par exemple de considérer des phénomènes de popularité (Figure 2.5(a)), de concentration (Figure 2.5(b)), de relais (Figure 2.5(c)) ou encore des relations transitives (Figure 2.5(d)) et cycliques (Figure 2.5(e)).

Le modèle de Markov permet donc de prendre en compte des relations plus complexes et réalistes (comme la transitivité par exemple) avec des dépendances plus fortes que celles des modèles de Bernoulli et "Dyad-independent". Cependant, Handcock [42] et Snijders [89] soulignent le fait que le modèle de Markov a tendance à produire des graphes peu vraisemblables dans le monde réel, comme des graphes (presque) complets ou (presque) vides. Cela signifie qu'un grand nombre de configurations de paramètre θ a tendance à produire des graphes complets ou vides. Handcock [42] montre ainsi que l'espace des paramètres permettant de générer des graphes réalistes est restreint comparé à l'espace de départ. Nous parlons ainsi de *dégénérescence* ou *quasi-dégénérescence* du modèle. Plus concrètement, cela signifie que le modèle de Markov n'est pas en mesure de reproduire n'importe quel graphe observé. Les méthodes de simulation que nous verrons dans la sous-section 2.3.1 sont sensibles à ce problème. Par exemple si un modèle pénalise les arêtes ($\theta_L < 0$) et encourage à la fois la production de triangle ($\theta_T > 0$), alors la chaîne de graphes simulés peut suivre un régime bimodal selon les termes de Snijders [89],

c'est-à-dire alterner entre des graphes fortement et faiblement denses. Ce problème peut-être mitiger en modifiant quelque peu la stratégie de simulation. Nous évoquerons plus en détails ces variantes dans la sous-section 2.3.1. De plus, une faible modification d'un paramètre peut entraîner un changement très important des statistiques suffisantes. Cela est d'autant plus flagrant pour des paramètres contrôlant des motifs de degré important, comme les triangles ou les étoile à n branches ($n \geq 2$). Par exemple, si nous augmentons légèrement le paramètre θ_T pour un modèle de Markov donné (2.8), nous observerons une forte augmentation du nombre de triangles ($T(y)$), mais aussi du nombre d'arêtes ($L(y)$) [57]. Le modèle souffre donc de surcroît d'instabilité. L'espace des paramètres produisant un graphe "réaliste" étant restreint, une modification importante d'un ou plusieurs paramètres peut rapidement conduire le modèle à produire des graphes vides ou complets. Ce problème de dégénérescence ne touche pas que les ERGMs, mais affecte aussi plus généralement des modèles d'interaction, comme le modèle de Ising [8, 97].

Enfin, Snijders *et al.* [92] proposent une autre statistique qui repose sur la distribution des degrés dans le graphe. Comme nous l'avons évoqué avec (2.9), les statistiques "k-star" peuvent s'écrire en fonction de la distributions des degrés avec $d_j(y)$, le nombre de nœuds de degré j dans le graphe y. La statistique proposée s'appuie toujours sur les hypothèses de Markov. Nous la désignons par le terme *Geometrically Weighted Degree* (GWD) et elle se définit de cette manière :

$$u_\alpha^{(d)}(y) = \sum_{k=0}^{n-1} e^{-\alpha k} d_k(y). \tag{2.10}$$

Elle est contrôlée par le paramètre θ_d ($\theta_d \times u_\alpha^{(d)}(y)$) de telle sorte qu'il pondère les degrés $d_j(y)$ pour $j = 0 \dots n-1$ suivant cette relation : $\theta_{d_j} = \theta_{d_{j-1}} e^{-\alpha}$. Les degrés de valeur élevée sont donc de plus en plus atténués. Ce qui mitige par conséquent l'effet "boule de neige" que nous avons décrit au-dessus, qui consiste pour un paramètre positif contrôlant une statistique de degré important à favoriser la production de graphes complets. La statistique $u_\alpha(y)$ peut s'écrire de manière équivalente comme fonction des statistiques "k-star" :

$$\begin{aligned} u_\lambda^{(s)}(y) &= S_2(y) - \frac{S_3(y)}{\lambda} + \frac{S_4(y)}{\lambda^2} - \dots + (-1)^{n-1} \frac{S_{n-1}(y)}{\lambda^{n-3}}, \\ &= \sum_{k=2}^{n-1} (-1)^k \frac{S_k(y)}{\lambda^{k-2}}, \end{aligned} \tag{2.11}$$

avec $\lambda = e^\alpha/(e^\alpha - 1)$. Notons une légère différence avec (2.10) : la statistique *"alternating-star"* $u_\lambda^s(y)$ (2.11) ne prend pas en compte les arêtes (ou étoiles S_1) contrairement à la statistique GWD. Le phénomène d'atténuation est quant à lui assuré, par l'alternance des statistiques positives et négatives. La statistique $u_\lambda^{(s)}(y)$ est contrôlée par le paramètre θ_{S_k} tel que : $\theta_{S_k} = -(1/\lambda)\theta_{S_{k-1}}$. Le paramètre λ (respectivement α) peut-être soit fixé arbitrairement, soit considéré comme un paramètre à estimer à part entière. Dans le second cas, le modèle est alors considéré comme un modèle exponentiel courbe (traduction littérale de *curved exponential model*) [49].

Le phénomène modélisé par la statistique *"alternating-star"* (ou GWD) est un phénomène de centralisation au niveau des nœuds de degrés élevés. Cette statistique caractérise la distribution des degrés dans le graphe. En considérant ici le paramètre

λ fixé (les auteurs de [92] ont fixé $\lambda = 2$), θ_{S_k} positif favorise l'émergence de nœuds de degrés importants. Un nœud i de faible degré aura tendance à être connecté à un nœud j de degré important plutôt qu'à un autre nœud de faible degré également. Il est intéressant de souligner que la distinction entre deux nœuds de degré important est faible pour cette statistique. Entre deux nœuds de degré important, le nœud i ne sera pas nécessairement connecté au nœud de degré le plus élevé. Inversement, θ_{S_k} négatif traduit une uniformisation de la distribution des degrés dans le graphe.

Une formulation de la statistique GWD existe aussi pour les graphes dirigés. Le cas des degrés entrants et celui des degrés sortants sont pris en compte séparément :

$$u_\alpha^{(id)}(y) = \sum_{k=0}^{n-1} e^{-\alpha k} d_k^{(in)}(y),$$
$$u_\alpha^{(od)}(y) = \sum_{k=0}^{n-1} e^{-\alpha k} d_k^{(out)}(y).$$

Ils sont respectivement nommés *geometrically weighted in-degrees* et *geometrically weighted out-degrees*.

Indépendance conditionnelle partielle et modèle du circuit social

Les problèmes de dégénérescence que nous avons évoqués un peu plus haut suggèrent que le modèle de Markov n'est pas adapté pour l'étude de certains graphes. Snijders *et al.* [92] en préambule de leur nouvelle spécification des ERGMs, soulignent notamment les limites du modèle de Markov pour représenter des graphes comportant des cliques plus denses et d'ordre plus important que les triangles. En effet, les motifs en triangle ne sont généralement pas équirépartis dans un graphe et sont souvent agrégés en grappes (cluster). Le modèle de Markov qui limite les dépendances entre les arêtes à leur voisinage ne considère pas ces clusters dans leur intégralité mais les interprète comme une composition de triangles ou d'étoiles.

Robins et Pattison [85] proposent une nouvelle hypothèse de dépendance, l'indépendance partielle conditionnelle (traduit littéralement de *partial conditional independence*) moins restrictive que les hypothèses de Markov introduites par Frank et Strauss [28]. Cette nouvelle hypothèse permet à deux arêtes d'être conditionnellement dépendantes sachant le reste du graphe, sans pour autant être voisines l'une de l'autre. La dépendance entre ces deux arêtes est conditionnée par la réalisation d'au moins une troisième arête qui les connecte toutes les deux. Nous retrouvons des travaux analogues pour la modélisation de processus spatiaux [2].

Snijders *et al.* [92] proposent l'hypothèse de dépendance du circuit social (traduction littérale de *social circuit dependence assumption*) qui s'inscrit dans le cadre de l'indépendance partielle conditionnelle de Robins et Pattison [85]. Selon cette hypothèse, deux arêtes (i, j) et (k, l) sont conditionnellement dépendantes sachant le reste du graphe si elles sont voisines (hypothèses de Markov) ou s'il existe deux arêtes (i, k) et (j, l) telle que $Y_{i,j} = Y_{k,l} = 1$. Ce second cas de figure est illustré par la Figure 2.6 et se nomme le circuit social.

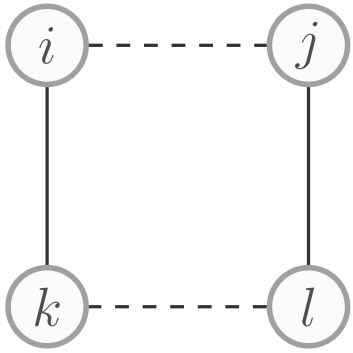


FIGURE 2.6 – Les hypothèses de dépendance du circuit social forment un cylce de 4 arêtes. La réalisation des arêtes (i, j) et (k, l) sont conditionnellement dépendantes sachant le reste du graphe si les arêtes (i, k) et (j, l) existent.

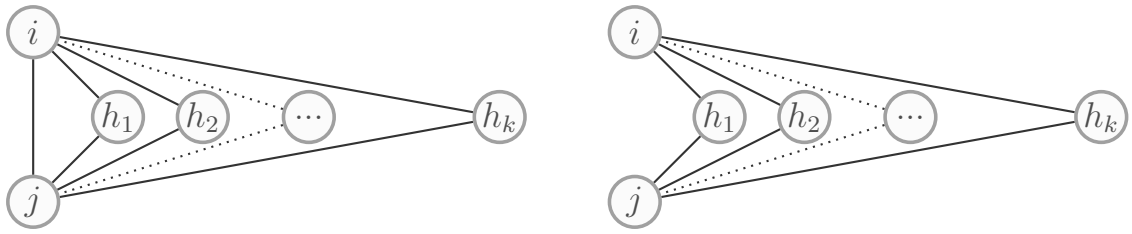
À partir de cette nouvelle dépendance conditionnelle, SNIJDERS *et al.* [92] proposent de prendre en compte deux nouveaux motifs : les "k-triangle" et les *"k-independent 2-paths"* illustrés par la Figure 2.7. Les motifs *"k-triangle"* (Figure 2.7(a)) représentent un ensemble de triangles ayant une base commune. Le *"1-triangle"* correspond au triangle tel que comptabilisé par $T(y)$, le *"2-triangle"* représente deux triangles ayant une base commune et ainsi de suite. Ce motif permet de représenter des graphes dont la distribution de triangles est biaisée et pour lesquels il existe des clusters formés d'une imbrication de triangles. Le nombre de *"k-triangle"* dans un graphe se comptabilise ainsi (pour $k \geq 2$) :

$$T_k(y) = \sum_{i<j} y_{i,j} \binom{S_2(y)}{k}.$$

Le coefficient binomial $\binom{S_2(y)}{k}$ dénombre l'ensemble des deux 2-chemins qui existent entre les nœuds i et j. Les *"k-independent 2-paths"* (Figure 2.7(b)) représentent également des clusters, mais d'ordre inférieur. Nous nommons également ce motif "shared partner" car il représente deux individus qui ne sont pas directement connectés, mais qui partagent des (k) partenaires. La différence avec le motif *"k-triangle"* réside dans le fait qu'il n'y a pas de lien entre les nœuds i et j, il n'y a pas de fermeture transitive. Le nombre de *"k-independent 2-paths"* s'exprime ainsi :

$$S_{k-2p} = \begin{cases} \sum_{i<j} \binom{S_2(y)}{k}, & \text{pour } k \neq 2 \\ \frac{1}{2} \sum_{i<j} \binom{S_2(y)}{k}, & \text{pour } k = 2 \end{cases}.$$

De manière analogue à la statistique *"alternating star"* (réciproquement GWD), l'ensemble des statistiques *"k-triangle"* sont contrôlées par un seul paramètre et pondérées par une suite géométrique. Ce qui permet d'une part de réduire le nombre de paramètres à estimer et d'autre part, de mitiger les problèmes de dégénérescence dont souffrent les modèles de Markov. La statistique résultante se nomme *"alternating k-triangle"* et



(a) "k-triangle" (b) "k-independent 2-paths"

FIGURE 2.7 – Motifs *"k-triangle"* et *"k-independent 2-paths"* introduits par SNIJDERS *et al.* [92]

s'exprime de cette manière :

$$u_\lambda^{t_k(y)} = 3T(y) - \frac{T_2(y)}{\lambda} + \frac{T_3(y)}{\lambda^2} - \cdots + (-1)^{n-3}\frac{T_{n-2}(y)}{\lambda^{n-3}}.$$

Le paramètre réel λ contrôle l'atténuation des motifs d'ordre important. Pour $\lambda > 1$, si le paramètre associé à cette statistique est positif, cela encourage l'occurrence de clusters dans le graphe. De la même manière, les motifs *"k-independent 2-paths"* sont comptabilisés au travers de la statistique *"alternating independent 2-path"* qui se définit :

$$u_\lambda^{s_{k-2p}(y)} = S_{1-2p}(y) - \frac{2}{\lambda}S_{2-2p}(y) + \cdots + \left(\frac{-1}{\lambda}\right)^{k-1} S_{k-2p}(y) + \cdots + \left(\frac{-1}{\lambda}\right)^{n-3} S_{(n-2)-2p}(y).$$

Une variante de ces statistiques existe aussi pour les graphes dirigés. Ces statistiques (pour les graphes dirigés) permettent de définir les relations de transitivités, de *"triad census"* telles que nous les avons présentées dans les Figures 2.5(a) à 2.5(e). Elles s'appuient sur les statistiques du modèle de Markov pour les graphes dirigés. Elles sont également présentées par SNIJDERS *et al.* [92].

2.2.3 Graphes multiniveaux

Jusqu'à présent, les modèles que nous avons présentés portent sur des graphes relativement élémentaires. Les individus représentés par les nœuds du graphe sont tous considérés de la même manière. De même, l'occurrence des motifs comptabilisés par les modèles est supposée homogène. Cette représentation illustre de nombreuses structures sociales comme nous l'avons souligné en introduction de ce chapitre. Néanmoins, elle est limitée pour représenter des structures plus complexes impliquant par exemple différents groupes d'individus. Dès lors, comment cette information d'affiliation (l'appartenance d'un individu à un groupe) peut-elle être encodée dans la représentation graphique ? Une première solution (simple) est de définir un attribut pour chaque individu renseignant sur son affiliation. Ce type de représentation a déjà été évoqué lorsque nous faisions référence au modèle *p2* (sous-section 2.2.2). Dans un premier temps nous pouvons considérer que les attributs associés aux nœuds représentent des caractéristiques exogènes qui ne dépendent pas du processus de création des liens (qui lui est endogène). En revanche, elles peuvent avoir une influence sur ce processus. Par exemple, nous pourrions tester l'hypothèse selon

laquelle les membres d'un même groupe (*i.e.* dont les attributs d'affiliation sont identiques) ont tendance préférentiellement à créer des liens entre eux. Ce type d'hypothèse peut-être testé avec les *social selections models* [83, 84]. Nous pouvons inversement considérer que les attributs sociaux peuvent être influencés par la structure du réseau. Un individu peut être influencé dans son attitude et sa façon d'agir (représentés par des attributs) par les individus avec qui il est en contact. Ce second cas de figure est cette fois-ci considéré par les *social influence models* [22, 86]. Ces modèles sont par exemple adaptés à l'étude des propagations épidémiques. De manière générale, la prise en compte d'attributs au niveau des nœuds enrichit l'observation et permet d'étudier des structures plus complexes.

Il est également possible d'affecter des attributs à des pairs de nœuds (dyades), que l'on nomme covariables dyadiques (*dyadic covariates*) [83]. Ces covariables représentent par exemple, un lien hiérarchique, l'appartenance à une même organisation ou encore la durée pendant laquelle deux individus ont été en contact ... Tout comme les attributs individuels, elles permettent d'étendre l'étude des réseaux à des structures sociales plus complexes.

Néanmoins, ces deux représentations restent limitées lorsqu'il s'agit d'étudier des structures sociales imbriquées. C'est par exemple le cas des structures organisationnelles [11, 14, 108]. Dans certaines situations, les individus sont organisés en équipes, qui sont elles-mêmes organisées en départements, puis regroupés en filiales et ainsi de suite. Cette organisation hiérarchique suggère l'existence de plusieurs réseaux organisés en strates successives : un réseau interpersonnel, un réseau interéquipe, un réseau interdépartement... Les nœuds du réseau interpersonnel représentent les individus, ceux du réseau interéquipe représentent les équipes et enfin ceux du réseau interdépartement, les départements. Ainsi, chaque réseau illustre des relations à différentes échelles. Nous pouvons observer à l'échelle individuelle des liens de collaboration ou d'échange de savoir entre les individus. D'autres liens de collaborations peuvent exister entre les équipes, illustrant par exemple des projets mener conjointement. Enfin, le graphe interdépartement peut quant à lui représenter une organisation institutionnalisée entre les départements ou régie par des accords. Les réseaux de strates adjacentes sont liés entre eux. Par exemple, les individus du réseau interpersonnel sont affiliés à des équipes, les équipes sont quant à elles organisées en départements. Ce type de structure est appelé graphe multiniveaux en référence à ses différentes strates. Nous entrevoyons ici la complexité de la structure qui fait intervenir plusieurs réseaux, représentant différents phénomènes et qui sont interdépendants. Nous remarquons aisément que l'ensemble de ces informations ne peuvent pas être contenues dans un graphe à un seul niveau.

Par souci de clarté, nous considérons un exemple de graphe multiniveaux plus simple, à deux niveaux qui est illustré par la Figure 2.8. Le premier niveau, le niveau micro, comprend le réseau interpersonnel, celui représentant la structure sociale tissée par les individus. Le niveau macro contient le réseau inter-organisationnel et représente les liens entre les organisations à une échelle macroscopique. Le troisième niveau (le niveau méso) qui s'intercale entre les deux autres contient le réseau transversal qui représente les relations entre les individus et les organisations. Isolé, le réseau transversal prend la forme d'un réseau biparti. Ici nous pouvons supposer que les relations de ce réseau illustrent un lien d'affiliation : un individu est affilié à (au moins) une organisation. D'autres formes de réseaux multiniveaux existent [77], la représentation de l'observation dépend

du phénomène étudié.

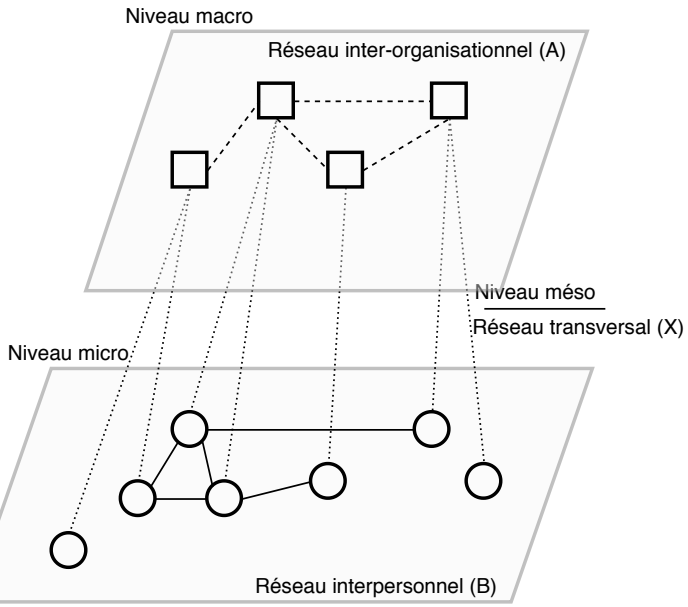


FIGURE 2.8 – Exemple d'un graphe multiniveaux à deux niveaux. La strate inférieure englobe le réseau interpersonnel, la strate supérieure, le réseau inter-organisationnel. Les cercles représentent les individus et les carrés, les organisations. Les relations entre les individus sont matérialisées par des traits pleins et celles entre les organisations par des traits discontinus. Les relations d'affiliation entre les individus et les organisations sont en pointillé.

En fonction des hypothèses initialement posée, plusieurs stratégies peuvent être envisagées pour étudier un réseau multiniveaux tel que celui présenté par la Figure 2.8. On peut par exemple considérer le réseau transversal comme fixe. C'est à dire, supposer que ce réseau représente un phénomène exogène. De cette manière, nous pouvons étudier les structures des réseaux interpersonnel et inter-organisationnel et la façon dont ils s'influencent sachant le réseau transversal [101]. Nous pouvons également faire l'hypothèse que les structures des trois réseaux sont endogènes et ainsi étudier la façon dont ces trois réseaux s'influencent mutuellement.

La notation utilisée pour décrire le réalisation d'un réseau multiniveaux tel que présenté par la Figure 2.8 est $M = m$. Elle peut se décomposer de cette manière $[A, X, B] = [a, x, b]$ où A, B et X représentent respectivement les réseaux inter-organisationnel, transversal et interpersonnel.

La forme générale des ERGMs présentée par (2.1) ne permet pas de prendre en compte plusieurs réseaux à la fois, ni les interdépendances entre ces derniers. Pour pallier cette

limitation Wang *et al.* [101] proposent une nouvelle formulation des ERGMs :

$$p(A=a, X=x, B=b|\theta) = \frac{1}{\kappa(\theta)} exp\Big[\theta_Q z_Q(a) + \theta_Q z_Q(b) + \theta_Q z_Q(x) + \theta_Q z_Q(a,x) + \theta_Q z_Q(b,x) + \theta_Q z_Q(a,x,b)\Big]. \tag{2.12}$$

Cette extension des ERGMs fait intervenir plusieurs groupes de statistiques ($z_Q(\cdot)$) s'appliquant sur chacun des graphes, mais aussi sur plusieurs graphes à la fois. Ces statistiques sont contrôlées par le vecteur de paramètres θ_Q.

Les premières statistiques désignées par $z_Q(a)$ et $z_Q(b)$ comptabilisent les motifs dans les réseaux A et B suivant les hypothèses de dépendance considérées telles que nous les avons énoncées dans la sous-section 2.2.2. Elles s'appliquent sur chaque réseau de manière indépendante sans prendre en compte les interdépendances entre les réseaux. Les statistiques représentées par $z_Q(x)$ s'appliquent sur le réseau transversal qui est un réseau biparti. Une spécification des ERGMs pour les réseaux bipartis est proposée par Wang *et al.* [102] et repris notamment par Wang [100]. Les statistiques suivantes représentées par $z_Q(a,x)$ et $z_Q(b,x)$ comptabilisent des motifs impliquant à la fois les liens d'affiliation du réseau transversal et les liens du réseau interpersonnel ou du réseau inter-organisationnel. Ce qui permet d'étudier l'influence des liens d'affiliation sur les deux autres réseaux et réciproquement. Enfin les dernières statistiques inclues dans $z_Q(a,x,b)$ font intervenir les trois réseaux. Ces dernières permettent d'étudier le réseau multiniveaux dans sa globalité et d'analyser l'influence mutuelle des différents niveaux. Cette extension des ERGMs introduit de nouvelles configurations impliquant plusieurs réseaux à la fois. Les auteurs donnent une liste exhaustive des statistiques prises en compte par le modèle [101]. Par ailleurs, ils soulignent également que ce modèle peut être entendu à des graphes de plus de deux niveaux.

Cette formulation des ERGMs permet de traiter de nouvelles applications qui ont trait notamment aux sciences sociales ou à la théorie du management. Pour n'en citer qu'une : Zappa et Lomi [108] montrent l'influence du réseau interunité dans une entreprise sur les liens interpersonnels et les relations de conseil entre les managers. Ils mettent en évidence le fait que les échanges interpersonnels se réalisent principalement au sein des unités. La structure organisationnelle contraint la manière dont les individus communiquent et échangent l'information. En outre, les relations hiérarchiques favorisent la communication interunité, tout comme les attributs communs partagés entre les individus (homophilie). Les auteurs soulignent l'intérêt de s'appuyer sur une approche multiniveaux pour prendre en considération l'ensemble de la structure organisationnelle et toutes les informations qu'elle recèle.

Nous proposons ici une introduction aux approches multiniveaux en nous concentrant plus particulièrement sur une extension des ERGMs dédiée aux graphes multiniveaux. Ces approches se trouvent être pertinentes pour l'étude de structures sociales hiérarchiques imbriquées. Elles permettent de mettre en perspective les comportements individuels observés et les structures qui s'établissent à différentes échelles. Les questions que soulèvent ces études et leur portée dépassent néanmoins le cadre de ce travail de thèse. Il

est à noter que d'autres modèles et approches [13, 90, 91] existent pour ce type de graphe et que nous n'avons pas couvert l'étendue des applications possibles [61].

Dans cette section, nous avons présenté différentes configurations des ERGMs. Elles reposent en premier lieu sur les hypothèses de dépendances qui définissent les structures locales comptabilisées par le modèle. L'intérêt d'un tel modèle est de pouvoir expliquer une observation. Autrement dit, l'intérêt est de pouvoir estimer les paramètres du modèle sachant un graphe observé. La constante de normalisation $\kappa(\theta)$, du fait de sa complexité, ne permet pas d'estimer facilement les paramètres du modèle. Il est donc nécessaire de recourir à certaines stratégies pour déterminer la valeur des paramètres la plus à même de reproduire l'observation que l'on nomme aussi maximum de vraisemblance. Dans la section suivante (section 2.3), nous présentons ces méthodes d'inférence. Nous introduisons cette section avec les procédures de simulation. Ces procédures consistent à générer un ensemble de graphes sachant un modèle et ses paramètres. Les graphes ainsi construits intègrent les propriétés du modèle. Les procédures de simulation sont utilisées dans les algorithmes d'estimation du maximum de vraisemblance présentés en seconde partie de la section 2.3.

2.3 Méthodes de simulation et d'inférence

2.3.1 Simulation avec les méthodes de Monte-Carlo par chaîne de Markov

Les procédures de simulation sont des algorithmes qui génèrent par modifications successives une chaîne de graphes, sachant un modèle dont les paramètres sont connus et fixés. L'intérêt de ces outils est de pouvoir échantillonner à partir de distributions incalculables, par exemple un ERGM. Dans notre cas, la distribution à échantillonner est $p(y|\theta)$ (2.1) avec θ connu.

On désigne plus généralement ces algorithmes par méthodes de Monte-Carlo par chaîne de Markov de l'anglais *Markov Chain Monte-Carlo* (MCMC) *methods* [106]. Les méthodes MCMC construisent par itérations successives une chaîne d'échantillons, ici de graphes : $\{y^{(i)}\}_{i=1\ldots n}$. Cette chaîne est une chaîne de Markov, l'état d'un échantillon dépend uniquement du précédent, autrement formulé :

$$\mathbb{P}(y^{(i)}|y^{(i-1)}, y^{(i-2)}, \ldots, y^{(0)}) = \mathbb{P}(y^{(i)}|y^{(i-1)}).$$

Une chaîne de Markov est un processus sans mémoire. L'échantillon $y^{(i)}$ est le résultat d'une modification appliquée à l'échantillon précédent $y^{(i-1)}$. Nous matérialisons cette modification par une transition de $y^{(i-1)}$ vers $y^{(i)}$. Cette transition se définie au travers du noyau de transition $T(y, A)$ tel que :

$$T(y, A) = T(y^{(i+1)} \in A|y^{(i)} = y),$$

où A est un sous-ensemble de l'espace d'états. Ainsi, une chaîne de Markov à l'étape n peut se définir par un état initial et une succession de transitions menant jusqu'à l'étape

n :

$$T(y^{(n)} \in A | y^{(0)} = y) = T^n(y, A),$$

où T^n peut se définir par récurrence sur n avec la multiplication à gauche comme $T^n = TT^{n-1}$.

Pour assurer la convergence de la chaîne vers la distribution stationnaire (que l'on désigne ici par π), la chaîne doit respecter plusieurs propriétés [30, 33, 66, 99].

Invariance La distribution π est invariante pour le noyau T si $\pi = T\pi$. Autrement dit, pour une chaîne de noyau T, si $y^{(i)}$ suit la loi de π, alors l'état suivant $y^{(i+1)}$ la suit également.

Irréductibilité À partir d'un état initial, la chaîne peut atteindre en n étapes tous les états dont la probabilité de réalisation est non-nulle suivant la distribution π.

Récurrence Un état est dit récurrent s'il peut être visité une infinité de fois à partir de tous les points départ de la chaîne. Une chaîne est quant à elle récurrente si tous les états atteignables ($\pi(A) > 0$) de la chaîne sont récurrents. Une chaîne récurrente irréductible qui a une distribution invariante est une chaîne récurrente positive.

Apériodicité Le noyau T est dit apériodique si pour toute partition $\{A_1, \ldots, A_d\}$ de A il n'existe pas de cycle tel que $\forall i < d, T(y_1 \in A_{i+1} | y_0 \in A_i) = 1$. En d'autres termes, le comportement de la chaîne n'admet pas de cycle.

Ces propriétés permettent de conclure à la convergence de la chaîne vers la distribution stationnaire π. Nous disons aussi que la chaîne est ergodique. Plus formellement, pour tout état initial $y \in A$, lorsque $n \to \infty$:

$$\left| T(y^{(n)} \in A | y^{(0)} = y) - \pi \right| \to 0,$$

Dans la suite de cette section, nous présentons deux algorithmes d'échantillonnage : l'Algorithme Metropolis-Hastings et l'échantillonneur de Gibbs qui permettent de simuler une distribution incalculable.

Metropolis-Hasting

L'algorithme Metropolis-Hastings [45] construit par itérations successives une suite d'échantillons $\{y^{(i)}\}_{i=0\ldots n-1}$. Chaque itération peut être décomposée en deux phases : une phase de proposition et une phase d'acceptation/rejet. La première phase consiste donc à proposer un nouveau candidat y' sachant l'étape courant y. Cette proposition est générée suivant la densité de proposition $q(y \to y')$. Par la suite, le candidat y' est acceptée avec la probabilité :

$$\alpha_{y \to y'} = \min \left[1, \frac{p(y'|\theta) q(y' \to y)}{p(y|\theta) q(y \to y')} \right]. \tag{2.13}$$

Dans le cas où la proposition est refusée, le nouvel état reprend l'état précédent y. La dynamique de cet algorithme reproduit les itérations du noyau de transition tel que nous

l'avons défini précédemment. Une attention particulière doit être portée sur le choix de la densité de proposition $q(y \to y')$ pour assurer le respect des propriétés de convergence énoncées ci-dessus. Dans notre cas, une solution assez simple qui garantit les propriétés de convergence est la distribution uniforme. Notons néanmoins qu'elle induit certains effets négatifs, à savoir, un fort taux de rejet et une corrélation élevée entre les échantillons. En outre, le choix de la distribution uniforme permet de simplifier le ratio d'acceptation (2.13) puisque la distribution est symétrique et donc : $q(y' \to y)/q(y \to y') = 1$.

Dans le cas d'un ERGM, la modification mise en jeu dans la phase de proposition est élémentaire et revient à modifier l'état d'une arête $y_{i,j}$. Le choix effectué est binaire et influe sur l'existence ou l'absence de l'arête en question. À la manière d'un lancer de pièce, la distribution de proposition détermine l'état "candidat" de l'arête : $y_{i,j} = 1$ (l'arête est activée), $y_{i,j} = 0$ (l'arête est désactivée). Ainsi, l'échantillon y' fait référence au graphe y intégrant l'état candidat de l'arête $y_{i,j}$. Le ratio d'acceptation (2.13) peut s'exprimer ainsi :

$$\alpha_{y \to y'} = \min \left[1, \frac{p(y'|\theta)}{p(y|\theta)} = \exp\left(\langle \theta, (t(y') - t(y)) \rangle\right) \right]. \tag{2.14}$$

On remarque que le ratio d'acceptation ne fait pas intervenir la constante de normalisation $\kappa(\theta)$, le calcul est donc facilement réalisable. Par ailleurs, le calcul des statistiques $t(\cdot)$ peut se limiter aux arêtes dépendantes de $y_{i,j}$ suivant les hypothèses de dépendances préalablement établies. L'Algorithme 1 détaille les différentes étapes de l'algorithme Metropolis-Hastings pour les modèles exponentiels.

Algorithme 1 Algorithme Metropolis-Hastings

function MetropolisHastings(y_0, *sim_iter*) ▷ y_0 : graphe initial, *sim_iter* : nombre d'itérations
  $y \leftarrow y_0$
  for $iter = 1 \dots sim\_iter$ **do**
    for i, j in $y.edges.index$ **do** ▷ Balayage séquentiel des arêtes
      $y' \leftarrow y$ ▷ Copie du graphe y
      $y'_{i,j} \sim q(y_{i,j} \to y'_{i,j})$ ▷ Nouvelle proposition
      $\alpha_{y \to y'} = \min\left[1, \exp\left(\langle \theta, (t(y') - t(y)) \rangle\right)\right]$ ▷ Calcul du ratio
      $accepted \leftarrow \mathcal{U}(0, 1)$
      if $\alpha_{y \to y'} > accepted$ **then**
        $y_{i,j} \leftarrow y'_{i,j}$ ▷ Proposition acceptée
      end if
    end for
    $y^{(iter)} \leftarrow y$ ▷ Sauvegarde du graphe
  end for
  return $\{y^{(i)}\}_{i=1 \dots sim\_iter}$
end function

Il existe d'autres variantes de l'algorithme Metropolis-Hastings visant à améliorer la convergence de la chaîne. Face au problème de dégénérescence et pour éviter que la chaîne ne reste bloquée dans un régime, Snijders [89] propose de faire des mises à jour

plus importantes, en suggérant d'inverser l'intégralité du graphe : les arêtes activées sont désactivées et vice versa. Ces inversions interviennent à intervalle de temps réguliers de telle sorte que la chaîne puisse explorer l'ensemble des états possibles. Ce type de stratégie est particulièrement utile lorsque la distribution échantillonnée a une forme bimodale. Morris, Handcock et Hunter [69] suggèrent une autre variante, l'échantillonneur "*tie no tie*" (TNT), dont la stratégie est de sélectionner de manière équiprobable entre les arêtes activées et désactivées l'arête à mettre à jour. Cela a pour effet d'améliorer les propriétés de mixage de la chaîne, en particulier lorsqu'il s'agit de simuler des graphes faiblement denses.

Échantillonneur de Gibbs

L'algorithme Metropolis-Hastings peut souffrir d'un fort taux de rejet, ce qui ralentit la convergence de la chaîne vers la distribution stationnaire. Pour contourner cette difficulté, il est possible, dans le cas présent, d'utiliser un autre algorithme de simulation : l'échantillonneur de Gibbs [32]. La différence principale entre les deux algorithmes réside dans la façon de choisir le nouveau candidat. Au lieu d'utiliser une distribution auxiliaire (ligne 6 de l'Algorithme 1) pour sélectionner une proposition qui pourrait être rejetée, l'échantillonneur de Gibbs choisit une nouvelle proposition sachcant l'état actuel d'après la probabilité $P(Y|y^c)$ (où y^c est la configuration actuelle). Ce qui revient à sélectionner directement la meilleure proposition et donc à accepter chaque nouvelle proposition de telle sorte que $\alpha_{y \to y'} = 1$. L'Algorithme 2 reprend l'ensemble des étapes de l'échantillonneur de Gibbs, ici adapté pour l'échantillonnage de graphes. Il partage de nombreuses étapes communes avec l'algorithme Metropolis-Hastings (Algorithme 1).

Algorithme 2 Échantillonneur de Gibbs

1: **function** GibbsSampler(y_0, sim_iter) ▷ y_0 : graphe initial, sim_iter : nombre d'itérations
2: $y \leftarrow y_0$
3: **for** $iter = 1 \dots sim\_iter$ **do**
4: **for** i, j in $y.edges.index$ **do** ▷ Balayage séquentiel des arêtes
5: $y'_{i,j} \sim P(\cdot|y)$ ▷ Nouvelle proposition
6: $y_{i,j} \leftarrow y'_{i,j}$ ▷ Mise à jour de l'arête
7: **end for**
8: $y^{(iter)} \leftarrow y$ ▷ Sauvegarde du graphe
9: **end for**
10: **return** $\{y^{(i)}\}_{i=1 \dots sim\_iter}$
11: **end function**

Dans le cas d'un graphe binaire, le tirage du nouvel état (ligne 5 de l'Algorithme 2) consiste à choisir entre deux états : $y_{i,j} = 1$ l'arête est activée et $y_{i,j} = 0$ l'arête est désactivée suivant les probabilités $P(Y_{i,j} = 1|y)$ et $P(Y_{i,j} = 0|y)$. Elles s'expriment de la

manière suivante :

$$P(Y_{i,j} = 1|y) = \frac{P(Y_{i,j} = 1, Y_{-i,j} = y_{-i,j}|\theta)}{P(Y_{i,j} = 1, Y_{-i,j} = y_{-i,j}|\theta) + P(Y_{i,j} = 0, Y_{-i,j} = y_{-i,j}|\theta)},$$
$$P(Y_{i,j} = 0|y) = \frac{P(Y_{i,j} = 0, Y_{-i,j} = y_{-i,j}|\theta)}{P(Y_{i,j} = 1, Y_{-i,j} = y_{-i,j}|\theta) + P(Y_{i,j} = 0, Y_{-i,j} = y_{-i,j}|\theta)},$$

où $Y_{-i,j} = y_{-i,j}$ signifie que tout le graphe y, excepté l'arête (i, j) reste inchangé. Notons que ces deux ratios peuvent se simplifier de telle sorte que la constante de normalisation $\kappa(\theta)$ n'intervient plus dans le calcul.

Dans cette sous-section (sous-section 2.3.1), nous avons présenté deux algorithmes de simulation : Metropolis-Hastings et l'échantillonneur de Gibbs. L'objectif est de générer un ensemble d'échantillons à partir d'une distribution donnée. Dans le cas présent, ils permettent de simuler un ensemble de graphes sachant un modèle et ses paramètres. Les graphes ainsi générés reproduisent les propriétés du modèle (hors cas de dégénérescence). Les propriétés de convergence garantissent que la chaîne d'échantillons résultante atteindra la distribution stationnaire après un nombre suffisant d'étapes. Les deux méthodes MCMC que nous avons présentées sont relativement peu coûteuses et permettent d'approcher la distribution dans un temps raisonnable. A contrario, des méthodes de simulation dites parfaites [47] comme la méthode *Coupling From the Past* (CFP) [81] permettent d'échantillonner la distribution exactement. Ces méthodes sont généralement plus coûteuses, néanmoins nécessaires dans certaines situations que nous aurons l'occasion d'évoquer dans la section 3.4. Les algorithmes de simulation que nous avons présentés sont utilisés pour estimer les paramètres du modèle. Ils interviennent notamment dans l'approximation du log de vraisemblance pour l'estimation du maximum de vraisemblance que nous détaillerons dans la sous-section suivante (sous-section 2.3.2). Ils sont aussi mis en œuvre pour l'échantillonnage d'une variable auxiliaire dans certains algorithmes d'inférence bayésienne, nous y reviendrons dans la section 3.4. Une fois les paramètres estimés, ces algorithmes permettent également de vérifier la qualité de l'estimation. En simulant le modèle avec les paramètres estimés, il est possible de vérifier que les échantillons simulés reproduisent bien l'observation. Si ce n'est pas le cas, cela signifie que l'estimation n'est pas de bonne qualité.

2.3.2 Inférence : estimation du maximum de vraisemblance

L'estimation du maximum de vraisemblance consiste à déterminer la valeur des paramètres d'un modèle (2.1) en mesure de reproduire le plus fidèlement possible l'observation qui est faite. Le maximum de vraisemblance se définit plus formellement comme étant la valeur $\hat{\theta}$ des paramètres dont la distribution des échantillons est centrée sur l'observation y :

$$\mathbb{E}_{\hat{\theta}} t(Y) - t(y) = 0.$$

On appelle cette équation, équation du moment [58, 89]. La principale difficulté concernant l'estimation du modèle réside dans l'évaluation de la constante de normalisation (2.2) qui est analytiquement incalculable.

Une manière de contourner le problème est d'utiliser l'estimation du maximum de vraisemblance de Monte-Carlo, de l'anglais *Monte-Carlo Likelihood Estimation* (MCMLE). Cette approche introduite par Geyer et Thompson [35] et Geyer [33] consiste à maximiser la fonction du log de vraisemblance qui se définit comme le logarithme de la densité de probabilité $p(y|\theta)$ (2.1) :

$$l(\theta) = \langle \theta, t(y) \rangle - \log(\kappa(\theta)),$$

où y est fixe et représente l'observation. Cette fonction du log de vraisemblance s'exprime plus généralement sous la forme d'un ratio. En considérant un paramètre θ_0 fixé et connun, nous définissons le ratio du log de vraisemblance :

$$l_{\theta_0}(\theta) = \langle (\theta - \theta_0), t(y) \rangle - \log \left(\frac{\kappa(\theta)}{\kappa(\theta_0)} \right). \tag{2.15}$$

Maximiser le ratio du log de vraisemblance, revient à maximiser le log de vraisemblance à une constante près. Cette représentation ne nous permet toujours pas de nous soustraire aux difficultés liées à la constante de normalisation. Ici, le ratio des constantes de normalisation $\kappa(\theta)/\kappa(\theta_0)$ reste analytiquement incalculable. Pour surmonter cette difficulté, les auteurs de l'approche proposent une approximation de ce ratio en partant de la définition de la constante de normalisation [3] :

$$\begin{aligned}
\frac{\kappa(\theta)}{\kappa(\theta_0)} &= \frac{1}{\kappa(\theta_0)} \int_{\mathcal{Y}} \exp(\langle \theta, t(y) \rangle) \mu(dy) \\
&= \int_{\mathcal{Y}} \frac{\exp(\langle \theta, t(y) \rangle)}{\exp(\langle \theta_0, t(y) \rangle)} \frac{\exp(\langle \theta_0, t(y) \rangle)}{\kappa(\theta_0)} \mu(dy) \\
&= \int_{\mathcal{Y}} \exp(\langle (\theta - \theta_0), t(y) \rangle) p(y|\theta_0) \mu(dy) \\
&= \mathbb{E}_{\theta_0} \exp(\langle (\theta - \theta_0), t(Y) \rangle), Y \sim p(y|\theta_0).
\end{aligned}$$

Cette définition du ratio des constantes de normalisation peut-être approximée grâce aux méthodes de Monte-Carlo que nous avons introduites dans la sous-section précédente (sous-section 2.3.1). Ainsi, en considérant un ensemble d'échantillons $y_1 \ \ldots y_n$ réalisations des variables $Y_1 \ldots Y_n$ indépendantes et identiquement distribuées (i.i.d) simulées à partir de $p(y|\theta_0)$, nous pouvons écrire :

$$\frac{\kappa(\theta)}{\kappa(\theta_0)} \approx \frac{1}{n} \sum_{i=1}^{n} \exp(\langle (\theta - \theta_0), t(y_i) \rangle) \cdot$$

De cette manière le ratio du log de vraisemblance :

$$l_{\theta_0}(\theta) = \langle (\theta - \theta_0), t(y) \rangle - log \left(\mathbb{E}_{\theta_0} \exp(\langle (\theta - \theta_0), t(Y) \rangle) \right)$$

est approximé par :

$$l_{n,\theta_0}(\theta) = \langle (\theta - \theta_0), t(y) \rangle - log \left(\frac{1}{n} \sum_{i=1}^{n} \exp(\langle (\theta - \theta_0), t(y_i) \rangle) \right). \tag{2.16}$$

3. Dans la situation où l'espace des configurations est continu et admet une mesure μ.

Le maximum $\hat{\theta}_n$ de (2.16) donne une approximation cohérente du maximum de vraisemblance $\hat{\theta}$. Nous pouvons aisément calculer le gradient de (2.16) :

$$\nabla l_{n,\theta_0}(\theta) = t(y) - \frac{\sum_{i=1}^{n} t(y_i) \exp(\langle (\theta - \theta_0), t(y_i) \rangle)}{\sum_{i=1}^{n} \exp(\langle (\theta - \theta_0), t(y_i) \rangle)} \tag{2.17}$$

Le dernier membre de (2.16) peut être identifié comme un échantillonnage préférentiel dont la pondération est donnée par :

$$w_{n,\theta,\theta_0}(y) = \frac{\exp(\langle (\theta - \theta_0), t(y) \rangle)}{\sum_{i=1}^{n} \exp(\langle (\theta - \theta_0), t(y_i) \rangle)}.$$

(2.17) peut ainsi se simplifier de cette manière :

$$\nabla l_{n,\theta_0}(\theta) = t(y) - \sum_{i=1}^{n} t(y_i) w_{n,\theta,\theta_0}(y_i)$$

De la même manière, il est possible de définir le hessien :

$$\begin{aligned} \nabla^2 l_{n,\theta_0}(\theta) &= -\left[\sum_{i=1}^{n} w_{n,\theta,\theta_0}(y_i) \langle t(y_i), t(y_i) \rangle - \left\langle \sum_{i=1}^{n} w_{n,\theta,\theta_0}(y_i) t(y_i), \sum_{i=1}^{n} w_{n,\theta,\theta_0}(y_i) t(y_i) \right\rangle \right] \\ &= -\mathbb{V}ar\, t(Y) w_{n,\theta,\theta_0}(Y) \end{aligned}$$

On remarque que le gradient et le hessien peuvent facilement se calculer et produire un estimateur cohérent. À partir de ces deux mesures, nous pouvons utiliser une méthode d'optimisation locale telle que le *Monte-Carlo Newton Raphson* (MCNR) pour déterminer le maximum de vraisemblance. Cette méthode met à jour par itérations successives le paramètre θ :

$$\theta_{k+1} = \theta_k - \left[\nabla^2 l_{n,\theta_0}(\theta_k) \right]^{-1} \nabla l_{n,\theta_0}(\theta_k),$$

de telle sorte qu'il converge à terme vers le maximum de vraisemblance.

Geyer [34] introduit deux mesures permettant d'évaluer la qualité de l'estimation. La première de ces mesures est l'écart-type asymptotique $\hat{\sigma}_\theta$ qui donne la différence entre les vrais paramètres du modèle et le maximum de vraisemblance qui sont tous les deux inconnus. Cette mesure se calcule à partir de l'information de Fisher :

$$B_n = -\nabla^2 l_{n,\theta_0}(\hat{\theta}_n).$$

L'écart-type asymptotique $\hat{\sigma}_\theta$ se calcule en prenant la racine carrée des valeurs diagonales de B_n^{-1}. La seconde mesure est l'écart-type de Monte-Carlo qui donne la différence entre l'estimation du maximum de vraisemblance de Monte-Carlo et le maximum de vraisemblance qui lui est inconnu : $\hat{\theta}_n - \hat{\theta}$. Elle permet d'évaluer la précision de l'approximation de Monte-Carlo. L'écart-type de Monte-Carlo $\hat{\sigma}_\theta^{MC}$ correspond à la racine carrée de la diagonale de :

$$\frac{1}{\sqrt{n}} B_n^{-1} A_n B_n^{-1}.$$

Les calculs des matrices A_n et B_n s'appuient sur un ensemble de n échantillons $Y_1 \dots Y_n$ préalablement simulés à partir de $p(y|\hat{\theta}_n)$. A_n peut s'exprimer ainsi :

$$A_n = \frac{C_n}{\left[\frac{1}{n}\sum_{i=1}^{n} \exp\left(t(Y_i)^T(\hat{\theta}_n - \theta_0)\right)\right]^2},$$

où C_n est la matrice de covariance empirique de $(t(y) - t(Y)) \exp\left(t(Y)^T(\hat{\theta}_n - \theta_0)\right)$.

Un des inconvénients majeurs de la méthode MCNR et plus largement du MCMLE de Geyer et Thompson [35] réside dans le choix du paramètre θ_0. Ce paramètre initial doit être suffisamment proche du maximum de vraisemblance $\hat{\theta}$. L'échantillonnage préférentiel sur lequel repose le calcul du gradient et du hessien est très sensible à cet aspect. Une des stratégies possibles pour définir la valeur du paramètre θ_0 est d'utiliser le maximum de pseudo-vraisemblance [8, 50, 98]. Cet estimateur est relativement pauvre pour des modèles s'appuyant sur des hypothèses de dépendances, puisqu'il fait l'hypothèse que les arêtes sont mutuellement indépendantes. Cette hypothèse d'indépendance facilite grandement le calcul, mais rend l'estimateur peu fiable lorsque les interactions sont fortes. Lorsque θ_0 est trop éloigné du maximum de vraisemblance, il est nécessaire de répéter l'opération d'échantillonnage de $p(y|\theta_0)$ avec θ_0 mis à jour avec la dernière valeur estimée de θ. Les propriétés du log de vraisemblance garantissent que cette stratégie permettra de converger à terme vers le maximum de vraisemblance, mais avec un coût important. Il existe d'autres algorithmes d'inférence qui s'appuie sur les méthodes MCMC pour approximer le maximum de vraisemblance. Snijders [89] propose une approche stochastique qui repose sur l'algorithme de Robbins-Monro [82]. Cette approche a l'avantage de ne pas être sensible au choix d'une valeur initiale de θ et ne nécessite pas un grand nombre d'échantillons. Geyer et Thompson [35] font néanmoins remarquer que l'approche stochastique ne fait pas une utilisation aussi efficace des méthodes MCMC que le MCMLE [51]. De manière générale, les problèmes liés à l'inférence des modèles exponentiels restent des questions ouvertes [35, 34, 33, 51, 48, 89]. Nous aurons l'occasion de revenir sur ces difficultés dans la section 3.4, où nous présenterons les méthodes d'inférence bayésiennes. Une revue étendue des méthodes pour l'inférence des ERGMs est donnée par Hunter, Krivitsky et Schweinberger [51].

2.4 Synthèse

Dans ce chapitre, nous avons présenté une famille de modèles statistiques : les ERGMs qui sont dédiés à l'étude des structures sociales. Le point de départ est donc l'observation d'une structure sociale sous la forme d'un graphe (ou réseau) représentant des individus et les liens tissés entre ces derniers. La raison principale qui motive l'usage de ces modèles est de déterminer quelles sont les interactions qui sont entrées en jeu dans la création de ces structures. La structure sociale dans son ensemble est alors vue comme une composition de structures locales impliquant un petit nombre d'acteurs. Identifier ces structures locales revient à identifier les phénomènes et les processus à l'origine de la structure globale. Les ERGMs se déclinent en plusieurs configurations qui comptabilisent différentes formes de structures locales. Cette diversité des configurations permet de confronter différentes

hypothèses sur les processus qui ont mené à la création de la structure observée. En somme les ERGMs sont des outils flexibles qui permettent de traiter une grande variété de questions. De plus, l'interprétation de ces modèles est relativement simple. À chaque type de structure comptabilisé est associé un paramètre qui renseigne sur sa présence dans le graphe.

L'estimation de ces modèles est cependant difficile, du fait de la constante de normalisation qui dénombre l'ensemble des configurations de graphes possibles. Plusieurs stratégies ont été mises en place dans la littérature pour estimer la valeur des paramètres (maximum de vraisemblance) à partir d'une observation. En réalité, ces approches proposent une approximation du maximum de vraisemblance, qui repose généralement sur les méthodes MCMC. Les méthodes MCMC font le cheminement inverse, elles simulent un ensemble de graphes à partir d'un modèle dont les paramètres sont connus. Nous avons présenté les deux algorithmes de simulation les plus connus : le Metropolis-Hastings et l'échantillonneur de Gibbs. Malgré ces stratégies d'approximation, le problème de l'estimation de ces modèles reste une question ouverte et n'admet pas de solutions évidentes.

Les ERGMs se trouvent à l'intersection de plusieurs champs disciplinaires. L'usage des ERGMs est motivée par des questions qui ont trait aux sciences sociales, à la théorie du management et se rapportent plus généralement à l'étude des réseaux sociaux. La modélisation statistique et les problèmes liés à l'inférence et à la simulation de ces modèles mobilisent des travaux en statistiques et en informatique. Il semble difficile de différencier ces deux aspects, tant ils sont mutuellement liés. Les besoins liés à l'application guident les recherches en modélisation statistique et sur les algorithmes. Inversement, les avancées sur les problématiques liées à l'inférence et la simulation ouvrent la voie vers de nouveaux champs d'application. Les travaux de thèses que nous présentons s'inscrivent dans cette démarche.

Dans le chapitre suivant (Chapitre 3) nous présentons une approche de modélisation et d'inférence. Cette approche partage de nombreux points communs avec les ERGMs. La forme générale du modèle est identique et l'objet observé reste les interactions entre des individus. Néanmoins, le formalisme diffère quelque peu. Nous ne considérons pas directement le graphe social, mais une représentation duale, le graphe des interactions. Ce changement de perspective nous amène à changer de formulation et à considérer le processus de création d'une arête comme un processus de coloration de graphe. Nous intégrons aussi dans ce modèle une information organisationnelle. Ces nouveaux éléments ne nous permettent pas de nous soustraire aux problèmes en rapport avec l'inférence. Nous proposons d'utiliser une méthode bayésienne pour estimer les paramètres et déterminer l'incertitude autour de cette estimation.

Le développement de cette approche est intrinsèquement lié à l'application ou tout du moins à une problématique concrète. Ce travail a en premier lieu été motivé par la volonté de comprendre la structure des interactions entre les membres de plusieurs organisations engagés dans une collaboration. Le Chapite 4 présente une application inhérente à cette problématique, qui porte sur les collaborations inter-organisationnelles entre les chercheurs. Nous montrons également dans le Chapitre 5 que l'approche que nous proposons peut s'appliquer à d'autres structures organisationnelles dans un tout autre contexte.

Chapitre 3

Description morpho-statistique des interactions dans les graphes

Sommaire

3.1 Motivations et aperçu de l'approche

La première raison qui a motivé le développement de cette approche est la volonté de comprendre la manière dont des individus appartenant à différentes organisations interagissent entre eux, au sein de ces organisations et entre ces organisations. Le cas d'usage que nous avions initialement envisagé est la collaboration inter-organisationnelle. Dans ce contexte, les motifs d'interactions semblent être une source d'information sur la manière dont les informations sont échangées entre les protagonistes et sur le contrôle appliqué sur ces échanges. Nous avons pris le parti de considérer la structure tissée par les collaborateurs à la fois comme vecteur d'informations et élément conditionnant la collaboration. Un des enjeux majeurs est d'étudier l'interface entre les organisations. Autrement dit, comment un groupe d'individus s'organise-t-il pour échanger avec un d'autres groupes ? Une des finalités que nous poursuivons dans ce travail se rapporte aux applications collaboratives et à la mise en œuvre de topologies pair-à-pair à même de supporter ces collaborations. Il semble d'autant plus intéressant d'étudier la structure de ces interactions, qu'elle peut être reprise par la suite comme topologie pair-à-pair dans le développement d'applications collaboratives distribuées.

L'objet d'étude initial est un graphe social tel que nous l'avons évoqué dans le chapitre précédent (Chapitre 2). Nous nous intéressons plus particulièrement aux liens tissés entre les individus. Nous opérons donc un changement de perspective pour mettre les interactions entre les individus au premier plan. De ce fait, nous n'utilisons pas directement le graphe social, mais le graphe des arêtes. Ce graphe des arêtes est une représentation duale du graphe social. Les arêtes du graphe social sont les nœuds du graphe des arêtes. Deux arêtes du graphe social qui partagent une extrémité commune sont reliées dans le graphe des arêtes. De cette manière, nous mettons en évidence la structure des interactions. Il est aussi nécessaire de prendre en compte la structure organisationnelle même si elle est simplifiée par rapport à celles considérées par les graphes multiniveaux (sous-section 2.2.3). Les organisations ne sont définies ici que par les membres qui les composent et n'ont pas d'autre signification. Ce ne sont en somme que des groupes d'individus. Dans la représentation associée au graphe des arêtes, nous distinguons deux types d'interactions : les liens intra-organisationnels et les liens inter-organisationnels. L'émergence d'un lien entre deux individus revient à labéliser le nœud correspondant dans le graphe des arêtes en fonction de l'affiliation des deux collaborateurs. Ce qui revient à considérer l'émergence d'un lien dans la collaboration comme un processus d'étiquetage [4] dans le graphe des arêtes. Quoi qu'il en soit, la structure du graphe des arêtes reste inchangée, seuls les labels associés aux nœuds changent en fonction de la configuration des liens de collaboration. La transformation que nous décrivons ici et le formalisme utilisé constituent la première étape de notre approche. Cette étape est décrite dans la section 3.2.

L'objectif de l'approche est toujours de caractériser l'occurrence des liens dans le graphe social. Ce qui revient à caractériser la distribution des labels dans le graphe des arêtes. Pour ce faire, nous utilisons un modèle d'interactions : le modèle de Potts. Ce modèle est très proche des ERGMs, mais nous l'empruntons au domaine des statistiques

4. Le fait d'associer un label/une étiquette à un nœud

spatiales. Nous considérons l'hypothèse de dépendance de Markov. Ce qui signifie que l'état d'un nœud dans le graphe des arêtes ne dépend que de l'état des nœuds voisins. Ce qui nous permet de considérer des configurations élémentaires impliquant deux nœuds voisins dans le graphe des arêtes. Par exemple deux nœuds étiquetés "intra-organisationnel". L'intérêt d'un tel modèle est de pouvoir caractériser l'occurrence de ces motifs dans le graphe des arêtes. Deux nœuds voisins dans le graphe des arêtes correspondent à une configuration "relais" dans le graphe social où un individu (au centre) est en contact avec deux autres individus de part et d'autre. En fonction de l'état des nœuds dans le graphe des arêtes, les individus dans le graphe social sont affiliés ou non à la même organisation. Par exemple, une configuration où un nœud "intra-organisationnel" est voisin d'un nœud "inter-organisationnel" équivaut dans le graphe social à une configuration, où un individu "relais" est à la fois connecté avec un membre de son organisation et un membre extérieur. Ainsi, le modèle de Potts nous permet d'étudier indirectement l'occurrence des configurations "relais" dans le graphe social. Cette notion de relais est importante, voire fondamentale, dans la diffusion d'information dans le cadre d'une collaboration. Elle se prête également à d'autres contextes, comme la propagation d'un virus au sein d'une population donnée. Nous proposons dans cette approche une modélisation relativement élémentaire qui encode à la fois la structure des interactions et prend en compte la structure organisationnelle.

La forme générale du modèle de Potts est identique à celle des ERGMs, nous faisons donc face aux mêmes difficultés quant à l'estimation des paramètres (sous-section 2.3.2). Nous proposons ici d'utiliser une approche bayésienne qui consiste à échantillonner la distribution des paramètres du modèle en mesure d'expliquer l'observation. Traiter la distribution *a posteriori* des paramètres permet de déterminer le maximum de vraisemblance et l'incertitude sur cette estimation. Il est malgré tout difficile d'échantillonner directement cette distribution (distribution *a posteriori*). Plusieurs stratégies ont été présentées dans l'état de l'art pour contourner cette difficulté. Nous les présentons dans la section 3.4. Dans le cadre de l'approche, nous utilisons l'algorithme ABC Shadow qui a déjà été utilisé pour la modélisation de processus spatiaux. Cet algorithme approxime la distribution *a posteriori*. Dans le cas d'ABC Shadow, l'approximation est contrôlée par des paramètres d'entrée. Ce qui permet de nous assurer que la distribution échantillonnée par ABC Shadow est suffisamment proche de la vraie distribution. Dans la section 3.5 nous démontrons la qualité de l'estimation fournie par ABC Shadow dans le cas d'un modèle analytiquement calculable et pour le modèle de Potts avec des paramètres connus. Cette étape est nécessaire pour ajuster au mieux les paramètres contrôlant l'erreur de l'approximation.

Cette approche est décrite ici sous le prisme des collaborations inter-organisationnelles. Néanmoins, elle peut s'appliquer à d'autres contextes impliquant également une structure organisationnelle comme nous le verrons dans le Chapitre 5.

3.2 Observation à partir graphe dual

Comme nous l'avons évoqué dans le Chapitre précédent (Chapitre 2), les structures sociales sont très souvent représentées sous la forme d'un graphe que l'on peut désigner par

la notation : $\mathcal{G}(\mathcal{V}, \mathcal{E})$. L'ensemble $\mathcal{V}$ désigne les nœuds du graphe et l'ensemble $\mathcal{E}$, les arêtes reliant les nœuds. Par la suite, nous désignerons ce graphe par le terme graphe social [103]. Les nœuds d'un graphe social représentent des individus. Une arête reliant deux nœuds dans le graphe représente un lien relationnel. La nature des liens et le type d'individu diffèrent d'un contexte à l'autre. Par exemple, un graphe social très couramment utilisé dans l'état de l'art représente des relations entre des moines en proie à un conflit interne au sein d'un monastère [87]. Les nœuds du graphe représentent les moines, les arêtes sont dirigées et illustrent des liens de soutien d'un moine à un autre. Un autre jeu de données également très connu illustre quant à lui les liens matrimoniaux tissés par les grandes familles Florentines (Médicis, Strozzi...) pendant la Renaissance [54]. De ce fait, la représentation graphique des structures sociales (graphe social) peut décrire de nombreuses situations. De plus, il est possible d'ajouter des attributs à chaque noeuds renseignant par exemple sur : l'age, le sexe, l'affiliation à une organisation ...

Nous observons ici des graphes sociaux représentant des collaborations. Tout comme dans la sous-section 2.2.3, nous considérons deux niveaux dans la collaboration. Le premier niveau est celui des individus. Les individus sont des nœuds reliés par des liens de collaboration. Le second niveau porte quant à lui sur les organisations. Chaque individu est affilié à une organisation. À la différence des approches multiniveaux (sous-section 2.2.3), nous ne considérons pas ici de structures à l'échelle organisationnelles. Les relations entre les organisations sont uniquement tissées par les liens individuels. Nous adoptons une représentation simplifiée qui considère un graphe de collaboration à l'échelle individuelle sur lequel ont été ajoutées des informations d'affiliation.

Le graphe est construit en ne considérant qu'un seul type de lien, les liens de collaboration entre les individus. Ensuite, nous faisons l'hypothèse que tous les collaborateurs sont affiliés à une et une seule organisation. Un attribut représentant l'affiliation est associé à chaque nœud. Les organisations sont uniquement représentées au travers des membres qui la composent. Les liens inter-organisationnels n'existent qu'au travers des liens individuels. Autrement dit, deux organisations collaborent à partir du moment où un membre de chacune collabore avec l'autre. Nous nous permettons de poser ces hypothèses, car ici l'objet de notre travail porte sur les structures des liens de collaboration entre individus. L'affiliation d'un individu à une organisation peut être vue comme un effet endogène propre à un individu.

Le graphe des individus nous suffit pour représenter les relations. Nous attachons à chaque nœud un attribut définissant l'affiliation du collaborateur. Un exemple de ce graphe est représenté par la Figure 3.1. Les individus sont représentés par les nœuds et les liens de collaboration, par les arêtes. Les nœuds sont colorés en fonction de leur affiliation, une couleur étant associée à chaque organisation.

Le graphe de collaboration décrit par la Figure 3.1 constitue notre graphe de départ, notre observation du monde réel. Les graphes sociaux mettent en avant les interactions entre les individus. Or ici, nous nous intéressons davantage à la structure des liens de collaboration, c'est-à-dire, aux liens eux-mêmes et à la façon dont ils interagissent avec les autres liens. Pour ce faire, nous nous appuyons sur une représentation duale du graphe de collaboration, le graphe des arêtes (*Line Graphe* dans la littérature anglaise).

À partir d'un graphe G, nous construisons le graphe des arêtes $L(G)$ de la manière suivante. L'ensemble des arêtes du graphe de départ G, correspondent aux nœuds du

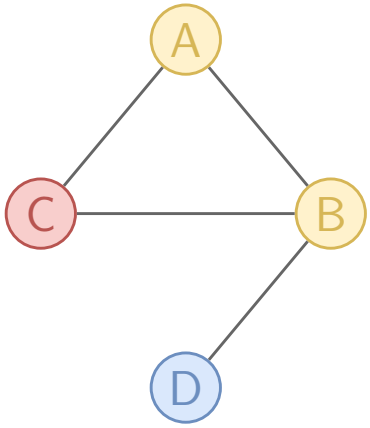


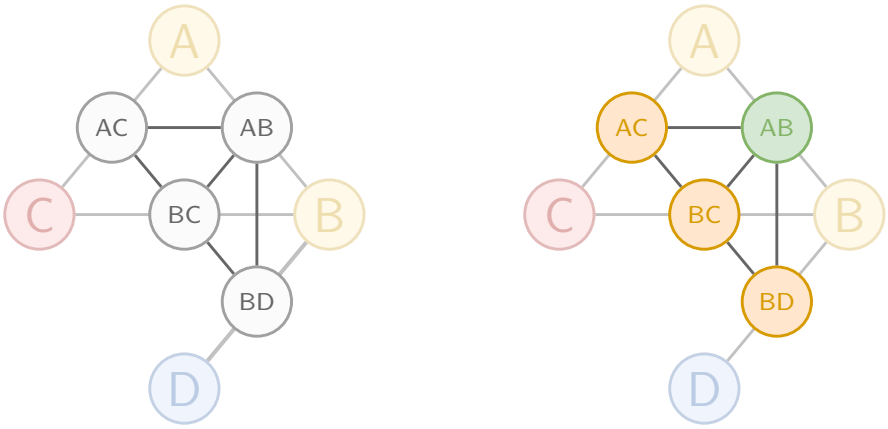
FIGURE 3.1 – Exemple d'un graphe de collaboration impliquant quatre individus : A, B, C et D. L'affiliation de chaque individu est représentée par une couleur. Les couleurs correspondent respectivement à : l'organisation 1 ●, l'organisation 2 ● et l'organisation 3 ●.

graphe des arêtes $L(G)$. Deux arêtes dans le graphe G sont dites "connectées" si elles ont une extrémité commune. Par exemple : les arêtes (i, j) et (j, k) sont connectées (par le nœud j), alors que (i, j) et (k, l) n'ont pas d'extrémité commune, elles ne sont donc pas connectées. Deux arêtes connectées dans le graphe G, correspondent à deux nœuds adjacents dans le graphe $L(G)$. Nous avons déjà eu l'occasion d'évoquer le graphe des arêtes dans la Section 2.2.2, lorsque nous avons présenté les dépendances de Markov. Le graphe de dépendance sous les hypothèses de Markov est également un graphe des arêtes. Comme le suggère HARARY [44], le graphe des arêtes $L(G)$ peut être considéré comme une transformation du graphe G. À un graphe G est associé un graphe des arêtes $L(G)$.

Le graphe des arêtes a déjà été utilisé dans la littérature pour différents propos. Dans le cadre des problèmes de détection de communautés, il peut arriver qu'un nœud soit à l'intersection de plusieurs communautés. Une solution permettant de prendre en compte ce chevauchement de communauté est de non plus considérer les nœuds du graphe, mais plutôt d'estimer les partitions entre les arêtes [27]. Cette approche est également utilisée pour l'étude de la propagation des maladies sexuellement transmissibles. Le graphe des arêtes permet entre autres de comparer le scénario où les partenaires entretiennent des relations monogames et le scénario où les partenaires entretiennent des relations concurrentes [70].

Reprenons l'exemple du graphe de collaboration illustré par la Figure 3.1. À partir de ce graphe, et d'après la définition donnée ci-dessus, nous sommes en mesure de construire le graphe des arêtes. La Figure 3.2(a) fournit ainsi la structure relationnelle des arêtes de notre exemple de graphe de collaboration. Nous ajoutons à cette structure une information sur la nature des liens de collaboration. Nous distinguons deux types de lien : les liens inter-organisationnels et les liens intra-organisationnels. Les liens inter-organisationnels désignés en orange ● dans la Figure 3.2(b) relient deux collaborateurs n'appartenant pas à la même organisation dans le graphe de collaboration. Dans le graphe de collaboration, les arêtes : (A, C), (B, C) et (B, D) sont liens inter-organisationnels puisqu'elles connectent des individus n'appartenant pas à la même organisation. Les liens intra-organisationnels

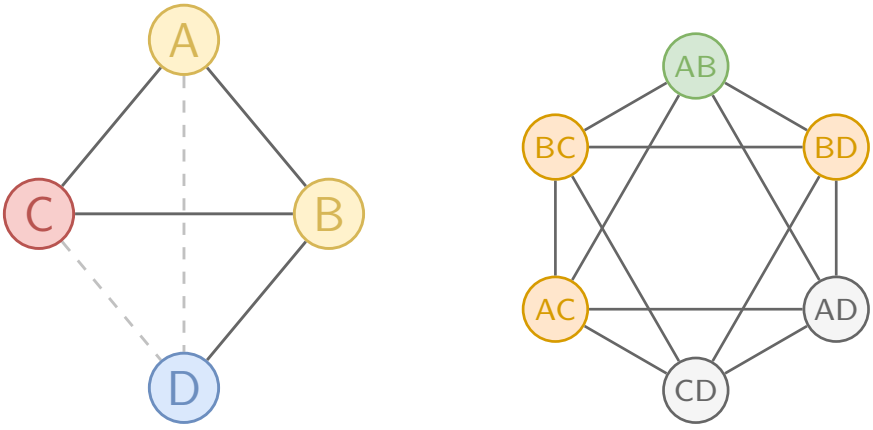
désignés en vert ● dans la Figure 3.2(b) relient deux collaborateurs appartenant à la même organisation dans le graphe de collaboration. Ici, dans le graphe de collaboration, seule l'arête (A, B) est un lien intra-organisationnel, puisque les deux collaborateurs situés de part et d'autre de l'arête sont issus de la même organisation.



(a) Graphe des arêtes (b) Graphe des arêtes coloré

FIGURE 3.2 – Construction d'un graphe des arêtes à partir d'un graphe de collaboration (Figure 3.1). Les nœuds verts ● représentent des liens intra-organisationnels et les nœuds orange ● représentent des liens inter-organisationnels.

Pour obtenir une représentation complète, nous devons en outre considérer les liens non existants dans le graphe observé (Figure 3.1), mais qui auraient pu se produire. Nous avons représenté ces liens en transparence dans la Figure 3.3(a). En considérant ces liens, nous obtenons le graphe des arêtes illustré par la Figure 3.3(b). Dans notre exemple, les liens non existants sont les arêtes (C, D) et (A, D) désignés par la couleur grise ● dans la Figure 3.3(b).



(a) Graphe social complet (b) Graphe des arêtes

FIGURE 3.3 – Prise en compte des liens non-existants dans le graphe observé. Les liens non-existants sont colorés en gris ● dans la Figure 3.3(b).

La Figure 3.3(b) constitue à présent notre représentation du monde réel. C'est cette représentation que nous considérerons dans la suite de notre approche. En substance, notre graphe des arêtes (Figure 3.3(b)) est un graphe dont les nœuds sont labélisés (ou colorés). Sa définition est donc légèrement différente de ce que nous avons pu évoquer jusqu'à présent.

Comme nous l'avons déjà évoqué en préambule du Chapitre 2, l'occurrence d'une arête dans un graphe social résulte d'un processus stochastique. Les nœuds du graphe présenté par la Figure 3.3(b), peuvent alors être considérés comme des variables aléatoires dont les valeurs sont définies par un ensemble de labels, en l'occurrence ici : *non_existant, intra-organisationnel, inter-organisationnel*. Ces trois labels représentent les trois états possibles d'une arête dans les graphes de collaboration observés.

Plus formellement, nous pouvons définir le graphe des arêtes (Figure 3.3(b)) de la manière suivante. Soit $\mathcal{G}$ le graphe des arêtes considéré, avec $\mathcal{V}$ l'ensemble des nœuds (ou de leur indice) tel que :

$$\mathcal{V} = \{1, \ldots, N\},$$

et $\mathcal{E}$ l'ensemble des arêtes :

$$\mathcal{E} = \{e_{ij} | i \sim j, \forall i, j \in \mathcal{V}\},$$

où $\sim$ est la relation de voisinage entre deux nœuds, c'est-à-dire que deux nœuds sont directement connectés par une arête. Par extension, le voisinage V_i d'un nœud i correspond à l'ensemble des autres nœuds ($i \notin V_i$) qui sont directement connectés à ce dernier :

$$V_i = \{j | i \sim j, \forall j \in \mathcal{V}\}.$$

On appelle système de voisinage V, l'ensemble des voisinages :

$$V = \{V_i | i \in \mathcal{V}\}.$$

Les nœuds du graphe sont étiquetés par un label $l \in \mathcal{L}$, où $\mathcal{L}$ définit l'ensemble des labels :

$$\mathcal{L} = \{\ell_1, \cdots, \ell_m\}.$$

L'état d'un nœud i, son étiquetage (le label qui lui est associé) est défini par la variable aléatoire Y_i, dont la réalisation se note : $Y_i = y_i$. Généralisé à l'ensemble du graphe $\mathcal{G}$, la réalisation de tous les nœuds se note quant à elle $Y = y$. Enfin, nous utilisons la notation $Y_{-i} = y_{-i}$, pour désigner la réalisation de tous les nœuds du graphe $\mathcal{G}$ à l'exception du nœud i.

Dans un cadre général, nous pouvons définir l'espace des configurations possibles Ω tel que : $\Omega = \mathcal{L}^{\mathcal{V}}$. Dans notre situation, il est nécessaire d'appliquer un conditionnement sur les labels. Le graphe des arêtes doit être cohérent vis-à-vis de l'observation. Les affiliations des collaborateurs sont immuables. De ce fait, le lien AB illustré par la Figure 3.3(a), peut être soit labélisé *non_existant* soit *intra-organisationnel*. En aucun cas ce lien ne peut-être labélisé *inter-organisationnel*, cela conduirait à une incohérence étant donné que ce lien connecte deux collaborateurs issus de la même organisation. Ce qui revient à dire que tous les labels ne peuvent pas être appliqués à tous les types de nœuds du graphe des arêtes. Nous pouvons considérer deux types de liens dans le graphe de collaboration et donc de nœuds dans le graphe des arêtes :

— les nœuds représentant des liens intra-organisationnels : $\mathcal{V}_{intra}$
— les nœuds représentant des liens inter-organisationnels : $\mathcal{V}_{inter}$

De telle sorte que :

$$\mathcal{V}_{intra} \cap \mathcal{V}_{inter} = \emptyset,$$
$$\mathcal{V} = \{\mathcal{V}_{intra}, \mathcal{V}_{inter}\}.$$

De la même manière, nous pouvons distinguer deux sous-ensembles de labels :

— les labels applicables aux nœuds représentant des liens intra-organisationnels : $\mathcal{L}_{intra} = \{non\_existant, intra - organisationnel\}$
— les labels applicables aux nœuds représentant des liens inter-organisationnels : $\mathcal{L}_{inter} = \{non\_existant, inter - organisationnel\}$

Étant donné le conditionnement sur les labels, l'espace des configurations possibles Ω se définit :

$$\Omega = \{\mathcal{L}_{intra}^{\mathcal{V}_{intra}}, \mathcal{L}_{inter}^{\mathcal{V}_{inter}}\}.$$

À travers ce formalisme nous retrouvons sensiblement la définition du graphe stochastique donnée dans le Chapitre 2. À la différence près que nous considérons les nœuds comme des variables aléatoires et non les arêtes. Ce changement de perspective est induit par la transformation du graphe initial en graphe des arêtes.

Nous empruntons ce formalisme aux *Markov Random Field*s (MRFs) [32] ou *champs de Markov* en français. Les MRFs décrivent un formalisme mathématique notamment utilisé en traitement de l'image [63, 106] et peuvent plus généralement s'appliquer à la description de processus spatiaux [7, 8, 29].

Les MRFs décrivent un ensemble discret de *sites* ayant chacun un état à valeur dans un ensemble fini d'états. Ce formalisme permet par exemple de représenter une grille de pixels. Les sites correspondent aux pixels, dont l'ensemble des valeurs diffèrent en fonction du type d'image traité. Dans la configuration la plus simple, les pixels peuvent avoir deux valeurs 0 ou 1, blanc ou noir (ce qui correspond à une image à deux niveaux : noir et blanc). Dans notre situation, les sites sont les nœuds dont les valeurs sont définies par l'ensemble $\mathcal{L}$.

Les champs de Markov, comme le nom le suggère, reposent sur les hypothèses de dépendance de Markov. Nous pouvons, encore une fois, faire un parallèle avec les notions abordées dans le Chapitre 2 (sous-section 2.2.2). Dans un cadre plus général, sous les hypothèses de Markov, la réalisation d'un site, ne dépend que de son voisinage (étant donné un système de voisinage). Rapporté à notre situation, l'étiquetage d'un nœud ne dépend que des labels des nœuds appartenant à son voisinage.

Sans hypothèse préalable, la probabilité de réalisation d'un nœud i se note :

$$P(Y_i = y_i | Y_{-i} = y_{-i}).$$

L'hypothèse de Markov, nous permet de réduire considérablement le calcul de cette probabilité ainsi :

$$P(Y_i = y_i | Y_{-i} = y_{-i}) = P(Y_i = y_i | Y_{V_i} = y_{V_i}).$$

Comme nous avons déjà pu l'évoquer, les hypothèses de dépendances nous permettent *in fine* de simplifier les calculs notamment pour la simulation est l'estimation de la (pseudo) vraisemblance (section 2.3).

Il est important de souligner que le graphe des arêtes d'un graphe complet, n'est pas complet. Nous pouvons aisément le constater dans notre exemple (Figure 3.3), en remarquant par exemple que le nœud AC n'est pas connecté au nœud BD dans la Figure 3.3(b). Harary [44] propose une formulation permettant de déterminer le nombre d'arêtes dans le graphe des arêtes en fonction du nombre d'arêtes dans le graphe initial et du degré de chaque nœud. À partir de cette formulation, nous pouvons généraliser ce constat à des graphes de toute taille. Le choix des hypothèses de Markov permet effectivement une simplification. Cela n'aurait pas été le cas si le graphe des arêtes était complet, puisque le voisinage d'un nœud comprendrait l'ensemble des autres nœuds du graphe.

Dans cette section, nous avons présenté une autre façon de considérer le graphe social, à travers une représentation duale : le graphe des arêtes. Ce graphe nous permet de mettre en avant directement la structure et les interactions entre les arêtes. C'est là notre principal centre d'intérêt, puisque nous étudions la structure des liens de collaboration. Chaque arête est étiquetée en fonction de son état dans le graphe social. À travers ce changement de perspective, nous pouvons considérer la réalisation d'une arête dans le graphe social comme un étiquetage de nœud dans le graphe des arêtes. Autrement dit, le processus de réalisation d'un lien de collaboration s'apparente à un processus spatial. La formulation et le formalisme diffèrent en apparence, mais la base reste inchangée. À partir de cette formulation, nous pouvons à présent définir un modèle décrivant les phénomènes que nous voulons mesurer au sein de ces observations (sous forme d'un graphe des arêtes).

3.3 Description du modèle de Potts [8]

Le graphe des arêtes que nous avons introduit dans la section précédente constitue notre observation. Nous l'avons défini comme un champ de Markov. Les nœuds sont des sites, qui peuvent être labélisés et dont la réalisation (l'attribution d'un label à un site) dépend exclusivement de l'état de leur voisinage. D'après le théorème d'Hammersley-Clifford [41] introduit pas Besag [8], la fonction de probabilité d'un champ de Markov Y se définit au travers d'une distribution de Gibbs :

$$p(Y = y|\theta) = \frac{\exp(U(y|\theta))}{\kappa(\theta)}$$

Où :

- θ est le vecteur des paramètres
- $U(\cdot|\cdot)$ est la fonction d'énergie
- $\kappa(\theta)$ est la constante de normalisation

Nous retrouvons la forme exponentielle des ERGMs. La famille de modèles ERGMs introduit par Frank et Strauss [28] est en réalité une application des MRF [8] à l'étude des réseaux sociaux. La fonction d'énergie $U(\cdot|\cdot)$ dépend des structures locales considérées. Elle se définit par la somme pondérée des statistiques des structures locales considérées (2.1). La constante de normalisation correspond à l'ensemble des configurations possibles sur Ω (2.2).

Le modèle de Potts est un des modèles les plus connus pour décrire des interactions dans l'espace, que ce soit entre des pixels, les nœuds d'un graphe ou des matériaux ferromagnétiques. Ce modèle est d'ailleurs une généralisation du modèle physique de Ising qui décrit les interactions entre des particules chargées soit positivement soit négativement (comme des spins par exemple) et qui sont soumises à un champs extérieur. Étant donné un ensemble discret de sites, la distribution de probabilité du modèle de Potts se définit comme :

$$p(Y = y|\theta) = \frac{1}{\kappa(J)} \exp(J \sum_{i \sim k} (\delta_{y_i y_k} - 1))$$

L'opérateur de Kronecker $\delta_{y_i y_k}$ retourne un si y_i et y_k ont le même label et zéro sinon. Ainsi, seuls les interactions entre des sites ayant des labels différents seront comptabilisées. Le paramètre J permet de réguler les interactions entre des sites de valeurs différentes. Lorsque les sites ne peuvent prendre que deux valeurs différentes de label, le modèle de Potts se rapporte à un modèle de Ising.

Le modèle que nous proposons est une extension du modèle de Potts. Ce modèle s'appuie sur la définition du champ de Markov présentée par la section 3.2. Pour alléger le formalisme, nous associons à chaque label un entier. Nous désignons les labels *non-existant, intra-organisationnel, inter-organisationnel* respectivement par les entiers 0, 1 et 2. La distribution de probabilité de ce modèle s'écrit ainsi :

$$\begin{aligned} p(Y = y|\theta) = \frac{1}{\kappa(\theta)} \exp\Big[& \theta_{11} \sum_{i \sim j} \mathbb{1}\{y_i = 1, y_j = 1\} \\ & + \theta_{12} \sum_{i \sim j} \mathbb{1}\{y_i = 1, y_j = 2\} \\ & + \theta_{22} \sum_{i \sim j} \mathbb{1}\{y_i = 2, y_j = 2\} \Big]. \end{aligned} \tag{3.1}$$

La fonction indicatrice $\mathbb{1}\{\cdot\}$ retourne un si la condition en entrée est vraie et zéro sinon.

Le modèle que nous proposons considère trois interactions différentes. Nous les retrouvons dans le vecteur des statistiques suffisantes $t(y)$:

$$\begin{aligned} t(y) &= [t_{11}(y), t_{12}(y), t_{22}(y)] \\ &= \left[\sum_{i \sim j} \mathbb{1}\{y_i = 1, y_j = 1\}, \sum_{i \sim j} \mathbb{1}\{y_i = 1, y_j = 2\}, \sum_{i \sim j} \mathbb{1}\{y_i = 2, y_j = 2\} \right]. \end{aligned} \tag{3.2}$$

La condition d'activation de la fonction $\mathbb{1}\{y_i = 1, y_j = 1\}$ est vérifiée lorsque deux nœuds adjacents dans le graphe des arêtes sont labélisés *intra-organisationnel*. Dans le graphe de collaboration, cette disposition correspond à la configuration où un individu collabore avec deux membres de l'organisation à laquelle il est affilié. La statistique t_{11} comptabilisant ce type d'interaction quantifie la propension pour un individu à collaborer avec d'autre membre de la même organisation. La condition d'activation de la fonction $\mathbb{1}\{y_i = 1, y_j = 2\}$ est quant à elle vérifiée lorsqu'un nœud labélisé *intra-organisationnel* est connecté à un nœud labélisé *inter-organisationnel*. Cette disposition dans le graphe des arêtes illustre une configuration où un individu est à la fois connecté à un membre

TABLE 3.1 – Représentation graphique des interactions contrôlées par le modèle. Les différents types de liens sont désignés de la manière suivante : $y_i = 1$ ● (intra-organisational), $y_i = 2$ ● (inter-organisational). Les trois organisations sont associées chacune à une couleur : jaune ●, rouge ● et bleu ●.

Désignation	Graphe des arêtes	Graphe de collaboration
Intra ($1 \leftrightarrow 1$)		
Hub ($1 \leftrightarrow 2$)		
Inter ($2 \leftrightarrow 2$)		

de son organisation et à un membre d'une autre organisation. En d'autres termes, la statistique t_{12} comptabilise la tendance à coopérer à la fois avec un membre de son organisation et un membre extérieur, jouant ainsi le rôle de *hub*. Enfin, la fonction indicatrice $\mathbb{1}\{y_i = 2, y_j = 2\}$ est activée lorsque deux nœuds adjacents sont labélisés *inter-organisationnel*. Ce qui se traduit dans le graphe de collaboration par une configuration où un individu collabore avec deux autres individus externes, issus d'une ou plusieurs organisations différentes. La statistique associée t_{22} quantifie la propension pour un membre d'une organisation à collaborer avec des membres extérieurs. La Table 3.1 résume notre propos et représente les interactions dans le graphe des arêtes et les configurations correspondantes dans le graphe de collaboration.

Dans les trois interactions que nous venons de mentionner, nous ne prenons pas en compte le label *non-existant*. Ceci tient du fait que les étiquetages *inter/intra-organisationnel*[5] et *non-existant* sont intrinsèquement liés. Le passage d'un nœud inter-organisationnel en *non-existant* ($y_i = 2 \rightarrow y_i = 0$) s'exprime comme une diminution du nombre de nœuds inter-organisationnels. En somme, considérer les intéractions impliquant des labels *non-existant* reviendrait à sur-contraindre le modèle.

Les paramètres appartenant au vecteur $\theta = [\theta_{11}, \theta_{12}, \theta_{22}]$ pondèrent les statistiques du vecteur des statistiques suffisantes $t(y)$. Si $\theta_{ij} > 0$ le modèle aura tendance à favoriser la configuration comptabilisée par la statistique t_{ij}. Inversement, si $\theta_{ij} < 0$ le modèle aura tendance à pénaliser la configuration associée à la statistique t_{ij} et diminuera sa valeur. Pour le cas limite $\theta_{ij} = 0$, le paramètre ne contrôle pas l'occurrence de la configuration associée à t_{ij}. Dans ce cas, il n'y a pas de tendance marquée. L'occurrence de la configuration est "purement" aléatoire, comparable à un lancer de pièces.

Tout comme le modèle de Potts, le modèle que nous proposons permet de décrire une situation observée. En l'occurrence ici, nous considérons des collaborations au travers du graphe des arêtes. Nous pouvons extrapoler des trois interactions spécifiées par le modèle, des caractéristiques structurelles liées au graphe de collaboration observé. Par exemple si nous constatons d'importantes taches (regroupement de nœuds verts) $1 \leftrightarrow 1$ dans le graphe des arêtes, cela signifie structurellement que les individus concernés ont tendance à collaborer avec les membres de leur équipe. Ce raisonnement s'applique également aux

5. En fonction du nœud considéré

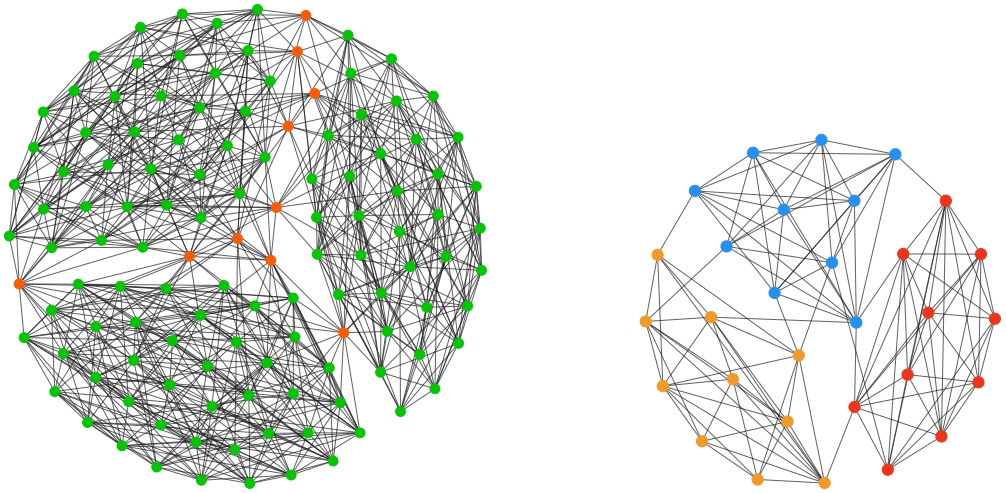
deux autres interactions. Des tâches $2 \leftrightarrow 2$ renseigneraient sur la tendance inverse, celle de collaborer exclusivement avec des acteurs extérieurs. Enfin, si au lieu de constater des aplats uniformes, nous observons une configuration hétérogène des couleurs 1 et 2 ($1 \leftrightarrow 2$), à la façon d'une image bruitée, nous pouvons déduire que la population étudiée à tendance à collaborer à la fois avec ces homologues, mais aussi avec des acteurs extérieurs.

Nous proposons de fixer arbitrairement les paramètres du modèle de façon à reproduire les trois tendances marquées que nous avons développées dans le précédent paragraphe. Pour les trois exemples qui vont suivre, nous considérons un graphe de collaboration comportant trois organisations ayant chacune 10 membres (soit en tout, 30 collaborateurs). Il est important de garder à l'esprit que nous ne manipulons pas directement le graphe de collaboration, mais le graphe des arêtes. Initialement, tous les nœuds du graphe des arêtes sont désactivés (associées au label *non-existant*). Une fois les paramètres fixés, nous simulons le modèle à l'aide d'un échantillonneur de Gibbs (présenté dans la sous-section 2.3.1). L'échantillonneur de Gibbs est exécuté sur 500 itérations, et produit 500 configurations de graphe différentes. En faisant la moyenne des statistiques suffisantes de chacune de ces configurations, nous obtenons les statistiques suffisantes du modèle appliqué au graphe de collaboration considéré [6]. Les statistiques suffisantes des échantillons simulés par l'échantillonneur de Gibbs sont détaillées dans l'Annexe A.

Le premier exemple tend à favoriser l'occurrence de l'interaction $1 \leftrightarrow 1$. Pour ce faire, nous avons sélectionné les paramètres suivants : $\theta = [0.15, -0.2, -0.2]$. Nous avons attribué à la configuration $1 \leftrightarrow 1$ un paramètre positif ($\theta_{11} = 0.15$). Pour accentuer le contraste, nous avons pénalisé les deux interactions restantes en attribuant à chacune d'elle un paramètre négatif ($\theta_{12} = \theta_{22} = -0.2$). La moyenne des statistiques suffisantes $\bar{t}(y) = [780, 162, 7]$ renseigne sur la prépondérance de l'interaction $1 \leftrightarrow 1$. La Figure 3.4(a) représente le dernier graphe des arêtes simulé par l'échantillonneur de Gibbs. Pour améliorer la lisibilité, nous avons retiré tous les nœuds labélisés *non-existant*. C'est une représentation relativement fidèle de ce que peut être une réalisation du modèle sachant ces paramètres. En effet, ces statistiques suffisantes ($t(y_{last}) = [740, 144, 6]$) sont relativement proches de la moyenne des statistiques suffisantes ($\bar{t}(y)$) obtenue à partir de tous les échantillons. Nous retrouvons dans la Figure 3.4(a) d'importants "amas" de nœuds verts représentant des liens intra-organisationnels. Cette configuration maximise le nombre d'interactions $1 \leftrightarrow 1$ et semble cohérent vis-à-vis du choix des paramètres. La Figure 3.4(b) illustre le graphe de collaboration correspondant à la Figure 3.4(a). Nous remarquons que les membres d'une même organisation sont fortement interconnectés. En revanche, les liens inter-organisationnels sont très faiblement présents. En d'autres termes, les collaborations sont fortement tournées vers l'intérieur. Nous procédons de la même manière pour les deux autres exemples.

Le second exemple précédemment évoqué a tendance quant à lui à avantager les occurrences de l'interaction $2 \leftrightarrow 2$. Nous avons sélectionné les paramètres du modèle comme suit : $\theta = [-0.1, -0.1, 0.08]$. Les statistiques suffisantes du modèle simulé sont $\bar{t}(y) = [1, 112, 5066]$. Il y a ici une prépondérance d'interactions $2 \leftrightarrow 2$. Nous pouvons l'observer dans la Figure 3.5(a) par la présence majoritaire (presque totale) de nœuds orange, représentant les liens inter-organisationnels. Nous constatons dans la Figure 3.5(b)

6. Au travers du grahe des arêtes



(a) Graphe des arêtes (b) Graphe de collaboration

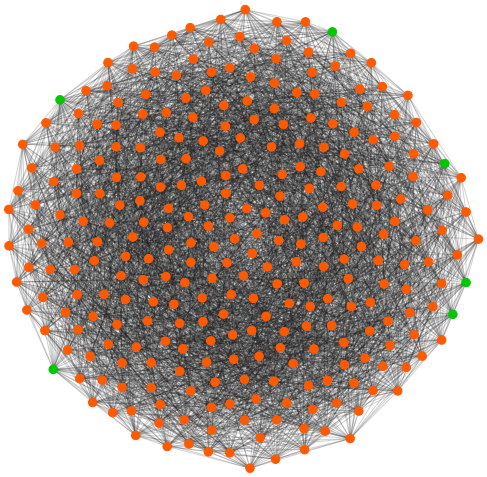
FIGURE 3.4 – Exemple d'une réalisation du modèle paramétré par $\theta = [0.15, -0.2, -0.2]$. Les statistiques suffisantes sont $t(y) = [740, 144, 6]$. [7]

que les frontières entre les trois organisations ne sont plus visibles. Les individus ont tendance à tisser (exclusivement) des liens avec des acteurs n'appartenant pas à la même organisation. Ce constat est d'autant plus marquant si nous regardons avec attention les liens entretenus par un individu (Figure 3.5(c)). Son voisinage est principalement constitué d'acteurs extérieurs. La tendance est inverse, la collaboration est tournée vers l'extérieur.

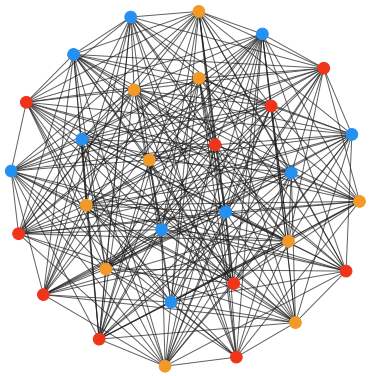
Enfin, le troisième exemple implique un grand nombre de hubs ($1 \leftrightarrow 2$), les paramètres sélectionnés sont : $\theta = [-0.2, 0.15, -0.25]$. Les statistiques suffisantes résultantes de la simulation ($\bar{t}(y) = [257, 708, 356]$) confirment la prédominance d'interactions $1 \leftrightarrow 2$. Contrairement aux Figures 3.4(a) et 3.5(a), la Figure 3.6(a) présente une distribution hétérogène des deux types de nœuds représentant les liens intra-organisationnels (en vert) et les liens intra-organisationnels (en orange). Cette disposition amplifie les interactions de type hub. Nous observons dans la Figure 3.6(b) qu'il n'y a pas non plus de frontière marquée entre les organisations. Les individus tissent des liens avec d'autres sans qu'il y ait de réelles discriminations sur l'affiliation. Nous constatons d'autant plus ce phénomène lorsque nous observons le voisinage d'un des individus (Figure 3.6(c)). Le voisinage est constitué à la fois d'individus appartenant à la même organisation et d'individus extérieurs. Dans ces circonstances, chaque individu joue le rôle de "hub" dans le sens où il maintient à la fois des liens avec les membres de son organisation (contrairement à l'exemple précédent) et des membres affiliés à d'autres organisations.

Nous venons de voir qu'à partir d'un modèle dont les paramètres sont fixés nous pouvons simuler des observations. Ces observations sont des graphes des arêtes représentant la distribution des liens au sein d'une collaboration. Nous avons présenté trois exemples de paramétrisation mettant en exergue les trois propriétés du modèle. Le parallèle avec le graphe de collaboration souligne les implications de ses propriétés sur

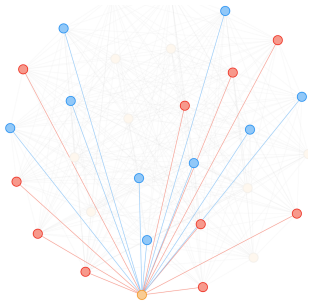
7. Les nœuds labélisés *non-existant* ont été retirés du graphe des arêtes. Les liens intra-organisationnels et inter-organisationnels sont respectivement représentés par des nœuds verts ● et orange ● dans le graphe des arêtes. Les trois organisations sont associées chacune à une couleur : jaune ●, rouge ● et bleu ●.



(a) Graphe des arêtes



(b) Graphe de collaboration

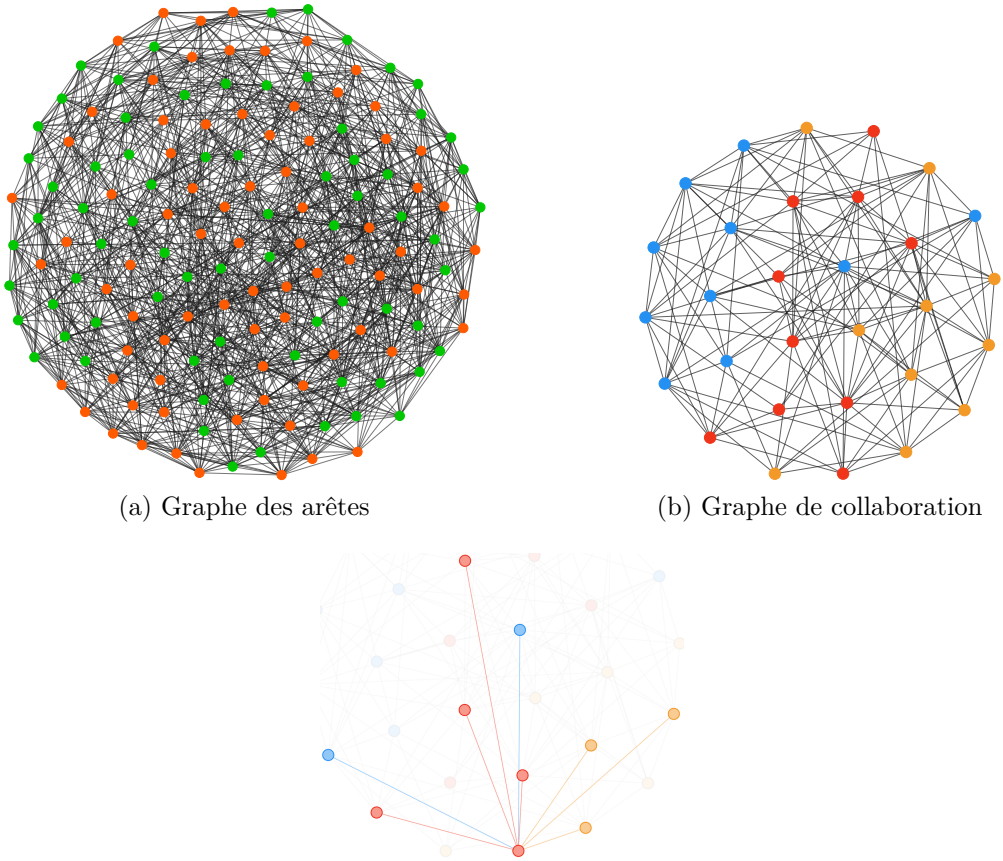


(c) Voisinage d'un nœud

FIGURE 3.5 – Exemple d'une réalisation du modèle paramétré par $\theta = [-0.1, -0.1, 0.08]$. Les statistiques suffisantes sont $t(y) = [1, 221, 5044]$ [7].

la structure de la collaboration. Le modèle que nous proposons est élémentaire et repose uniquement sur les hypothèses de dépendance de Markov. Cependant, il est pertinent au regard du contexte. Les interactions comptabilisées par les statistiques suffisantes (Table 3.1) impliquent à chaque fois trois collaborateurs dans le graphe de collaboration. Le collaborateur qui se trouve au centre joue alors le rôle de relais entre le collaborateur de gauche et celui de droite. Cette notion de relais est fondamentale dans les réseaux pair-à-pair. Nous la retrouvons dans de nombreux mécanismes comme les DHT, les protocoles de propagation épidémiques, les réseaux sociaux décentralisés... En somme, à travers notre modèle, nous pouvons mesurer les propensions d'occurrence de ces relais au sein et entre les organisations.

La forme générale du modèle et plus particulièrement la constante de normalisation $\kappa(\theta)$, ne permettent pas d'estimer directement le modèle à partir d'une observation. Les outils d'inférence comme les méthodes de Monte-Carlo par chaîne de Markov peuvent se révéler coûteux surtout pour le calcul du log de vraisemblance par échantillonnage préférentiel (*importance sampling*). D'autant que la qualité de l'estimation dépend du choix de paramètres initiaux. Nous avons déjà eu l'occasion discuter de ces aspects dans la section 2.3. Le problème reste le même avec le modèle que nous proposons et le choix d'un



(a) Graphe des arêtes

(b) Graphe de collaboration

(c) Voisinage d'un nœud

FIGURE 3.6 – Exemple d'une réalisation du modèle paramétré par $\theta = [-0.2, 0.15, -0.25]$. Les statistiques suffisantes sont $t(y) = [314, 773, 377]$ [7].

outil d'inférence stable et suffisamment peu couteux est une question ouverte [34, 33, 35]. Dans la section suivante (section 3.4), nous proposons d'utiliser une méthode alternative, un algorithme d'inférence bayésienne : ABC Shadow [96]. L'inférence bayésienne est plus informative et offre la possibilité de mieux exploiter les résultats au travers notament de tests statistiques comme nous aurons l'occasion de le voir dans le Chapitre 4. Mais elle ne nous permet pas directement de nous soustraire au problème de la constante de normalisation. Des approches [6, 67, 72] sont proposées dans l'état de l'art pour mitiger ce problème. Nous les présentons et les mettons en perspective avec l'algorithme ABC Shadow dans la section suivante.

3.4 Inférence bayésienne [5]

La loi de Bayes, du révérend Thomas Bayes, nous permet de définir la distribution *a posteriori* des paramètres. Cette distribution se note $p(\theta|y)$ et peut être vue comme la distributions des modèles susceptibles d'expliquer l'observation y. La formule de la probabilité conditionnelle établie par la loi de Bayes définie la distribution *a posteriori* de

cette manière :

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} \times p(\theta). \tag{3.3}$$

La distribution du modèle $p(y|\theta)$ (ou vraisemblance) définit la distribution des observations étant donnés les paramètres du modèle (section 3.3). Dans notre cas, elle prend la forme d'une distribution de Gibbs :

$$p(y|\theta) = \frac{\exp(U(y|\theta))}{\kappa(\theta)}. \tag{3.4}$$

La distribution $p(\theta)$ représente la connaissance a priori que l'on possède sur les paramètres θ, elle est définie sur l'espace des paramètres Θ. Enfin, $p(y)$ est la distribution marginale des observations sur l'ensemble des paramètres possibles et peut se définir ainsi :

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)d\theta.$$

On remarque que $p(y)$ ne dépend que des membres placés au numérateur de (3.3). Nous pouvons ainsi écrire (3.3) sous sa forme proportionnelle :

$$p(\theta|y) \propto \frac{\exp(U(y|\theta))}{\kappa(\theta)} \times p(\theta), \tag{3.5}$$

où le symbole $\propto$ signifie "proportionnel à". En substance, l'objectif de l'inférence bayésienne est d'estimer la distribution *a posteriori* des paramètres sachant une observation donnée.

L'inférence bayésienne et le maximum de vraisemblance que nous avons présentés dans la section 2.3 sont deux approches qui considèrent toutes les deux un modèle paramétrique. Ces deux approches s'appuient néanmoins sur deux philosophies différentes. D'une part, le maximum de vraisemblance ne s'appuie pas sur une connaissance a priori des paramètres. D'autre part pour cette approche, le modèle le plus probable est proposé comme explication de l'observation. L'approche bayésienne repose quant à elle sur une connaissance a priori des paramètres. De plus, cette approche considère tous les modèles possibles comme une potentielle explication de l'observation. La pertinence de ces modèles pour expliquer l'observation est pondérée par la distribution *a posteriori*. L'inférence bayésienne est plus informative et peut-être vue comme une généralisation de l'approche par maximum de vraisemblance. Néanmoins, lorsque $p(\theta)$ suit une loi uniforme (n'est pas informative), les deux approches sont strictement équivalentes. À partir de la distribution *a posteriori*, nous pouvons calculer le *Maximum a posteriori* (MAP) :

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|y),$$

la configuration du modèle la plus représentée par la distribution et qui peut être considérée comme un estimateur du maximum de vraisemblance.

Les méthodes de Monte-Carlo, telles que Metropolis-Hastings ou l'échantillonneur de Gibbs (sous-section 2.3.1) permettent d'échantillonner directement la distribution *a posteriori* des paramètres. L'algorithme Metropolis-Hastings que nous avons présenté

dans la sous-section 2.3.1 propose à chaque itération une nouvelle valeur θ' à partir de la distribution de proposition $q(\theta \to \theta')$. Cette nouvelle valeur est conservée ou rejetée en fonction du ratio d'acceptation $\alpha_{\theta\to\theta'}$. Cependant, son évaluation présuppose le calcul du ratio des constantes de normalisation $(\kappa(\theta)/\kappa(\theta'))$:

$$\begin{aligned}\alpha_{\theta\to\theta'} &= \min\left[1, \frac{p(\theta'|y)q(\theta'\to\theta)}{p(\theta|y)q(\theta\to\theta')}\right], \\ &= \min\left[1, \frac{\exp(U(y|\theta'))p(\theta')q(\theta'\to\theta)}{\exp(U(y|\theta))p(\theta)q(\theta\to\theta')} \times \frac{\kappa(\theta)}{\kappa(\theta')}\right].\end{aligned} \tag{3.6}$$

Dans ces conditions, la distribution *a posteriori* ne peut pas être échantillonnée avec un Metropolis-Hastings standard. De même, l'échantillonneur de Gibbs nécessite le calcul de la probabilité conditionnelle permettant de mettre à jour θ $(p(\cdot|\theta, y))$, qui n'est pas non plus disponible sous une forme calculable. Ces deux approches ne sont pas directement utilisables pour échantillonner la distribution *a posteriori* étant donnée la forme du modèle utilisé.

Une première solution proposée par Møller *et al.* [67] consiste à utiliser une variable auxiliaire x qui partage le même espace d'état que l'observation y. Møller *et al.* [67] proposent une variante du Metropolis-Hastings standard. L'objectif de cette stratégie est de supprimer du ratio d'acceptation (3.6) le ratio des constantes de normalisation. De manière semblable à (3.5), la distribution *a posteriori* avec la variable auxiliaire s'écrit :

$$p(\theta, x|y) \propto p(x|\theta, y) \times \frac{\exp(U(y|\theta))}{\kappa(\theta)} \times p(\theta).$$

Chaque pas du Metropolis Hastings met à jour à la fois θ et x. La distribution de proposition $q((\theta, x) \to (\theta', x'))$ porte alors sur le couple (θ, x) et peut se décomposer comme suit :

$$q((\theta, x) \to (\theta', x')) = q_1(\theta'|\theta, x)q_2(x'|\theta', \theta, x).$$

Il en est de même pour le ratio d'acceptation qui prend la notation $\alpha_{(\theta,x)\to(\theta',x')}$:

$$\alpha_{(\theta,x)\to(\theta',x')} = \min\left[1, H_{(\theta,x)\to(\theta',x')}\right].$$

Les deux distributions de proposition q_1 et q_2 peuvent être choisies de manière à rendre le calcul de $H_{(\theta,x)\to(\theta',x')}$ faisable. Tout d'abord, nous pouvons définir q_1 comme indépendante de x :

$$q_1(\theta'|\theta, x) = q_1(\theta'|\theta).$$

De plus, q_1 peut être considérée comme symétrique de telle sorte que :

$$q_1(\theta'|\theta) = q_1(\theta|\theta').$$

Enfin, la distribution q_2 permettant de mettre à jour la variable auxiliaire $(x \to x')$ prend la forme de la vraisemblance (3.4) mais ne dépend que de θ' :

$$q_2(x'|\theta', \theta, x) = q_2(x'|\theta') = \frac{\exp(U(x'|\theta'))}{\kappa(\theta')}.$$

Ainsi, le ratio du Metropolis-Hastings dans sa forme développée puis simplifiée s'écrit :

$$\begin{aligned} H_{(\theta,x)\to(\theta',x')} &= \frac{p(x'|\theta',y)p(y|\theta')p(\theta')}{p(x|\theta,y)p(y|\theta)p(\theta)} \times \frac{q((\theta',x')\to(\theta,x))}{q((\theta,x)\to(\theta',x'))}, \\ &= \frac{p(x'|\theta',y)\exp(U(y|\theta'))p(\theta')}{p(x|\theta,y)\exp(U(y|\theta))p(\theta)} \times \frac{\cancel{\kappa(\theta)}}{\cancel{\kappa(\theta')}} \times \frac{\cancel{q_1(\theta'|\theta)}\exp(U(x|\theta))}{\cancel{q_1(\theta|\theta')}\exp(U(x'|\theta'))} \times \frac{\cancel{\kappa(\theta')}}{\cancel{\kappa(\theta)}}, \\ &= \frac{p(x'|\theta',y)\exp(U(y|\theta'))p(\theta')}{p(x|\theta,y)\exp(U(y|\theta))p(\theta)} \times \frac{\exp(U(x|\theta))}{\exp(U(x'|\theta'))}. \end{aligned}$$

Après simplification, tous les termes du ratio sont calculables. L'algorithme 3 résume les différentes étapes d'un pas du Metropolis-Hastings suivant la méthode proposée par Møller *et al.* [67]. Il doit être répété autant de fois que nécessaire afin d'obtenir suffisamment d'échantillons. Notons que la variable auxiliaire x doit être échantillonnée grâce à un algorithme d'échantillonnage exact [47, 81] (ligne 3 de l'algorithme 3). Une

Algorithme 3 Algorithme proposé par Møller *et al.* [67]

1: **function** MH_step((θ, x), y) ▷ (θ, x) : le dernier état échantilloné, y : l'observation
2: $\theta' \sim q_1(\theta'|\theta)$ ▷ Le symbole $\sim$ fait référence à l'échantillonnage
3: $x' \sim q_2(x'|\theta')$ ▷ x nécessite d'être échantillonnée parfaitement
4: $\alpha_{(\theta,x)\to(\theta',x')} \leftarrow \min\left\{1, H_{(\theta,x)\to(\theta',x')}\right\}$
5: $accepted \leftarrow \mathcal{U}(0,1)$
6: **if** $\alpha > accepted$ **then**
7: **return** (θ', x')
8: **else**
9: **return** (θ, x)
10: **end if**
11: **end function**

question reste en suspens, le choix de la distribution $p(x|\theta, y)$. Plusieurs stratégies sont suggérées par les auteurs pour définir cette distribution. La première consiste à reprendre encore une fois la vraisemblance (3.4) mais en fixant le paramètre à $\tilde{\theta}$. Cette valeur ($\tilde{\theta}$) peut correspondre à un estimateur de θ par exemple le maximum de (pseudo) vraisemblance. La seconde proposition suggère d'utiliser une approximation de $\exp(U(x|\theta))/\kappa(\theta)$ (qui serait le cas idéal). Le choix de cette distribution est donc relativement libre et ne trouve pas de solution générique. Même si ce choix ne remet pas en cause la convergence de la chaîne à terme vers la distribution stationnaire $p(\theta, x|y)$, il peut tout de même affecter ses propriétés de mixage.

Murray, Ghahramani et MacKay [72] proposent également une variante du Metropolis-Hastings, l'*Exchange algorithm*, basée sur l'échantillonnage d'une variable auxiliaire. L'objectif est identique, rendre le ratio d'acceptation (3.6) calculable. L'idée de l'algorithme est la suivante. À chaque nouvelle proposition θ', une nouvelle variable auxiliaire x est échantillonnée suivant $p(x|\theta')$:

$$x \sim p(x|\theta'), \quad p(x|\theta') = \frac{\exp(U(x|\theta'))}{\kappa(\theta')}.$$

Cette variable auxiliaire est en quelque sorte une représentation du modèle paramétré par le paramètre θ'. La distribution échantillonnée est la suivante :

$$\begin{aligned} p(\theta', x, \theta|y) &\propto p(y|\theta)p(\theta)q(\theta'|\theta, y)p(x|\theta'), \\ &\propto \frac{\exp(U(y|\theta))}{\kappa(\theta)} p(\theta)q(\theta'|\theta, y)\frac{\exp(U(x|\theta'))}{\kappa(\theta')}. \end{aligned} \tag{3.7}$$

On retrouve dans (3.7) la vraisemblance, la distribution a priori et la distribution de la variable auxiliaire $p(x|\theta')$. La distribution de proposition $q(\theta'|\theta, y)$ peut être choisie arbitrairement et même être indépendante de y. L'algorithme 4 reprend l'ensemble des étapes d'un pas de l'*Exchange algorithm* proposé par Murray, Ghahramani et MacKay [72]. L'algorithme est organisé en deux phases. La première phase (lignes 2

Algorithme 4 *Exchange algorithm* [72]

1: **function** EA_step(θ, y) ▷ θ : le dernier paramètre échantilloné, y : l'observation
2: $\theta' \sim q(\theta'|\theta, y)$
3: $x \sim p(x|\theta')$
4: $\alpha_{\theta\leftrightarrow\theta'} \leftarrow \min\{1, H_{\theta\to\theta'}\}$
5: $accepted \leftarrow \mathcal{U}(0, 1)$
6: **if** $\alpha > accepted$ **then**
7: **return** θ'
8: **else**
9: **return** θ
10: **end if**
11: **end function**

et 3) consiste à proposer un paramètre θ' suivant la distribution $q(\theta'|\theta, y)$ et à simuler une variable auxiliaire avec ce paramètre suivant la même densité que celle du modèle. Dans un second temps (à partir de la ligne 4), l'algorithme accepte ou non la permutation $\theta \leftrightarrow \theta'$ suivant la valeur du ratio $\alpha_{\theta\leftrightarrow\theta'}$. De cette manière, le ratio se simplifie et devient facilement calculable :

$$\begin{aligned} \alpha_{\theta\leftrightarrow\theta'} = \min\{1, H_{\theta\leftrightarrow\theta'}\}, \quad H_{\theta\leftrightarrow\theta'} &= \frac{p(y|\theta')p(\theta')q(\theta|\theta', y)}{p(y|\theta)p(\theta)q(\theta'|\theta, y)} \times \frac{p(x|\theta)}{p(x|\theta')}, \\ &= \frac{\exp(U(y|\theta'))p(\theta')q(\theta|\theta', y)}{\exp(U(y|\theta))p(\theta)q(\theta'|\theta, y)} \times \frac{\exp(U(x|\theta))}{\exp(U(x|\theta'))}. \end{aligned} \tag{3.8}$$

Les différents membres du ratio d'acceptation (3.8) permettent d'expliquer comment la permutation est évaluée et acceptée ou rejetée. Tout d'abord, la variable auxiliaire est une indication de la pertinence de la proposition. Comme expliqué dans Murray [71], le ratio $\exp(U(x|\theta))/\exp(U(y|\theta))$ (3.8) évalue la validité de x par rapport à θ. Si ce ratio est supérieur à 1, cela signifie que la variable auxiliaire est privilégiée par l'ancien paramètre θ. Le ratio $\exp(U(y|\theta'))/\exp(U(x|\theta'))$ (3.8) évalue quant à lui la pertinence de la proposition θ' par rapport à l'observation y. En somme, une proposition n'est acceptée que si elle génère une variable auxiliaire satisfaisante et

qu'elle est pertinente vis-à-vis de l'observation. Les auteurs soulignent également que le ratio $\exp(U(x|\theta))/\exp(U(x|\theta'))$ peut être interprété comme l'approximation du ratio des constantes de normalisation $\kappa(\theta)/\kappa(\theta')$ par échantillonnage préférentiel (section 2.3). Pour obtenir un nombre suffisamment conséquent d'échantillons, l'algorithme 4 devra faire l'objet d'appels successifs.

L'*Exchange algorithm* est relativement simple et a donné lieu à une implantation dédiée à l'inférence des ERGMs [15, 17]. Caimo et Friel [15] utilisent en outre une stratégie, l'*adaptative direction sampling (ADS)* [36], exécutant parallèlement plusieurs chaînes qui interagissent entre elles de manière à améliorer les propriétés de mixage.

L'*Exchange algorithm* peut néanmoins rejeter des propositions pertinentes, car trop éloignées du paramètre précédemment sélectionné. Ce comportement impose une progression lente vers la distribution stationnaire que l'on pourrait qualifier "d'à petits pas". Pour augmenter le taux d'acceptation et accélérer la convergence vers la distribution stationnaire, Murray, Ghahramani et MacKay [72] proposent une variante de l'algorithme modifiant par transitions successives la variable auxiliaire pour la rendre plus conforme à θ. Les auteurs fournissent la preuve de la correction de l'algorithme. Une discussion sur cette solution ainsi qu'une revue plus détaillée de approches MCMC pour l'échantillonnage de distributions non calculables sont présentées dans les travaux de thèse de Murray [71].

Les méthodes de calcul bayésien approché ou Approximate Bayesian Computation *(ABC)* désignent un ensemble d'algorithmes permettant d'échantillonner approximativement la distribution *a posteriori* d'un modèle incalculable [1, 6, 37, 65]. Ces méthodes sont relativement simples à implanter. Elles nécessitent néanmoins des stratégies adaptées pour approcher au mieux la distribution *a posteriori*. L'algorithme 5 est l'un de ces algorithmes. Cet algorithme construit itérativement la distribution *a posteriori* approchée. À chaque itération, un nouveau paramètre est proposé (ligne 3, Algorithme 5). Un échantillon x est simulé à partir du modèle paramétré avec cette nouvelle proposition. La proposition est acceptée uniquement si les statistiques suffisantes de l'échantillon simulé sont suffisamment proches de celles de l'observation y (ligne 5, Algorithme 5). En résumé, l'algorithme 5 conserve toutes les propositions θ susceptibles de produire des échantillons proches de l'observation y suivant la valeur des statistiques suffisantes. Ce type d'algorithme peut souffrir d'un fort taux de rejets et retenir peu de propositions. D'autres stratégies [9] existent pour atténuer ces difficultés. Retenons que de manière générale, les méthodes ABC doivent contrôler la distance entre l'observation et les échantillons générés par l'algorithme.

L'algorithme ABC Shadow proposé par Stoica *et al.* [96] s'inspire directement des méthodes ABC et de la solution proposée par Møller *et al.* [67]. Comme son nom le suggère, ABC Shadow est une méthode d'échantillonnage approchée. Comparée à la solution de Møller *et al.* [67], elle offre de meilleures propriétés de mixage ainsi qu'un contrôle accru sur l'approximation par rapport aux méthodes ABC. Par ailleurs, Stoica *et al.* [95] ont récemment proposé un algorithme de recuit simulé qui s'appuie sur la dynamique d'ABC Shadow.

ABC Shadow est présenté par l'Algorithme 6. Toutes les preuves mathématiques sur la convergence de l'algorithme sont fournies dans Stoica *et al.* [96]. La méthode proposée est générale dans le sens où elle peut échantillonner des distributions *a posteriori*

Algorithme 5 Algorithme ABC

```
1: function ABC(y, m)
2:     for i = 1...m do
3:         θ ~ p(θ)
4:         x ~ p(x|θ)
5:         if d(t(y), t(x)) ≤ ε then
6:             θ^(i) ← θ
7:         end if
8:     end for
9:     return {θ^(i)}_{i=1...m}
10: end function
```

en supposant uniquement leur différentiabilité continue par rapport aux paramètres du modèle. L'algorithme est initialisé avec les paramètres suivants :

θ_0 = la valeur initiale de θ
y = l'observation, *i.e.* les statistiques suffisantes
m = le nombre de pas que doit exécuter ABC Shadow
Δ = le paramètre de contrôle de l'erreur

Le paramètre Δ définit les bornes de la distribution de proposition qui suit une loi uniforme centrée sur θ (ligne 5, Algorithme 6). Les preuves mathématiques concernant la validité de cette distribution de proposition sont également données par Stoica *et al.* [96]. Dans un premier temps, l'algorithme échantillonne une variable auxiliaire x suivant le modèle paramétré avec θ_0 (à la ligne 3, Algorithme 6). Contrairement à la méthode proposée par Møller *et al.* [67], il n'est pas nécessaire d'utiliser un algorithme d'échantillonnage parfait pour générer la variable auxiliaire. Nous pouvons dès lors utiliser des méthodes moins coûteuses comme un échantillonneur de Gibss ou l'algorithme Metropolis-Hastings (sous-section 2.3.1). À chaque itération de l'algorithme, un nouveau paramètre θ' est proposé suivant la distribution que nous avons précédemment décrite. Cette proposition est acceptée ou rejetée suivant la valeur du ratio $\alpha_{\theta\to\theta'}$:

$$\begin{aligned}
\alpha_{\theta\to\theta'} &= \min\left[1, \frac{p(\theta'|y)}{p(\theta|y)} \times \frac{p(x|\theta)q(\theta'|\theta)}{p(x|\theta')q(\theta|\theta')}\right], \\
&= \min\left[1, \frac{\exp(U(y|\theta'))p(\theta')}{\exp(U(y|\theta))p(\theta)} \times \frac{\cancel{\kappa(\theta)}}{\cancel{\kappa(\theta')}} \times \frac{\exp(U(x|\theta))}{\exp(U(x|\theta'))} \times \frac{\cancel{\kappa(\theta')}}{\cancel{\kappa(\theta)}} \times \frac{\cancel{\mathcal{U}_\Delta(\theta \to \theta')}}{\cancel{\mathcal{U}_\Delta(\theta' \to \theta)}}\right], \\
&= \min\left[1, \frac{\exp(U(y|\theta'))p(\theta')}{\exp(U(y|\theta))p(\theta)} \times \frac{\exp(U(x|\theta))}{\exp(U(x|\theta'))}\right].
\end{aligned}$$

Si nons considérons la fonction d'énergie $U(\cdot|\cdot)$ comme étant le produit scalaire des paramètres et des statistiques suffisantes $t(\cdot)$, nous retrouvons le ratio simplifié présent à la ligne 6 de l'Algorithme 6. En cas de rejet de la proposition, l'algorithme conserve son état précédent. La distribution échantillonnée par l'algorithme suit approximativement la distribution *a posteriori* $p(\theta|y)$ avec une erreur limite contrôlée par les paramètres m et Δ. La valeur de Δ doit être ajustée de manière précise pour atteindre un juste

Algorithme 6 Algorithme ABC Shadow

function ABC_SHADOW(θ_0, y, m, Δ)
  $\theta \leftarrow \theta_0$
  $x \sim p(x|\theta_0)$
  for $\_ = 1 \dots m$ **do**
    $\theta' \sim \mathcal{U}_\Delta(\theta \to \theta')$
    $\alpha_{\theta \to \theta'} \leftarrow \min\{1, \exp[(t(y) - t(x))(\theta' - \theta)](p(\theta')/p(\theta))\}$
    $accepted \leftarrow \mathcal{U}(0, 1)$
    if $\alpha_{\theta \to \theta'} > accepted$ **then**
      $\theta \leftarrow \theta'$
    end if
  end for
  return θ
end function

compromis entre la qualité de l'approximation et de bonnes propriétés de mixage. Il est également nécessaire de trouver ce juste milieu pour le paramètre m. Si le nombre m de pas exécutés par l'algorithme est trop important, l'algorithme s'éloignera de la distribution. Inversement, une valeur de m est trop petite peut avoir un effet négatif sur les propriétés de mixage. Les paramètres doivent être fixés raisonnablement lors d'une phase préalable d'ajustement. Stoica *et al.* [96] prouvent que pour un paramètre m fixé, il existe une valeur positive de Δ de telle sorte que la distribution résultante d'ABC Shadow est suffisamment proche de la distribution *a posteriori*.

Algorithme 7 Routine principale

function MAIN(θ_{prior}, y, m, Δ, $iters$)
  $\theta^{(0)} \leftarrow \theta_{prior}$
  for $i = 1 \dots iters$ **do**
    $\theta^{(i)} \leftarrow ABC\_SHADOW(\theta_{i-1}, y, m, \Delta)$
  end for
  return $\{\theta^{(i)}\}_{i=0 \dots iter}$
end function

Pour obtenir un nombre suffisamment important d'échantillons, ABC Shadow devra faire l'objet d'appels successifs comme le montre l'Algorithme 7. Le nombre d'appels est défini par le paramètre $iters$. Le paramètre θ_{prior} correspond à la valeur initiale de θ. Le paramètre θ_0 d'ABC Shadow (Algorithme 6) est mis à jour à chaque appel avec le dernier paramètre échantillonné.

ABC Shadow est comparable à l'algorithme proposé par Murray, Ghahramani et MacKay [72]. Le mécanisme d'échange de l'*Exchange Algorithm* permet d'avoir un Metropolis-Hastings qui converge à l'infini, sous condition d'avoir une simulation exacte [47, 81] de la variable auxiliaire. En outre, ABC Shadow est une méthode approximative qui ne nécessite pas de simulation exacte. Caimo et Friel [15] reprennent l'*Exchange Algorithm* et s'affranchissent de la simulation exacte. L'implantation proposée

est identique à celle d'ABC Shadow. Notons qu'ABC Shadow permet explicitement de contrôler l'approximation en instrumentant la loi de proposition avec le paramètre Δ. C'est un point qui n'est pas développé par Caimo et Friel [15]. Le rapprochement entre ces deux algorithmes constitue une perspective intéressante du point de vue mathématique.

L'inférence bayésienne consiste à échantillonner la distribution *a posteriori*. C'est en d'autres termes la distribution des paramètres d'un modèle en mesure d'expliquer l'observation donnée. Cette distribution fournit plus d'informations que le maximum de vraisemblance et permet d'étayer l'exploitation des résultats par des tests statistiques et des mesures d'incertitudes comme nous le verrons dans le Chapitre 4. Les méthodes traditionnelles comme l'algorithme Metropolis-Hastings ou l'échantillonneur de Gibbs ne permettent pas d'échantillonner directement la distribution *a posteriori*. Nous restons confrontés au problème de l'évaluation de la constante de normalisation qui est analytiquement incalculable. Différentes stratégies existent dans l'état de l'art pour contourner le problème [6, 72]. Nous proposons d'utiliser ABC Shadow, un algorithme qui s'inspire à la fois de la solution proposée par Møller *et al.* [67] et des méthodes ABC. Cet algorithme permet de mieux contrôler l'approximation et offre de meilleures propriétés numériques. Certains de ces paramètres doivent être ajustés au mieux pour assurer la qualité de l'approximation et de bonnes propriétés de mixage. Dans la section suivante (section 3.5), nous validons expérimentalement le choix de ces paramètres initiaux. Nous vérifions que l'algorithme offre une bonne approximation de la distribution *a posteriori* et qu'il est en mesure d'estimer les paramètres du modèle présenté dans la section 3.3. Au-delà des applications que nous traitons dans ce travail de thèse, il est important de souligner que l'inférence bayésienne ouvre la voie vers d'autres perspectives. Pour n'en citer qu'une : la sélection du modèle le plus pertinent pour expliquer une observation [16, 55].

3.5 Estimation de modèles simulés

L'algorithme ABC Shadow nécessite avant toute chose d'ajuster au mieux les paramètres Δ et m qui conditionnent à la fois la qualité de l'estimation et les propriétés de mixage de la chaîne estimée. Nous nous inspirons de la stratégie proposée par Stoica *et al.* [96] pour valider le choix de ces paramètres.

Cette approche consiste tout d'abord à utiliser une distribution analytiquement calculable qui peut à la fois être échantillonnée par ABC Shadow, mais aussi par un algorithme plus traditionnel comme le Metropolis-Hastings. La distribution échantillonnée par ABC Shadow est une approximation de la distribution *a posteriori* qui est directement échantillonnée par le Metropolis-Hastings. Dans la sous-section 3.5.1, nous comparons ces deux distributions échantillonnées pour vérifier que l'approximation est suffisamment proche de la distribution *a posteriori*.

Une fois cette première étape validée, nous voulons nous assurer que l'algorithme ABC Shadow est à même d'estimer le modèle que nous avons présenté dans la section 3.3. À partir d'un modèle dont les paramètres ont été fixés, nous pouvons simuler une observation comme nous l'avons fait dans la section 3.3. Dans la sous-section 3.5.2, nous exécutons

TABLE 3.2 – Statistiques des distribution *a posteriori* échantillonnée à partir de la loi binomiale (3.9)

	Q_5	Q_{25}	Q_{50}	$\overline{\theta}$	Q_{75}	Q_{95}	MAP	$\hat{\sigma}_\theta$	$\hat{\sigma}_\theta^{MC}$
ABC (θ)	-0.153	-0.69	-0.383	-0.392	-0.075	0.345	-0.408	0.454	0.021
MH (θ)	-0.15	-0.672	-0.371	-0.377	-0.071	0.353	-0.3718	0.453	0.02

ABC Shadow sur des observations simulées et vérifions que les distributions résultantes sont centrées sur les paramètres préalablement fixés.

3.5.1 Loi binomiale

On considère une observation (numérique) y générée par une loi binomiale de paramètres N et p. Replacé dans le contexte des graphes, cela pourrait correspondre à un processus de coloriage indépendant qui suit une Distribution de Bernoulli de paramètre p dans un graphe bicolore de taille N. Le paramètre N est connu et nous voulons estimer le paramètre p. La vraisemblance d'une loi binomiale s'écrit (sous sa forme exponentielle) :

$$p(y|\theta) = \binom{N}{y} p^y (1-p)^{(N-y)} = \exp\left[y\theta - N\log(1+e^\theta) + \log\binom{N}{y}\right] \tag{3.9}$$

où $\theta = \log(p/(1-p))$ et $p \in]0,1[$. Notons que tous les termes de la fonction de vraisemblance sont calculables. Dans le cadre de l'expérience nous avons fixé $N = 20$ et $p = 0.4$ ($\theta = -0.405$). Suivant ces deux paramètres, l'observation obtenue par la loi binomiale est $y = 8$. C'est à partir de cette observation qu'ont été exécutés les algorithmes Metropolis-Hastings et ABC Shadow.

L'algorithme Metropolis-Hastings échantillonne la distribution (3.9). L'algorithme exécute $m = 100$ pas pour chaque paramètre échantillonné. La distribution de proposition $p(\theta)$ suit une loi uniforme sur l'intervalle $[-100, 100]$, centrée sur la valeur actuelle de θ et de taille $\Delta = 0.005$. Pour simuler une connaissance a priori médiocre, le paramètre θ_0 a été délibérément fixé loin de la valeur réelle de θ : $\theta_0 = 1$. La procédure a été exécutée de manière à estimer 1.002×10^6 échantillons. Les 2000 premiers échantillons constituent une phase transitoire et sont donc évincés de la distribution. À cela s'ajoute une opération de sous-échantillonnage (*sub-sampling*) qui conserve tous les 100 échantillons et améliore de ce fait les propriétés de mixage. La distribution résultante est une chaîne de paramètres $(\theta^{(t)})_{t=1,\dots,T=10^4}$ de 10^4 échantillons.

La configuration d'ABC Shadow partage de nombreux points communs avec celle du Metropolis-Hastings. Le premier étant la distribution de proposition qui est identique à celle du Metropolis-Hastings. La variable auxiliaire est quant à elle directement simulée à partir de la distribution (3.9). La procédure décrite par l'algorithme 7 a été exécutée sur l'observation simulée avec $m = 100$ et $iters = 1.002 \times 10^6$. De manière analogue, ABC Shadow a échantillonné une chaîne $(\theta^{(t)})_{t=1,\dots,T=10^4}$ dont les 2×10^3 premiers échantillons ont été supprimés (phase transitoire) et sur laquelle un sous-échantillonnage qui conserve tous les 100 échantillons a été appliqué.

La Figure 3.7 représente les traces des distributions respectivement échantillonnées par ABC Shadow (à gauche) et par Metropolis-Hastings (à droite). En complément de cette représentation et pour faciliter la comparaison entre les deux distributions, la Figure 3.8 décrit les distributions sous la forme de boites à moustache (à gauche) et d'un diagramme quantile-quantile (à droite). Nous remarquons que les différents quantiles des deux distributions sont mutuellement alignés (Figure 3.8 gauche) suggérant ainsi qu'elles sont très proches l'une de l'autre. Ce constat se confirme par l'alignement des échantillons sur la droite rouge dans le diagramme quantile-quantile (Figure 3.8 droite).

La Table 3.2 représente les statistiques calculées à partir des deux distributions échantillonnées. Les similitudes entre les valeurs numériques des quantiles sur la ligne supérieure (ABC Shadow) et inférieure (Metropolis-Hastings) corroborent nos observations précédentes. Notons par ailleurs que les deux distributions convergent toutes les deux vers la valeur exacte du paramètre θ ($\theta = -0.405$) et ce, même avec un mauvais a priori ($\theta_0 = 1$). Le MAP correspond à la valeur du paramètre la plus représentée dans la distribution. Elle peut-être comprise comme étant la configuration du modèle la plus pertinente pour expliquer l'observation y. Le MAP est calculé en sélectionnant le maximum du noyau de densité des distributions. Pour les deux distributions échantillonnées, il est proche du paramètre de référence, voire presque égal, pour celui issu d'ABC Shadow. La Table 3.2 fournit également l'écart-type asymptotique ($\hat{\sigma}_\theta$) et l'écart-type de Monte-Carlo ($\hat{\sigma}_\theta^{MC}$) que nous avons déjà eu l'occasion de présenter dans la section 2.3. Nous considérons le MAP comme un estimateur du maximum de vraisemblance de Monte-Carlo. Dans ces conditions, l'écart-type asymptotique estime la différence entre le vrai paramètre du modèle et l'estimateur du maximum de vraisemblance qui sont tous les deux inconnus. L'écart-type de Monte-Carlo estime quant à lui la différence entre l'estimateur du maximum de vraisemblance (qui est inconnu) et l'estimateur du maximum de vraisemblance de Monte-Carlo. Ces deux calculs se basent sur 10^4 observations simulées à partir des MAPs. Il est possible de calculer l'intervalle de confiance à partir de l'écart-type de Monte-Carlo.

3.5.2 Échantillonnage *a posteriori* sur le modèle de Potts

Nous considérons à présent le modèle d'interactions que nous avons proposé dans la section 3.3. La constante de normalisation incalculable placée au dénominateur ne nous permet pas d'échantillonner directement sa loi *a posteriori* avec l'algorithme Metropolis-Hastings comme nous avons pu le faire dans la précédente section (sous-section 3.5.1). La stratégie d'évaluation proposée par Stoica *et al.* [96] suggère de comparer le maximum de la distribution *a posteriori* échantillonnée par ABC Shadow avec les vrais paramètres du modèle simulé pour générer l'observation.

Dans un premier temps, nous fixons tous les paramètres d'interactions à 0 : $\theta_{11} = \theta_{12} = \theta_{22} = 0$. De la sorte, l'effet des interactions est inhibé et les occurrences des trois motifs ($1 \leftrightarrow 1$, $1 \leftrightarrow 2$, $2 \leftrightarrow 2$) sont équiprobables. Dans cette configuration le modèle est comparable à un modèle de Bernoulli pour les graphes de probabilité 1/3. Contrairement à la situation présentée dans la section 3.3, l'observation simulée représente une organisation composée de 12 personnes et 8 collaborateurs extérieurs $size = (12, 8)$. Nous nous concentrons sur une seule organisation et les liens qu'elle peut entretenir avec

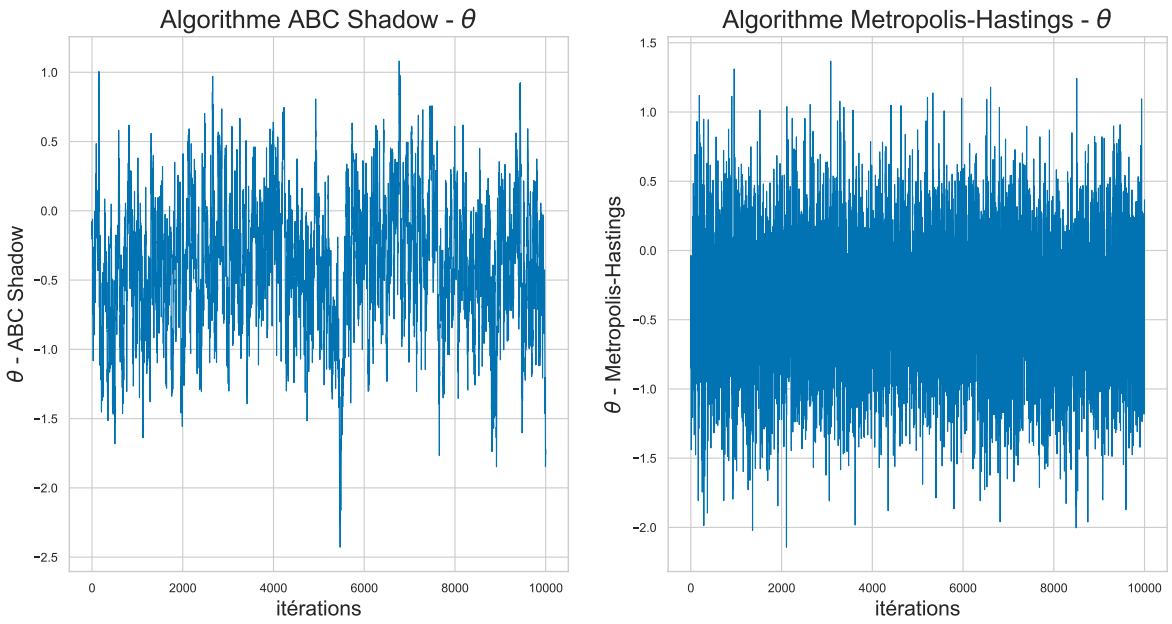


FIGURE 3.7 – Traces laissées par les distributions *a posteriori* échantillonnées par ABC Shadow (à gauche) et Metropolis-Hastings à droite.

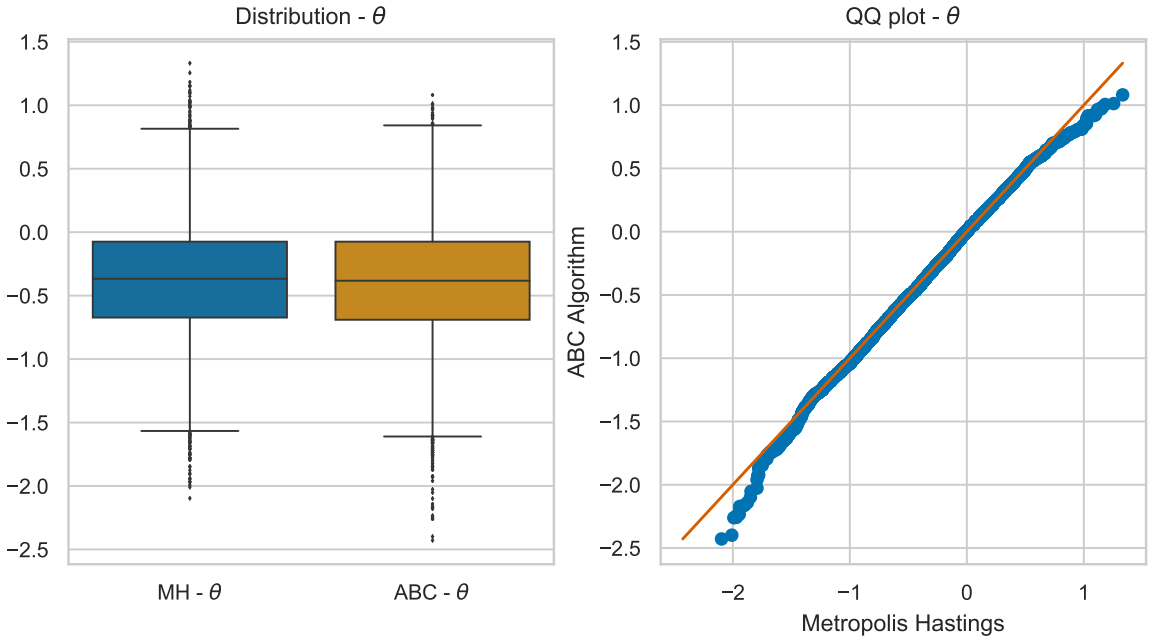


FIGURE 3.8 – Comparaison des distributions respectivement échantillonnées (à partir de (3.9)) par les algorithmes Metropolis-Hastings et ABC Shadow au travers d'un diagramme en boîte (à gauche) et d'un diagramme quantile-quantile (à droite).

des partenaires extérieurs. Nous ne considérons que les liens impliquant au moins un membre de l'organisation. Les liens entre deux collaborateurs extérieurs sont ignorés.

TABLE 3.3 – Statistiques sur la distribution échantillonée par ABC avec la configuration sans intéraction. Les valeurs des MAPs sont très proches du vrai vecteur $\theta = [0, 0, 0]$.

	Q_{50}	$\overline{\theta}$	MAP	$\hat{\sigma}_\theta$	$\hat{\sigma}_\theta^{MC}$
ABC (θ_{11})	-0.05	-0.06	-0.03	0.08	7.11×10^{-4}
ABC (θ_{12})	0.05	0.06	0.04	0.10	9.04×10^{-4}
ABC (θ_{22})	-0.09	-0.11	-0.06	0.16	1.67×10^{-3}

De même, les interactions comptabilisées par les statistiques suffisantes ne sont calculées qu'au niveau des membres de l'organisation. Nous réutiliserons ce conditionnement sur l'observation dans le Chapitre 4 et nous l'appliquerons à des données réelles. L'observation a été générée à partir de 10^3 échantillons calculés par un échantillonneur de Gibbs. En faisant la moyenne des statistiques suffisantes de tous les échantillons nous obtenons les statistiques suffisantes de l'observation simulée qui seront fournies en entrée d'ABC Shadow : $\overline{t}(y) = [164.75, 263.49, 83.76]$. La distribution a priori $p(\theta)$ intégrée dans ABC Shadow suit une loi uniforme sur le volume $[-4, 4] \times [-4, 4] \times [-4, 4]$. Les paramètres m et Δ sont respectivement fixés à $m = 200$ et $\Delta = [0.005, 0.005, 0.005]$. Tout comme dans la sous-section 3.5.1, ABC Shadow a été exécuté pour produire 1.002×10^6 échantillons. Les 2000 premiers échantillons de la phase transitoire ont été évincés. Pour améliorer les propriétés de mixage, un sous-échantillonnage conservant tous les 100 échantillons a été appliqué sur la chaîne restante. La chaîne résultante est une chaîne de 10^4 vecteurs : $(\theta^{(t)})_{t=1,\dots,T=10^4}$. À chaque itération, une variable auxiliaire x est échantillonnée (ligne 3, Algorithme 6) en utilisant 200 itérations d'un échantillonneur de Gibbs (de manière analogue à l'observation simulée).

À partir de la distribution échantillonnée nous avons calculé la moyenne, la médiane, le MAP, l'écart-type asymptotique ainsi que l'écart-type de Monte-Carlo de chaque distribution marginale. Le calcul des deux écarts-types a nécessité la génération de 10^4 échantillons simulés à partir des MAPs. Toutes ces valeurs sont présentées dans la Table 3.3. La Figure 3.9 présente les distributions marginales obtenues. Elle se présente sous la forme de trois histogrammes représentants les trois distributions marginales échantillonnées. Chaque distribution est projetée sur des plans en 2 dimensions suivant les combinaisons de paramètres : $(\theta_{11}, \theta_{22})$, $(\theta_{11}, \theta_{12})$ et $(\theta_{12}, \theta_{22})$. Plus la zone est sombre plus la densité est élevée. Nous remarquons que les lignes bleues qui représentent les MAPs sont proches des lignes vertes qui correspondent aux vrais paramètres du modèle. En effet, les valeurs numériques obtenues pour le MAP ($\hat{\theta} = [-0.03, 0.04, -0.06]$) sont très proches des vrais paramètres $\theta = [0, 0, 0]$.

Nous considérons à présent le modèle avec des effets d'interactions (c'est-à-dire $\theta_{ij} \neq 0$). Nous avons arbitrairement fixé les paramètres de cette manière : $\theta_{11} = -0.50$, $\theta_{12} = 0.20$ et $\theta_{22} = 0.30$. Sur le même principe que l'expérience précédente, nous avons simulé une observation impliquant une organisation composée de 12 individus potentiellement en relation avec 8 membres extérieurs : $size = (12, 8)$. Après 10^3 itérations d'un échantillonneur de Gibbs, les statistiques suffisantes de l'observation simulées sont en moyenne $\overline{t}(y) = [78.88, 360.73, 295.55]$. Les paramètres Δ et n sont inchangés par rapport à l'expérience précédente. La méthode de simulation de la variable auxiliaire reste la

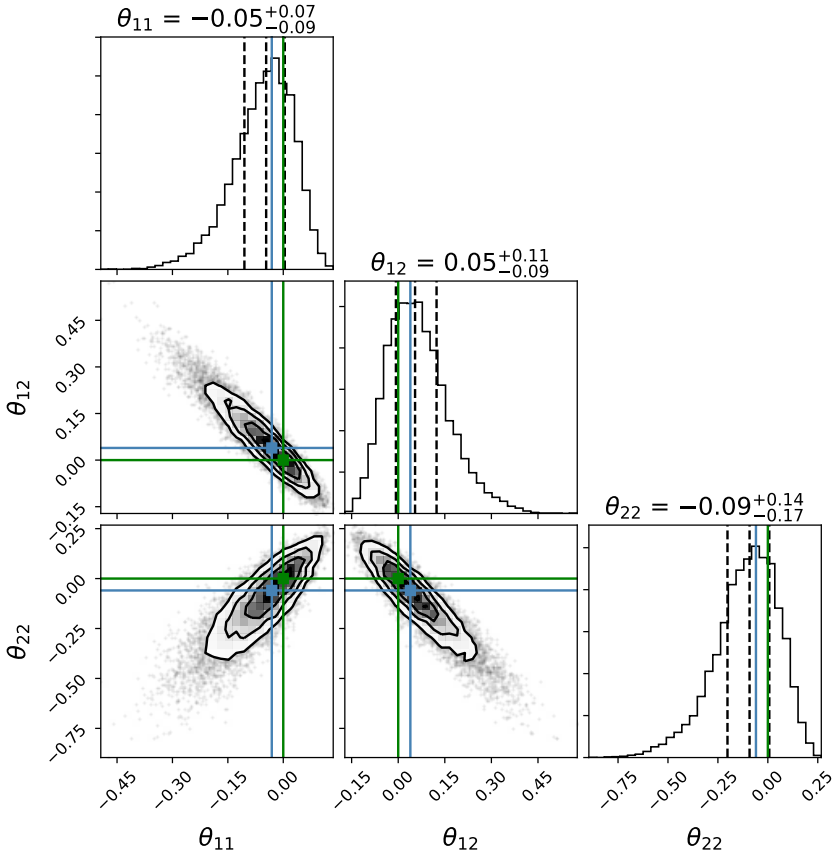


FIGURE 3.9 – Distributions marginales de l'échantillonnage *a posteriori* du modèle de Potts présenté dans la section 3.3 avec les paramètres $\theta_{11} = \theta_{12} = \theta_{22} = 0$. Les lignes bleues représentent les MAPs et les lignes vertes correspondent aux vraies valeurs des paramètres.

TABLE 3.4 – Statistiques sur la distribution échantillonée par ABC à partir d'une observation simulée avec $\theta = [-0.50, 0.20, 0.30]$. Les valeurs des MAPs sont très proches du vrai vecteur ($\theta = [-0.50, 0.20, 0.30]$).

	Q_{50}	$\overline{\theta}$	MAP	$\hat{\sigma}_\theta$	$\hat{\sigma}_\theta^{MC}$
ABC (θ_{11})	-0.66	-0.70	-0.54	0.36	8.43×10^{-3}
ABC (θ_{12})	0.27	0.29	0.21	0.17	3.77×10^{-3}
ABC (θ_{22})	0.26	0.26	0.29	0.12	2.32×10^{-3}

même. La distribution échantillonnée à la même dimension que celle qui a été décrite pour la configuration précédente : $(\theta^{(t)})_{t=1,\dots,T=10^4}$. Les statistiques calculées à partir des distributions marginales sont présentées dans la Table 3.4.

Sur le même principe que la Figure 3.9, la Figure 3.10 représente les distributions marginales échantillonnées à partir de l'observation simulée avec $\theta = [-0.50, 0.20, 0.30]$. Nous pouvons faire le même constat que pour la configuration sans intéraction, les lignes

représentant les MAPs et celles représentant les vraies valeurs de θ sont très proches les unes des autres et les intersections se chevauchent.

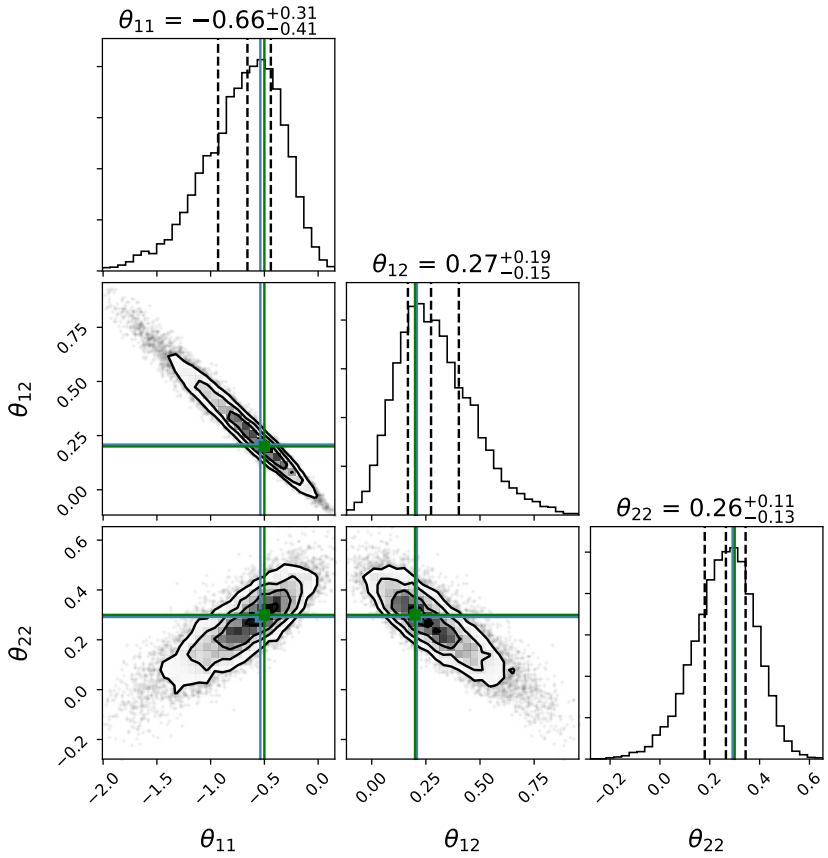


FIGURE 3.10 – Distributions marginales de l'échantillonnage *a posteriori* du modèle de Potts présenté dans la section 3.3 avec les paramètres $\theta_{11} = -0.50$, $\theta_{12} = 0.20$ et $\theta_{22} = 0.30$. Les lignes bleues représentent les MAPs et les lignes vertes correspondent aux vraies valeurs des paramètres.

3.6 Synthèse

Les interactions entre des individus appartenant à une population donnée peuvent intuitivement être représentées sous la forme d'un graphe. Les nœuds représentent les individus et les arêtes, les liens entre ces derniers. Cette représentation est tout à fait adaptée à l'étude des collaborations. Nous nous intéressons plus particulièrement aux collaborations impliquant différents groupes d'individus ou organisations. Nous cherchons à comprendre la manière dont ces individus interagissent au sein de leur organisation et avec des membres extérieurs. Dans ce contexte, nous proposons une approche permettant d'inférer une structure morphologique des interactions à partir d'un graphe observé.

Nous proposons tout d'abord dans la section 3.2, une transformation duale du graphe initial en graphe des arêtes. Dans cette représentation, les arêtes du graphe initial sont

les nœuds. Deux arêtes adjacentes du graphe initial (connectées au même nœud) sont représentées par deux nœuds reliés par une arête dans le graphe des arêtes. Ce changement de perspective met en avant la structure des interactions. Les nœuds du graphe des arêtes sont étiquetés en fonction du type d'interaction qu'ils représentent. Ainsi, la structure des interactions d'un graphe de collaboration donné se traduit par une configuration d'étiquetage particulière. Ce qui revient à définir le processus de création des interactions comme un processus spatial d'étiquetage de graphe.

À partir de ce formalisme que nous empruntons aux MRFs, nous décrivons dans la section 3.3 le modèle qui définit les phénomènes que nous voulons mesurer au travers de nos observations. Nous proposons un modèle élémentaire qui repose sur les hypothèses de dépendance de Markov. Ce modèle est une extension du modèle de Potts qui comptabilise trois types d'interaction. Ces interactions nous permettent d'identifier des situations où un individu est connecté à deux autres collaborateurs, jouant ainsi le rôle de relais entre ces deux collaborateurs. En fonction du type d'interactions considéré, les collaborateurs situés de part et d'autre du relais peuvent avoir la même affiliation ou appartenir à différentes organisations. De cette manière, le modèle nous permet d'évaluer les tendances à tisser des liens au sein et entre les organisations.

La forme du modèle induit la présence d'une constante de normalisation qui est placée au dénominateur et qui rend le modèle analytiquement incalculable. Nous proposons dans la section 3.4 d'utiliser un algorithme d'inférence bayésienne : ABC Shadow, qui approche la distribution *a posteriori* du modèle. Comparé aux autres solutions de l'état de l'art, ABC Shadow offre de bonnes performances numériques et un meilleur contrôle sur l'approximation. Certains de ses paramètres nécessitent d'être ajustés pour assurer une bonne approximation. Les auteurs de l'algorithme proposent une stratégie pour valider le choix des paramètres initiaux et la qualité de l'approximation. Nous l'avons mise en œuvre dans la section 3.5. Cette stratégie consiste à comparer la distribution échantillonnée par ABC Shadow (distribution approchée) à la distribution *a posteriori* exacte dans le cas d'un modèle calculable. L'algorithme est ensuite évalué sur des observations simulées à partir de notre modèle de Potts.

Dans le Chapitre suivant (Chapitre 4), nous appliquons notre approche à un cas d'étude réel : les collaborations entre chercheurs. Le contexte porte plus précisément sur les collaborations tissées par les chercheurs dans le cadre de leur activité de publication scientifique. L'objectif de ce travail est de montrer la capacité de l'approche à identifier des structures de collaboration propres à ce contexte. De cette manière, nous souhaitons évaluer la tendance pour un chercheur à collaborer avec des membres de son équipe et avec des collaborateurs extérieurs. À partir de données bibliographiques issues d'une archive ouverte, nous avons tracé les graphes de collaboration, appliqué notre modèle et exécuté ABC Shadow. L'exploitation des distributions échantillonnées par ABC Shadow nous a permis d'étudier les collaborations au sein de différentes équipes de recherche et d'identifier des caractéristiques communes à plusieurs équipes.

Chapitre 4

Application sur des collaborations entre chercheurs

Sommaire

4.1 Problématique

Les liens sociaux créés entre les chercheurs dans la cadre de leurs activités de recherche sont fondamentaux pour le partage de la connaissance, des compétences et des expertises nécessaires à la production de résultats scientifiques. Nous pouvons observer ces liens à travers les publications scientifiques qu'ils co-écrivent. Ces liens constituent le réseau des chercheurs qui ont rédigé au moins un article scientifique ensemble. D'autres liens peuvent être également envisagés, comme les liens de citation entre deux articles.

Certaines études se sont déjà emparées du sujet. Price [80] s'est par exemple intéressé aux pratiques des chercheurs quant à leurs manières de référencer d'autres articles dans leurs publications. Newman [74] a montré que les réseaux des coauteurs dans différentes disciplines formaient un "petit monde", dans le sens où la distance moyenne entre deux auteurs est faible par rapport à la taille totale du réseau. Ce travail a également mis en évidence certaines différences structurelles entre les disciplines témoignant de différentes pratiques et stratégies de publication.

Les données bibliographiques sont des objets pertinents pour l'étude des collaborations scientifiques. De ces données, nous pouvons extraire les interactions entre les chercheurs et construire un réseau social des chercheurs. Ce réseau social présente un enjeu puisqu'il détermine les collaborations et *in fine* la production de connaissances. En effet, le contenu des publications, les résultats et leur portée ne seront pas les mêmes en fonction des relations entretenues par un chercheur avec ses homologues.

Les chercheurs sont très souvent rattachés à un laboratoire ou une université. Ces organisations peuvent être aussi subdivisés en équipes ou départements de recherche. Cette structure organisationnelle est souvent cohérente vis-à-vis des thématiques de recherche et des problématiques traitées. Ainsi les membres d'une même équipe travaillent sur des problématiques similaires et appartiennent aux mêmes communautés scientifiques. Les équipes de recherches offrent un cadre pour la collaboration. Nous pouvons intuitivement penser que les chercheurs peuvent plus facilement communiquer et interagir au sein d'une même équipe. Enfin, nous pouvons supposer que les équipes de recherche, de par leur rayonnement, favorisent la création et la collaboration avec d'autres équipes, à l'échelle nationale, européenne ou internationale.

Il nous parait intéressant de confronter ces intuitions à la réalité et de placer les structures sociales dans un contexte organisationnel. Notre objectif est d'étudier les liens tissés par les chercheurs à l'échelle d'une équipe de recherche. Notre étude considère les interactions entre chercheurs ayant co-rédigé une publication (lien de coauteur). Notre approche diffère des travaux précédemment cités [74, 80] puisqu'elle fait intervenir les affiliations des chercheurs à une équipe ou un laboratoire. Ainsi, en considérant des chercheurs appartenant à la même équipe, nous pouvons étudier les liens qu'ils tissent avec leurs collègues (chercheurs de la même équipe) et leurs homologues extérieurs (chercheurs affiliés à d'autres équipes ou laboratoires). En nous appuyant sur le modèle d'interaction décrit par la section 3.3, nous souhaitons identifier des motifs structurels caractéristiques de ces réseaux de collaborations.

Notre étude s'articule autour de trois questions :

Dans quelle mesure les membres d'une équipe ont-ils tendance à collaborer avec leurs homologues internes ? Les chercheurs peuvent occuper une place différente au sein de leur équipe en fonction notamment du statut, de leur position hiérarchique et du domaine de recherche. Cette position s'illustre au travers des liens qu'ils tissent avec leurs collègues. Nous souhaitons identifier des individus clés dans la collaboration, des individus qui collaborent avec de nombreux autres chercheurs au sein de l'équipe. Plus généralement, nous voulons caractériser l'occurrence d'un lien interne dans la collaboration et déterminer si cette occurrence reflète une réelle tendance ou est le fait d'un processus purement aléatoire.

Quelles structures caractérisent les interfaces inter-organisationnelles ? De nombreuses productions scientifiques voient le jour dans le cadre de projets nationaux, européens ou internationaux (ANR, H2020, ...). Des communautés scientifiques se créent autour d'une thématique ou d'une problématique donnée. En somme, les liens de collaboration ont très souvent tendance à dépasser les frontières d'un laboratoire. À l'échelle d'une équipe comment ces liens inter-organisationnels sont-ils distribués ? Est-ce que seule une petite minorité des membres entretiennent des liens avec l'extérieur ? De manière plus absolue, est-ce que les équipes considérées ont tendance à tisser des relations inter-organisationnelles ?

Existe-t-il des motifs récurrents communs entre les équipes ? Les équipes de recherche peuvent adopter différentes postures dans leur façon de collaborer en interne et d'interagir avec d'autres équipes. Certaines équipes peuvent par exemple privilégier les collaborations internes alors que d'autres vont exclusivement collaborer avec d'autres équipes externes au laboratoire. Ces différentes postures sont reflétées par la distribution des liens entre les chercheurs et sont donc structurellement identifiables. Nous souhaitons vérifier s'il y a des motifs récurrents parmi les équipes considérées. En d'autres termes : les équipes interagissent-elles de la même manière ? Nous souhaitons par ailleurs vérifier si ces récurrences peuvent être liées à la thématique de recherche ou totalement décorrélées.

L'objectif de ce chapitre est de répondre à ces questions en mettant en œuvre l'approche décrite par le Chapitre 3 sur un jeu de données réel. Le contexte, les sources et l'ensemble des processus qui ont permis la construction du jeu de données sont décrits dans la Section suivante (section 4.2). Une fois le jeu de données constitué, nous pouvons commencer l'étude des collaborations entre chercheurs. Cette étude s'organise en deux parties. Une première partie porte sur les collaborations initiées par les chercheurs d'une même équipe avec leurs homologues internes et externes (section 4.3). Nous cherchons dans cette partie à répondre aux deux premières questions précédemment mentionnées. À partir des résultats obtenus pour chaque équipe, nous souhaitons identifier des motifs structurels communs à plusieurs équipe. Dans la section 4.4, nous présentons une classification des résultats de chaque équipe. Cette classification est mise en perspective avec les thématiques de recherche des équipes. Une synthèse de ces résultats est présentée dans la section 4.5.

4.2 Construction du jeu de données

Les chercheurs sont régulièrement amenés à publier dans des revues scientifiques ou dans des conférences (sous la forme de *proceedings*). Ces publications sont l'aboutissement d'un travail de recherche très souvent collectif. Ces collaborations peuvent dépasser les frontières des équipes de recherche et des laboratoires, notamment lorsque les problématiques traitées sont transdisciplinaires. Les liens tissés à l'occasion de ces collaborations forment un maillage social qui sous-tend le travail de recherche. L'étude de ces collaborations au travers des publications scientifiques offre ainsi la possibilité de reconstruire ce maillage et donc d'étudier le graphe de collaboration.

Différentes plateformes comme *Google Scholar* ou *dblp (Digital Bibliography & Library Project)* recensent et indexent les publications scientifiques. Elles facilitent le travail de recherche bibliographique et permettent de trouver facilement une publication en fonction de son titre, de ses auteurs ou de la problématique traitée. Ces moteurs de recherches ne sont pas les seuls dispositifs à faciliter l'accès aux publications. Les archives ouvertes comme *arXiv* ou *HAL* mettent à disposition librement des publications scientifiques. En parallèle des soumissions dans une revue ou une conférence, les chercheurs déposent souvent leurs articles sur ces archives ouvertes. C'est une pratique de plus en plus courante qui offre la possibilité à chacun d'accéder à la publication sans devoir s'acquitter d'un abonnement [8] auprès des éditeurs scientifiques. Cela permet de surcroît de rendre la publication plus visible, puisqu'elle est accessible en tant que document de travail avant que le processus de révision (parfois long) des revues soit terminé. Nous trouvons également sur ces archives d'autres types de publication comme des posters, des manuscrits, des brevets ...

Les archives ouvertes sont donc des sources de données intéressantes du fait de leur popularité et de la diversité des documents hébergés. L'archive ouverte HAL facilite tout particulièrement la récupération des données en mettant à disposition une base de connaissances des métadonnées des documents hébergés [9]. Nous avons tiré profit de ce dispositif pour collecter notre jeu de données. Notre étude porte sur les collaborations initiées par les chercheurs d'un laboratoire de recherche en informatique, le Loria (Laboratoire lorrain de recherche en informatique et ses applications). Les domaines de recherche du Loria sont vastes et vont de l'informatique fondamentale à l'informatique appliquée. Ce laboratoire comporte 28 équipes organisées en cinq départements qui définissent les principaux axes de recherche :

— Département 1 : Algorithmique, calcul, image et géométrie
— Département 2 : Méthodes formelles
— Département 3 : Réseaux, systèmes et services
— Département 4 : Traitement automatique des langues et des connaissances
— Département 5 : Systèmes complexes, intelligence artificielle et robotique

Parmi tous les types de documents déposés sur HAL, nous avons choisi de nous intéresser aux articles de revues et de conférences. En proportion, ce sont les documents les plus fréquement déposés sur HAL. Mais outre la quantité, c'est la nature de ces documents

8. exorbitant
9. http://data.archives-ouvertes.fr

qui nous intéresse. Les articles de revues sont très souvent l'aboutissement de travaux de recherche de longue haleine qui nécessitent l'expertise de plusieurs chercheurs et qui pousse au travail collectif. Les articles de conférences sont quant à eux plus répandus et sont également l'illustration d'un travail collaboratif.

La Figure 4.1 représente un extrait des métadonnées associées à un article de conférence qui sont stockées dans la base de connaissances de HAL. Nous nous concentrons particulièrement sur les auteurs et leurs affiliations respectives. Les auteurs sont les collaborateurs et nous considérons ici que les auteurs d'un même article ont tous œuvré ensemble (collaboré) pour rédiger cette article et mener à bien le travail scientifique sous-jacent. À partir de ces deux postulats, nous pouvons tracer un graphe social des collaborations entre chercheurs. Les nœuds de ce graphe sont les auteurs (les chercheurs). Deux chercheurs qui ont tous deux co-écrit un article sont connectés dans le graphe. Suivant ces deux règles, les auteurs de la publication (Figure 4.1) sont tous interconnectés.

HAL est une des seules plates-formes à notre connaissance qui fournit des informations organisationnelles. Comme illustré par la Figure 4.1, HAL renseigne les affiliations de chaque auteur, et ce à plusieurs échelles. Les chercheurs sont généralement affiliés à des laboratoires de recherche ou à des universités c'est le cas du deuxième et du quatrième auteur. Certains laboratoires ont une organisation interne plus complexe. Ils peuvent être divisés en équipes de recherche, ces équipes peuvent même être organisées en départements, lorsque le laboratoire est de taille importante. C'est le cas du Loria qui a fourni ces informations organisationnelles à HAL. La première affiliation correspond à une des équipes du Loria, l'équipe COAST qui appartient au *"Department of Networks, Systems and Services"* (département 3), comme indiqué par la capture d'écran (Figure 4.1). La base de connaissances de HAL définit sémantiquement et met en relation ces différentes structures et sous-structures qui composent un organisme de recherche et facilite de ce fait la collecte des données organisationnelles.

Les données collectées concernent tous les articles de revues ou de conférences dont au moins un des auteurs est membres du Loria. Nous avons choisi comme échelle de temps l'année et collecté tous les articles déposés sur la plate-forme en 2018. À partir de la liste des auteurs de chaque article, nous pouvons tracer le graphe social des collaborations entre chercheurs comme nous l'avons fait pour l'exemple de la Figure 4.1.

La Figure 4.2 représente le graphe résultant. Les nœuds du graphe sont des chercheurs et deux chercheurs qui ont co-rédigé un article (collaboré) sont reliés par une arête. Nous avons intégré dans ce graphe des informations organisationnelles de première ordre, à savoir, les affiliations aux : laboratoires, universités et instituts de recherche en général. Les nœuds sont colorés en fonction de l'affiliation du chercheur associé. Certains chercheurs ont de multiples affiliations, c'est notamment le cas du premier auteur de l'article présenté dans notre exemple (Figure 4.1). Si une de ces multiples affiliations fait référence au Loria, nous considérons que cet auteur est membre du Loria. Ce qui est le cas de l'auteure mentionnée. Dans l'autre cas, nous considérons l'affiliation multiple comme une affiliation à part entière. D'autres chercheurs n'ont pas renseigné d'affiliation, nous les avons par conséquent associés à une affiliation fictive `UNKOWN`. Toutes les affiliations (`UNKOWN` comprise) sont associées à une unique couleur permettant ainsi de les distinguer. Par exemple, tous les chercheurs affiliés au Loria sont colorés en jaune dans le graphe.

Nous avons désigné sur le graphe (Figure 4.2) deux zones. La première, en bas à droite,

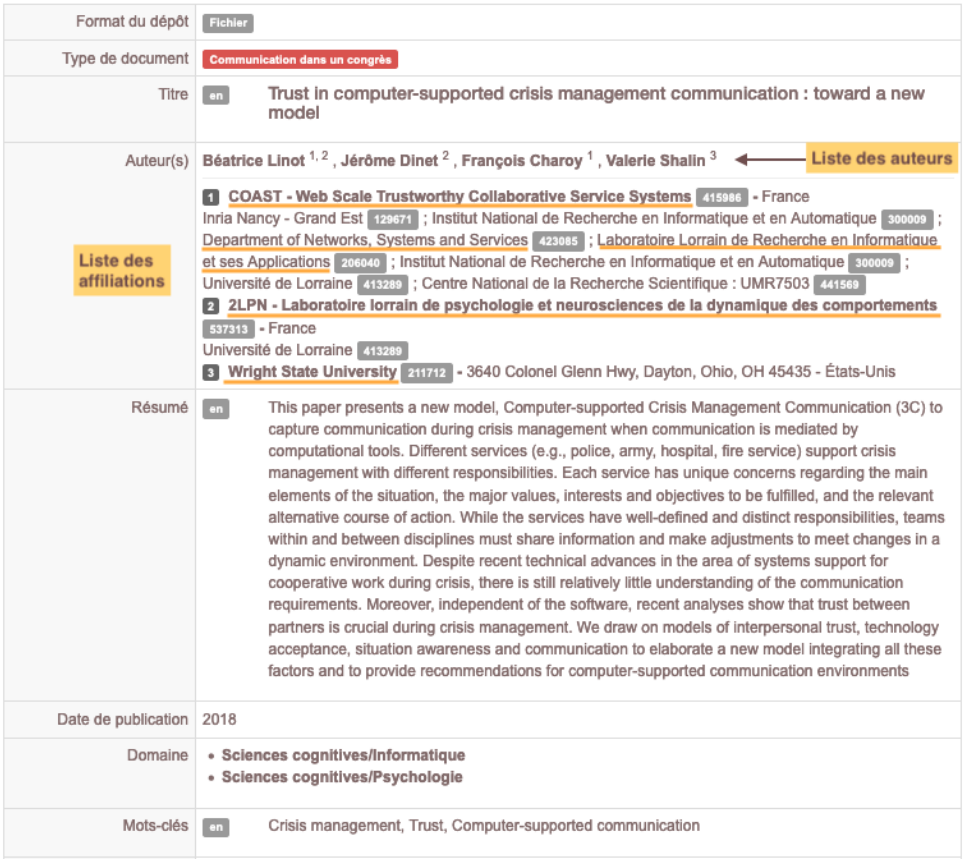


FIGURE 4.1 – Exemple des métadonnées référencées par HAL pour un article de conférence (extrait)

représente l'exemple que nous avons précédemment évoqué pour illustrer les méta-données disponibles sur HAL (Figure 4.1). L'article de conférence présenté dans cet exemple est associé à l'identifiant HAL `hal-02280866`. Nous retrouvons dans cette zone les quatre auteurs de l'article. Les deux nœuds colorés en jaunes désignent les deux auteurs affiliés au Loria. Les deux autres auteurs sont associés à des couleurs distinctes, désignant le laboratoire 2LPN et l'université Wright State. La seconde zone située au centre, représente un sous-graphe complet de 41 nœuds. Ce sous-graphe représente le réseau de collaboration impliquant les 41 auteurs d'un article de revue identifié par : `hal-01863922`. Parmi ces auteurs, il y a un chercheur affilié au Loria.

Lorsque nous collectons des données, il y a souvent une différence entre la représentation que nous nous en faisons et leur état réel. La réalité "du terrain" (en l'occurrence bibliographique) nous rattrape et provoque des inconsistances et des erreurs d'observation. Par exemple, nous avons considéré initialement que les auteurs recensés étaient uniques. D'autant que HAL propose un mécanisme permettant d'agréger plusieurs identités en une seule. De plus, de nombreuses vérifications humaines sont effectuées pour maintenir autant que possible la cohérence de la base de données. Malgré tout, certains auteurs ont plusieurs identités. Cela peut être notamment dû à une erreur au moment du dépôt de l'article. Par ailleurs, certains chercheurs changent de nom

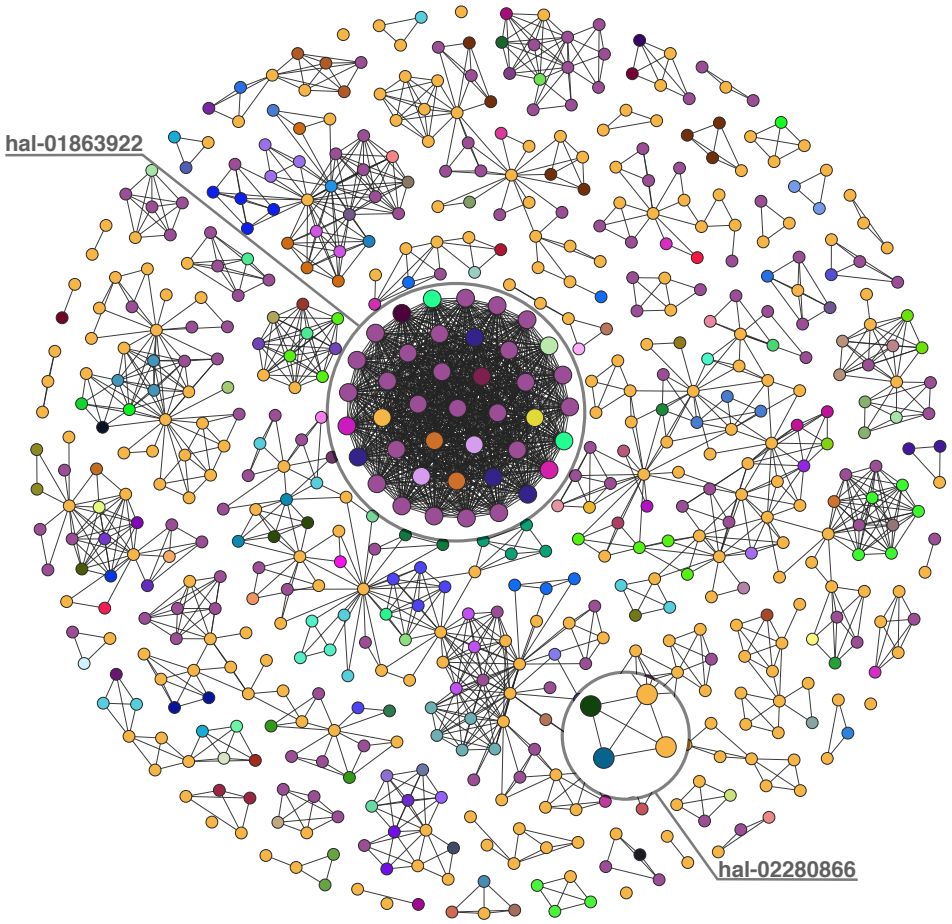


FIGURE 4.2 – Graphe de collaborations impliquant des chercheurs du Loria

d'usage du fait de l'évolution de leur situation personnelle. Toutes ces identités ne sont pas forcément agrégées et produisent des identités multiples dans les données collectées. Un chercheur qui a plusieurs identités non agrégées sera représenté dans le graphe par plusieurs nœuds. Ce type d'erreur impacte négativement la cohérence du graphe social généré. Une comparaison sur le nom d'usage et l'affiliation permet de détecter la majeure partie des identités dupliquées. La comparaison stricte des noms n'est pourtant pas suffisante. Elle ne permet pas de détecter des identités multiples lorsque le nom d'usage a changé. Il arrive par moment que le nom de naissance soit ajouté (inversement retiré) sur plusieurs identités référençant le même chercheur. Pour de tels cas, nous appliquons une méthode de comparaison approximative (*fuzzy matching*). Si le score de cette méthode est satisfaisant (les deux noms de famille ont une très forte ressemblance) et que les affiliations correspondent, alors nous considérons que les identités correspondent à un seul et même chercheur. Dans le cas où il y a ambiguïté et que l'algorithme n'a pas suffisamment de données pour trancher, une vérification "humaine" est nécessaire. Fort heureusement ces cas sont très limités (moins d'une vingtaine). Cette procédure de nettoyage a été mise en œuvre en amont de la génération du graphe.

Le graphe obtenu représente les collaborations initiées par les chercheurs au travers

des publications déposées sur HAL et publiées en 2018. Ce graphe nous permet d'avoir une vision à plusieurs niveaux. Nous avons associé à chaque nœud des affiliations aux laboratoires et aux équipes de recherche. Ce qui nous permet d'étudier les collaborations de chaque équipe du Loria de manière indépendante comme présenté dans la Section suivante (section 4.3). Ces différents nivaux encodés dans le graphe ouvrent la voie à d'autres applications que nous ne traitons pas ici. Nous pourrions par exemple envisager d'étudier les collaborations non plus à l'échelle individuelle, mais plutôt à l'échelle des équipes. Nous pourrions également étendre la démarche à d'autres laboratoires et construire le graphe des collaborations "inter laboratoire".

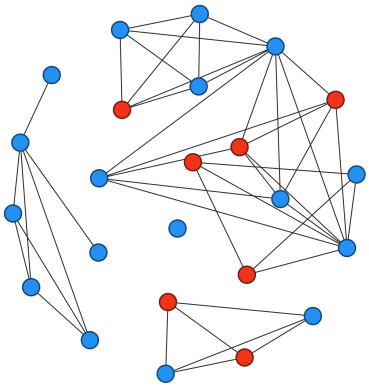
4.3 Étude des collaborations au sein des équipes de recherche

À partir du graphe de collaboration du Loria (Figure 4.2), nous pouvons dissocier les graphes des différentes équipes de recherche. Nous avons construit ces graphes en considérant tous les membres de chaque équipe ainsi que leurs voisins directs. En d'autres termes, chaque graphe d'équipe est construit à partir des réseaux personnels (*ego-networks*) de ses membres.

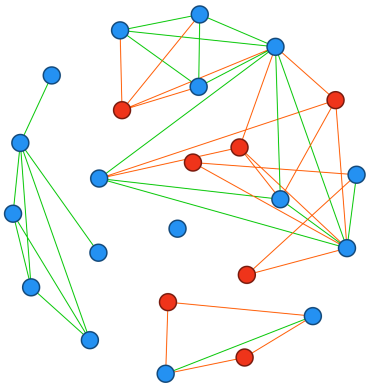
Nous avons sélectionné 17 équipes pour lesquelles les données recueillies produisent un graphe de collaboration suffisamment important [10] [59]. Chaque graphe d'équipe est un objet pour lequel nous appliquons l'approche décrite par le Chapitre 3. Un pré-traitement sur chaque graphe de collaboration est nécessaire afin de retrouver le formalisme décrit par la section 3.2. Ce pré-traitement illustré par la Figure 4.3 est scindé en trois étapes. À partir des affiliations de chaque nœud du graphe de collaboration, nous pouvons différencier deux types de collaborateurs : les membres du Loria • et les chercheurs affiliés à d'autres centres de recherche • (Figure 4.3(a)). Cette dichotomie nous permet d'identifier les deux types de liens présentés dans la section 3.2 : les liens intra-organisationnels • et les liens inter-organisationnels •. La seconde étape consiste à retirer les liens entre les membres qui ne sont pas affiliés au Loria (Figure 4.3(b)). Notre étude se concentre particulièrement sur la façon dont les membres d'une équipe donnée interagissent avec leurs alter ego (internes ou externes). C'est pourquoi nous prenons uniquement en compte les interactions impliquant au moins un membre du Loria. Enfin, pour la même raison que nous venons d'évoquer, nous comptabilisons les interactions entre les arêtes uniquement au niveau des chercheurs affiliés au Loria (Figure 4.3(c)). Ces trois étapes de pré-traitement nous permettent de définir les liens et leurs interactions qui structurent le réseau de collaboration d'une équipe. Nous obtenons en quelque sorte un graphe "ego centré" d'une équipe.

Le nombre d'auteurs référencés présente d'importantes variations d'une équipe à l'autre comme le montre les statistiques décrites par la Table 4.1. Comparés à la moyenne et à la médiane, les écarts types et les *Median absolute deviation*s (Mads) sont importants et indiquent par conséquent une grande dispersion du nombre d'auteurs qu'ils soient internes ou externes. C'est une caractéristique du jeu de données à considérer lorsque

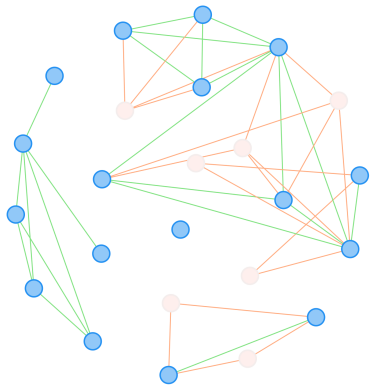
10. Pour chaque graphe, le nombre total de collaborateurs (internes/externes) est supérieur à 15



(a) *Étape 1 :* Nous distinguons deux types de nœuds, les membre du Loria (en bleu) et les collaborateurs affiliés à d'autres centres (en rouge).



(b) *Étape 2 :* Nous conservons uniquement les arêtes impliquant un chercheur du Loria



(c) *Étape 3 :* Les interactions entre les arêtes ne sont considérées qu'au niveau d'un chercheur du Loria.

FIGURE 4.3 – Un exemple de pré-traitement appliqué à un graphe de collaboration. Le graphe présenté par la Figure 4.3(a) représente le graphe des co-auteurs de l'équipe COAST. Les nœuds bleus ● représentent les membres du Loria, les nœuds rouges ● représentent les collaborateurs extérieurs. Les liens inter-organisationnels sont colorés en orange ●, et les liens intra-organisationnels sont quant à eux en vert ●.

nous comparons différentes équipes. Par ailleurs, nous souhaitons préciser que le nombre d'auteurs (membres du Loria) n'est pas une métrique représentative de l'effectif d'une équipe. Les données recueillies ne reflètent pas non plus l'activité de recherche d'une équipe et ne pourraient être considérées comme une indication juste du nombre et de la qualité des publications.

À partir de (3.2) (section 3.3), nous pouvons calculer les statistiques suffisantes de chaque graphe de collaboration. Ces statistiques suffisantes font partie des paramètres d'entrée de notre algorithme d'estimation. La taille du graphe de chaque équipe et leurs

TABLE 4.1 – Statistiques sur le nombre de collaborateurs (internes et externes) par équipe

	Moyenne	Médiane	Écart type	Mad
Membres du Loria	14.88	12.0	6.39	3.0
Collaborateurs extérieurs	22.35	18.0	15.51	8.0

TABLE 4.2 – Statistiques suffisantes des graphes de collaboration des 17 équipes

	t_{11}	t_{12}	t_{22}	Membres du Loria	Collaborateurs externes
BISCUIT	18	33	82	12	16
CAPSID	104	111	79	13	16
CARAMBA	3	4	5	9	9
COAST	55	61	36	17	7
GAMBLE	7	88	176	8	24
LARSEN	173	259	171	21	16
MADYNES	151	109	65	21	19
MAGRIT	29	0	8	10	8
MULTISPEECH	141	561	648	26	70
NEUROSYS	26	106	102	11	18
ORPAILLEUR	324	374	202	31	33
PESTO	18	36	282	12	33
RESIST	54	60	40	15	11
SEMAGRAMME	1	15	801	7	45
SIMBIOT	123	61	30	12	10
SMarT	78	301	325	11	19
VERIDIS	29	42	112	17	26

statistiques suffisantes sont présentés dans la Table 4.2. En effet, le paramètre y de l'algorithme ABC Shadow (Algorithme 6) encode à la fois les statistiques suffisantes et la taille du graphe observé. Mis à part le paramètre y, l'algorithme ABC Shadow a été exécuté sur les 22 graphes avec la même configuration. N'ayant aucune connaissance a priori de la valeur des paramètres, nous avons fixé $\theta_0 = [0, 0, 0]$. Le vecteur Δ qui définit le volume de la distribution de propositions a quant à lui été fixé à $\Delta = [0.005, 0.005, 0.005]$. Nous avons configuré l'algorithme de telle sorte que la longueur de la *shadow chain* soit $m = 200$ itérations. La variable auxiliaire est échantillonnée suivant un échantillonneur de Gibbs (Algorithme 2) qui exécute 500 itérations. L'algorithme ABC Shadow est paramétré ici pour échantillonner une chaîne de $iters = 1.002 \times 10^6$ échantillons. Les 2×10^3 premiers échantillons sont une étape transitoire (burn-in) et sont donc supprimés. Pour améliorer les propriétés de mélange et diminuer l'auto-corrélation, nous appliquons sur la chaîne résultante une procédure de sous-échantillonnage qui conserve tous les 10^3 échantillons. La distribution que nous traitons contient donc 10^3 échantillons.

Les distributions résultantes des paramètres θ_{11}, θ_{12} et θ_{22} sont représentées par les Figures 4.4, 4.5 et 4.6. En complément de ces figures, la moyenne, la médiane et l'écart type nous permettent de caractériser les distributions. Nous calculons le MAP, en prenant le maximum du noyau de densité des distributions. Le MAP correspond à la valeur de paramètre la plus vraisemblable parmi toutes les valeurs échantillonnées (sachant que notre loi a priori est uniforme). La moyenne, l'écart-type et la MAP de chaque paramètre échantillonné sont décrits par la Table 4.3.

Nous avons également calculé l'écart-type asymptotique ($\hat{\sigma}_\theta$) et l'écart-type de Monte-

TABLE 4.3 – Moyennes, médianes et estimateurs (MAP) calculés à partir des distributions échantillonnées par l'algorithme ABC Shadow

	$\overline{\theta_{11}}$	Q_{50} θ_{11}	MAP θ_{11}	$\overline{\theta_{12}}$	Q_{50} θ_{12}	MAP θ_{12}	$\overline{\theta_{22}}$	Q_{50} θ_{22}	MAP θ_{22}
BISCUIT	0.031	0.052	0.085	-0.372	-0.368	-0.364	-0.236	-0.233	-0.228
CAPSID	0.084	0.087	0.091	-0.245	-0.244	-0.237	-0.171	-0.168	-0.159
CARAMBA	-0.469	-0.352	-0.122	-1.199	-1.125	-0.961	-1.330	-1.283	-1.228
COAST	-0.088	-0.081	-0.060	-0.387	-0.386	-0.387	-0.042	-0.034	-0.026
GAMBLE	-0.989	-0.914	-0.692	0.078	0.066	0.054	-0.164	-0.161	-0.157
LARSEN	0.005	0.007	0.009	-0.204	-0.205	-0.204	-0.096	-0.094	-0.089
MADYNES	-0.008	-0.006	-0.004	-0.385	-0.383	-0.383	-0.364	-0.360	-0.344
MAGRIT	-0.036	-0.029	-0.018	-7.284	-6.666	-5.146	-0.812	-0.780	-0.724
MULTISPEECH	-0.010	-0.003	0.015	-0.134	-0.135	-0.137	-0.247	-0.247	-0.247
NEUROSYS	-0.121	-0.094	-0.022	-0.097	-0.104	-0.108	-0.216	-0.210	-0.200
ORPAILLEUR	-0.013	-0.013	-0.013	-0.254	-0.254	-0.252	-0.306	-0.304	-0.303
PESTO	0.109	0.127	0.150	-0.271	-0.265	-0.247	-0.173	-0.172	-0.174
RESIST	-0.018	-0.009	0.006	-0.381	-0.379	-0.371	-0.251	-0.242	-0.221
SEMAGRAMME	-0.780	-0.342	0.325	-0.178	-0.156	-0.126	-0.044	-0.044	-0.045
SIMBIOT	0.079	0.081	0.084	-0.333	-0.331	-0.326	-0.138	-0.129	-0.113
SMarT	0.076	0.087	0.109	-0.071	-0.074	-0.077	-0.015	-0.013	-0.009
VERIDIS	-0.004	0.008	0.023	-0.428	-0.425	-0.421	-0.396	-0.394	-0.396

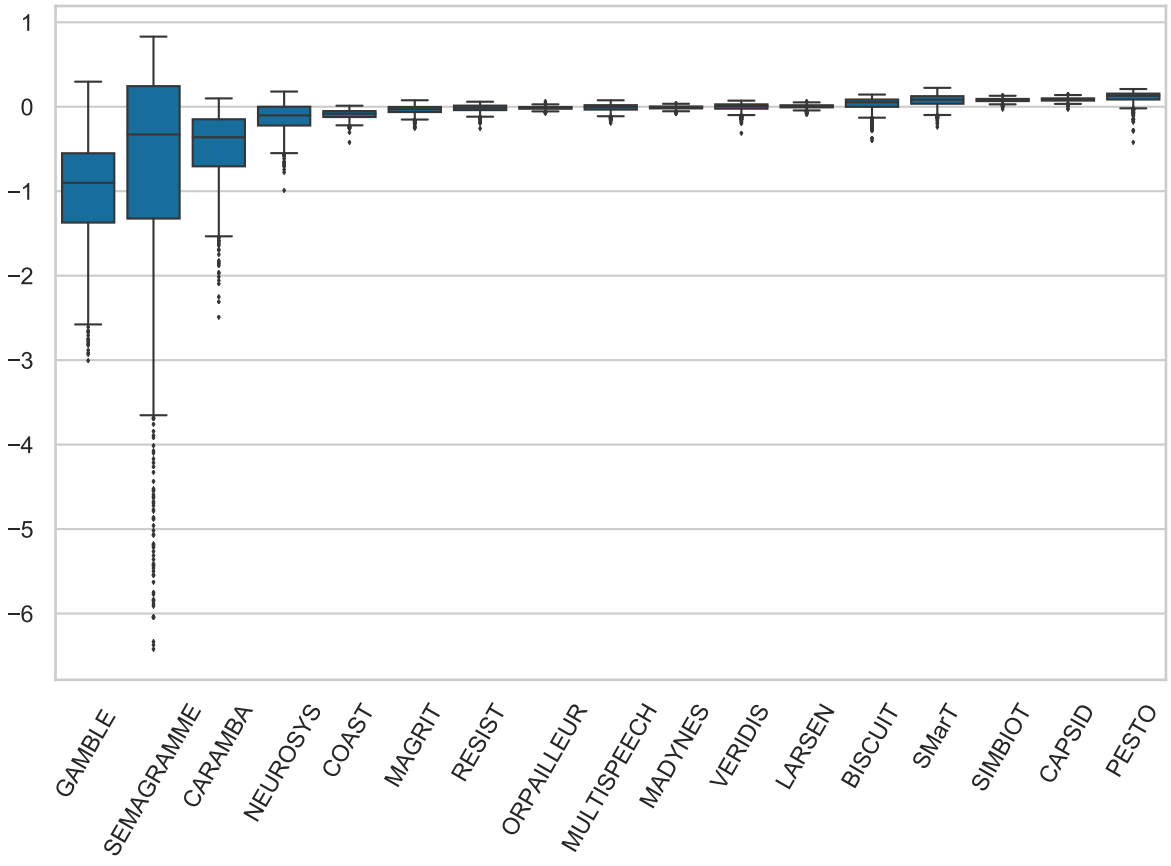


FIGURE 4.4 – Distributions *a posteriori* du paramètre θ_{11} pour chaque équipe (représentées par des boîtes à moustaches)

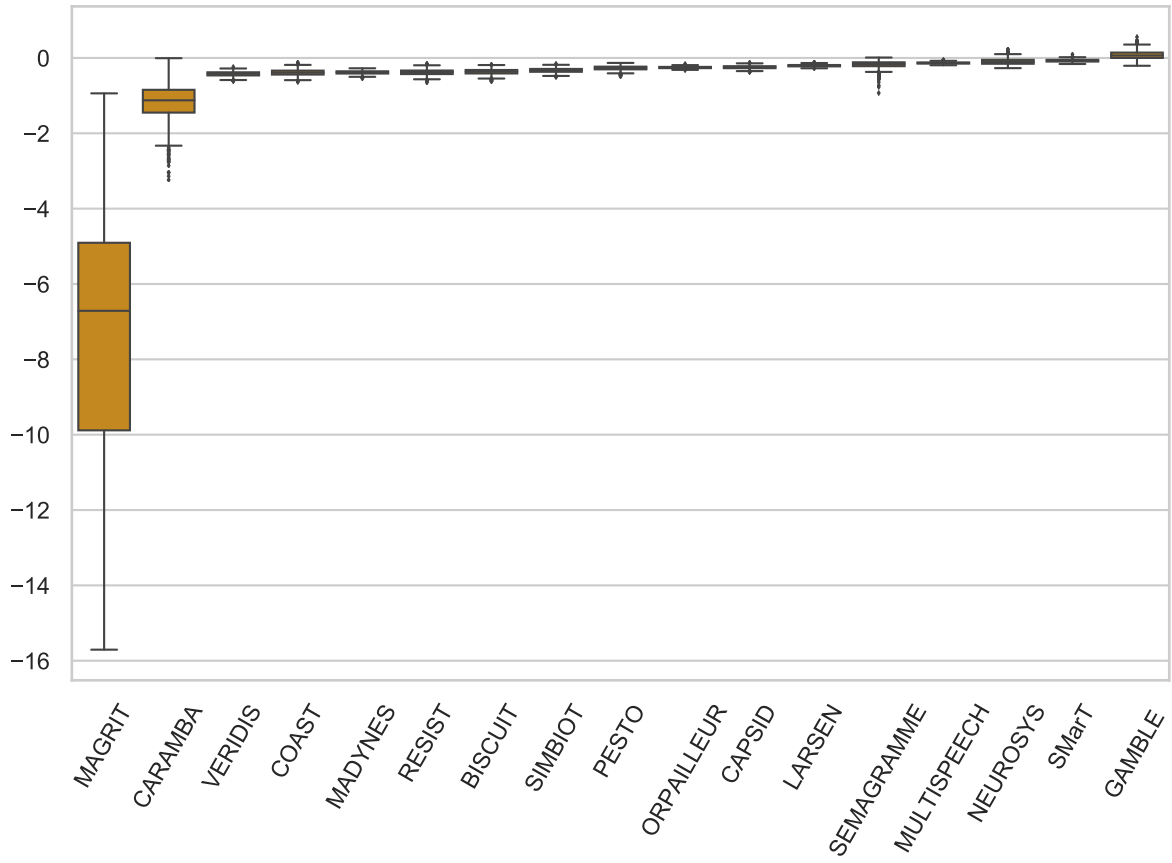


FIGURE 4.5 – Distributions *a posteriori* du paramètre θ_{12} pour chaque équipe (représentées par des boîtes à moustaches)

Carlo ($\hat{\sigma}_\theta^{MC}$) (section 2.3) pour chaque distribution échantillonnée. Toutes ces valeurs sont décrites dans la Table 4.4. Tout comme pour la section 3.5, le calcul de ces valeurs s'appuie sur des échantillons simulés à partir du MAP (Table 4.3) avec un échantillonneur de Gibbs. Pour décorréler les 10^7 échantillons générés par l'échantillonneur de Gibbs, nous avons appliqué une opération de sur-échantillonnage conservant tous les 10^3 échantillons. Le calcul des deux écarts-types a donc utilisé 10^4 échantillons. Ces mesures nous permettent de vérifier la bonne convergence des chaînes échantillonnées et de déterminer l'erreur de l'approximation. Nous remarquons par exemple que ces deux erreurs sont relativement élevées comparées aux autres pour les distributions du paramètre θ_{11} de l'équipe SEMAGRAMME et celle du paramètre θ_{12} de l'équipe MAGRIT. Ce qui suggère que ces deux chaînes de paramètre échantillonnées n'ont pas convergé et que l'estimation n'est pas suffisamment fiable. D'après la Table 4.2, les statistiques suffisantes observées pour l'équipe MAGRIT sont $t(y) = [29, 0, 8]$ et celles de l'équipe SEMAGRAMME sont $t(y) = [1, 15, 801]$. Nous pouvons constater que les statistiques t_{11} de l'équipe SEMAGRAMME et t_{12} de l'équipe MAGRIT sont très faibles, voire nulles, pour l'équipe MAGRIT. Nous pouvons vraisemblablement envisager que nous sommes confrontés à un problème de dégénérescence du modèle. Comme nous avons déjà

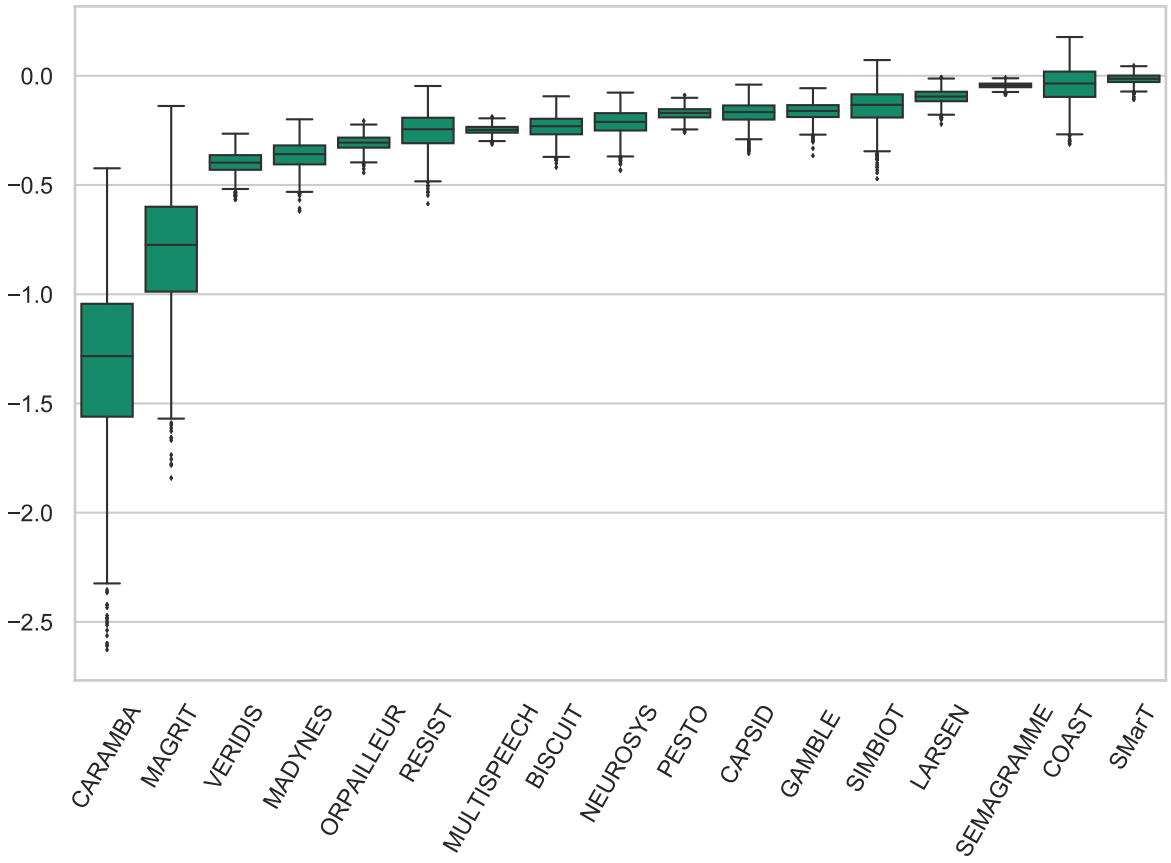


FIGURE 4.6 – Distributions *a posteriori* du paramètre θ_{22} pour chaque équipe représentées par des boîtes à moustaches)

eu l'occasion de l'évoquer dans la sous-section 2.2.2, les modèles d'interaction peuvent souffrir de dégénérescence [8, 97]. Dans les deux cas de figure que nous venons de voir, les statistiques suffisantes associées aux paramètres incriminés sont (presque) nulles. Il y a donc un décalage entre les phénomènes observés (le graphe dans son ensemble) et les phénomènes mesurés par les statistiques suffisantes. Ce qui laisse donc entendre que le modèle n'est pas en capacité d'expliquer les phénomènes observés. Nous employons alors le terme de "dégénérescence parfaite" pour qualifier cette situation. Les limites inhérentes au modèle et les problèmes de dégénérescence font partie des discussions que nous aborderons dans le Chapitre 6. À partir de l'écart-type de Monte-Carlo, nous pouvons déterminer les Intervalles de confiance sur nos estimateurs (MAPs). La Table 4.5 recense les MAPs estimé et les intervalles de confiance associés.

Les valeurs des paramètres renseignent sur la tendance à obtenir le motif structurel contrôlé par ce dernier dans le graphe observé. Par exemple, si le paramètre θ_{11} est positif, cela indique qu'il y a une propension positive à observer les interactions de type $1 \leftrightarrow 1$ et inversement si le paramètre est négatif. Si le paramètre est nul, cela signifie que l'occurrence du motif n'est pas contrôlée par le paramètre. En d'autres termes, il résulte d'un processus aléatoire à la manière d'un lancer de pièce. D'après les Figures 4.4, 4.5 et

TABLE 4.4 – Erreurs sur les estimations : l'écart-type asymptotique ($\hat{\sigma}_\theta$) et l'écart-type de Monte-Carlo ($\hat{\sigma}_\theta^{MC}$)

	$\hat{\sigma}_{\theta_{11}}$	$\hat{\sigma}_{\theta_{12}}$	$\hat{\sigma}_{\theta_{22}}$	$\hat{\sigma}_{\theta_{11}}^{MC}$	$\hat{\sigma}_{\theta_{12}}^{MC}$	$\hat{\sigma}_{\theta_{22}}^{MC}$
BISCUIT	1.963e-01	8.850e-02	5.44e-02	2.924e-03	8.155e-04	3.22e-04
CAPSID	2.967e-02	4.136e-02	4.835e-02	8.099e-05	1.301e-04	1.868e-04
CARAMBA	2.837e-01	3.374e-01	4.345e-01	7.01e-03	1.012e-02	1.75e-02
COAST	4.497e-02	7.670e-02	9.876e-02	2.284e-04	5.621e-04	8.406e-04
GAMBLE	6.543e-01	1.277e-01	4.696e-02	3.65e-02	6.340e-03	1.631e-03
LARSEN	3.145e-02	3.299e-02	3.328e-02	6.088e-05	7.646e-05	8.454e-05
MADYNES	2.838e-02	4.998e-02	5.840e-02	9.530e-05	2.154e-04	2.793e-04
MAGRIT	9.268e-02	1.097e+01	2.229e-01	1.397e-03	1.204e+01	1.462e-02
MULTISPEECH	4.966e-02	2.623e-02	1.953e-02	1.247e-04	5.338e-05	3.486e-05
NEUROSYS	1.523e-01	8.197e-02	6.534e-02	1.334e-03	5.676e-04	4.196e-04
ORPAILLEUR	2.302e-02	2.879e-02	3.478e-02	4.652e-05	6.45e-05	9.842e-05
PESTO	6.586e-01	6.924e-02	2.406e-02	4.074e-02	3.067e-03	2.254e-04
RESIST	4.408e-02	7.126e-02	8.792e-02	2.156e-04	4.614e-04	6.597e-04
SEMAGRAMME	2.66e+00	5.855e-02	9.73e-03	7.007e-01	9.874e-03	1.937e-04
SIMBIOT	1.860e-02	5.462e-02	7.81e-02	8.774e-05	2.881e-04	5.301e-04
SMarT	7.420e-02	3.702e-02	2.545e-02	2.822e-04	1.195e-04	7.135e-05
VERIDIS	2.132e-01	8.216e-02	5.170e-02	3.576e-03	8.594e-04	3.07e-04

TABLE 4.5 – Intervalles de confiances (95%) sur les MAPs estimés, calculés à partir de l'écart-type de Monte-Carlo (Table 4.4)

	CI 95% θ_{11}	CI 95% θ_{12}	CI 95% θ_{22}
BISCUIT	0.084862 ± 5.848e-03	-0.363798 ± 1.631e-03	-0.22772 ± 6.44e-04
CAPSID	0.09124 ± 1.62e-04	-0.236851 ± 2.603e-04	-0.159431 ± 3.736e-04
CARAMBA	-0.12204 ± 1.402e-02	-0.961354 ± 2.025e-02	-1.228014 ± 3.499e-02
COAST	-0.060408 ± 4.568e-04	-0.386647 ± 1.124e-03	-0.025746 ± 1.681e-03
GAMBLE	-0.691552 ± 7.3e-02	0.054031 ± 1.268e-02	-0.157195 ± 3.263e-03
LARSEN	0.008574 ± 1.218e-04	-0.20449 ± 1.529e-04	-0.089105 ± 1.691e-04
MADYNES	-0.003632 ± 1.906e-04	-0.383245 ± 4.308e-04	-0.343631 ± 5.586e-04
MAGRIT	-0.017512 ± 2.795e-03	-5.145882 ± 2.408e+01	-0.724146 ± 2.925e-02
MULTISPEECH	0.014521 ± 2.493e-04	-0.137195 ± 1.068e-04	-0.247163 ± 6.972e-05
NEUROSYS	-0.022392 ± 2.667e-03	-0.108188 ± 1.135e-03	-0.199922 ± 8.393e-04
ORPAILLEUR	-0.013457 ± 9.304e-05	-0.251772 ± 1.29e-04	-0.303298 ± 1.968e-04
PESTO	0.149809 ± 8.147e-02	-0.247339 ± 6.135e-03	-0.17373 ± 4.509e-04
RESIST	0.005723 ± 4.311e-04	-0.370663 ± 9.228e-04	-0.221469 ± 1.319e-03
SEMAGRAMME	0.324512 ± 1.401e+00	-0.126172 ± 1.975e-02	-0.045249 ± 3.873e-04
SIMBIOT	0.084257 ± 1.755e-04	-0.325838 ± 5.763e-04	-0.112576 ± 1.060e-03
SMarT	0.108987 ± 5.644e-04	-0.077073 ± 2.390e-04	-0.009359 ± 1.427e-04
VERIDIS	0.02327 ± 7.151e-03	-0.421076 ± 1.719e-03	-0.395927 ± 6.14e-04

TABLE 4.6 – Résultats du test de Student

	TS($\theta_{11}, 0$)	p-val1
MADYNES	-15.273	$\leq 10^{-6}$
LARSEN	5.820	$\leq 10^{-6}$
MAGRIT	-24.166	$\leq 10^{-6}$
MULTISPEECH	-7.951	$\leq 10^{-6}$
ORPAILLEUR	-26.193	$\leq 10^{-6}$
RESIST	-12.985	$\leq 10^{-6}$
VERIDIS	-1.911	5.631e-02

4.6 nous observons que la majeure partie des distributions sont négatives ou proches de zéro. Nous déduisons que de manière générale les tendances d'occurrence des trois motifs : $1 \leftrightarrow 1$, $1 \leftrightarrow 2$, $2 \leftrightarrow 2$ ont une faible tendance d'apparaître dans les graphes observées. Ce premier résultat s'explique par le fait que les 17 graphes observés sont faiblement denses. Replacée dans le contexte de l'activité de publications d'articles scientifiques, cette faible propension corrobore l'intuition selon laquelle un scientifique ne peut pas rédiger des articles avec tous ses homologues. La rédaction et le travail de recherche nécessaire en amont sont des tâches prenantes et difficiles. Ce qui limite par conséquent le nombre de publications et de collaborations sous-jacentes.

Certaines distributions sont très proches de zéro, voire presque centrées sur zéro. Nous observons ce résultat en particulier pour les distributions θ_{11} des équipes suivantes : LARSEN, MADYNES, MAGRIT, MULTISPEECH, ORPAILLEUR, RESIST et VERIDIS. Pour ces distributions, nous pouvons légitimement nous demander si le paramètre θ_{11} contrôle le motif $1 \leftrightarrow 1$. Autrement dit, est-ce que l'occurrence de ce motif ($1 \leftrightarrow 1$) pour ces équipes est le résultat d'une tendance marquée ou d'un processus "purement" aléatoire ? En des termes plus formels, nous voulons tester si l'espérance mathématique des échantillons qui composent ces distributions est significativement différente de zéro. Pour ce faire, nous utilisons un test paramétrique de Student. Nous posons l'hypothèse nulle $\mathcal{H}_0$ et l'hypothèse alternative $\mathcal{H}_1$ que nous voulons soumettre au test ainsi :

$$\mathcal{H}_0 : \mathbb{E}[\theta] = 0,$$
$$\mathcal{H}_1 : \mathbb{E}[\theta] \neq 0.$$

Les résultats du test sont présentés dans la Table 4.6. L'hypothèse nulle est rejetée pour les équipes : LARSEN, MULTISPEECH et ORPAILLEUR. Pour l'équipe VERIDIS, la valeur p (*p-value*) est légèrement supérieure au seuil de confiance de 95%. Pour cette équipe nous ne pouvons pas rejeter l'hypothèse nulle.

La Figure 4.7 présente les projections des MAPs estimés (Table 4.3) sur trois plans en deux dimensions. Chaque point projeté sur un plan représente une équipe. Les abscisses et les ordonnées de chaque plan représentent les valeurs des MAPs associés à un paramètre du modèle. Les lignes grises en pointillés délimitent les propensions d'occurrence positives et négatives des motifs contrôlés. Prenons comme exemple le premier plan (θ_{11}, θ_{22}) (situé tout à gauche), les équipes sont positionnées selon la valeur du MAP θ_{11} en abscisse et la valeur du MAP θ_{22} en ordonnée. La ligne verticale délimite les tendances, négative à gauche et positive à droite d'observer une occurrence du motif $1 \leftrightarrow 1$. De même, la ligne horizontale délimite les propensions d'occurrence positives (en haut) et négatives (en bas)

d'observer une occurrence du motif $2 \leftrightarrow 2$. Parmi toutes les équipes observées, aucune ne présente une tendance d'occurrence positive pour le motif $2 \leftrightarrow 2$. C'est pour cela qu'il n'y a aucun point (*i.e.* équipe) au-dessus de la ligne horizontale.

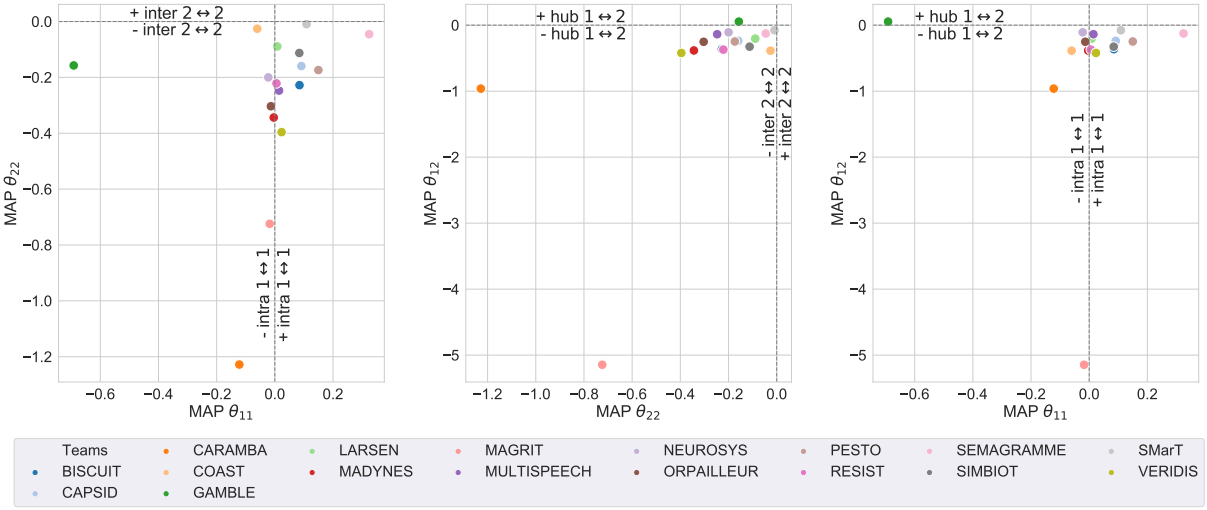
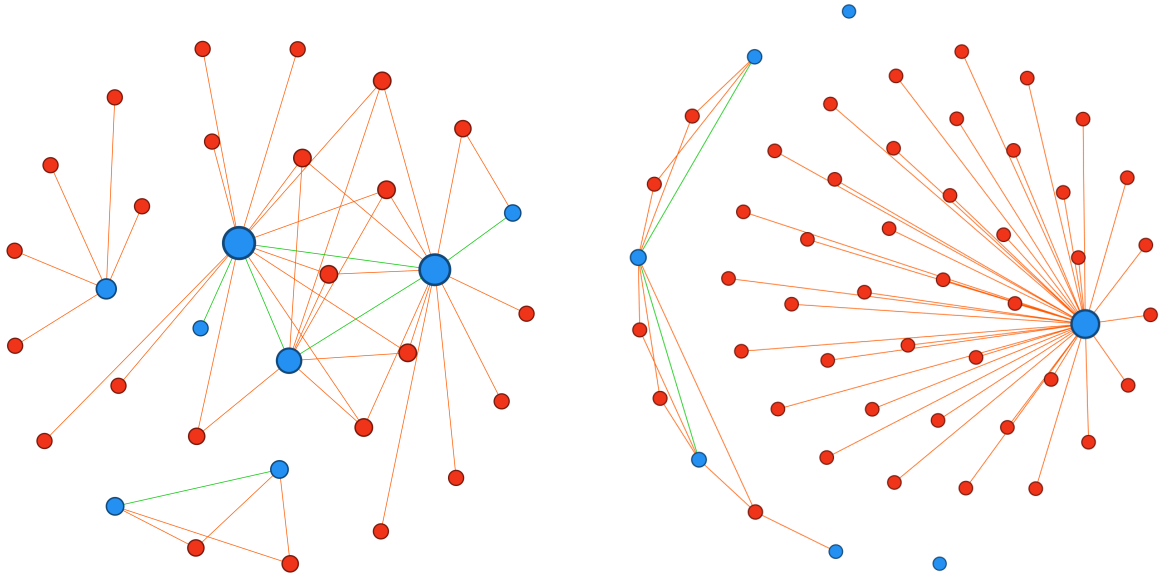


FIGURE 4.7 – Projections des MAP sur des plans de 2 dimensions

À partir des trois plans : $(\theta_{11}, \theta_{22})$, $(\theta_{22}, \theta_{12})$ et $(\theta_{11}, \theta_{12})$ illustrés par la Figure 4.7, nous avons un aperçu des propensions d'occurrence des motifs contrôlés pour l'ensemble des équipes. La majeure partie des équipes sont situées en dessous du seuil nul pour le motif $1 \leftrightarrow 2$ (ou hub). Cela préfigure une tendance générale : pour une majorité des équipes observées, le motif hub est moins susceptible de se produire que par pur hasard. Cette observation peut s'étendre au motif $2 \leftrightarrow 2$, pour lequel aucun des estimateurs calculés (MAPs) n'est positif. Les motifs impliquant des liens inter-organisationnels ($1 \leftrightarrow 2$ et $2 \leftrightarrow 2$) ont généralement une faible probabilité d'occurrence. Cela renforce notre intuition sur la faible occurrence des liens de collaboration. Les collaborations avec des équipes extérieures sont encore plus coûteuses et nécessitent du temps pour que ces travaux collectifs portent leurs fruits et que les résultats soient publiés.

Une exception subsiste tout de même, l'équipe GAMBLE se situe légèrement au-dessus du seuil nul pour le motif $1 \leftrightarrow 2$. Ce qui dénote une tendance inverse, légèrement positive. Nous remarquons tout d'abord dans la Table 4.2 que le graphe de collaboration de l'équipe GAMBLE contient une importante proportion de nœuds extérieurs (3/4 des nœuds). Cette disposition favorise l'émergence de liens inter-organisationnels constituants du motif hub. Ce n'est bien sûr pas la seule raison. Par exemple, le graphe de collaboration de l'équipe SEMAGRAMME comporte lui aussi (en proportion) un grand nombre de collaborateurs extérieurs. Il est nécessaire d'observer la façon dont les deux types de collaborateurs (internes/externes) interagissent les uns avec les autres. Les graphes de collaboration des équipes sont illustrés par la Figure 4.8. Nous avons appliqué sur ces graphes le prétraitement décrit par la Figure 4.3 et repris le même code de couleurs. Les membres de l'équipe GAMBLE (Figure 4.8(a)) représentés en bleu maintiennent presque tous des connexions avec des collaborateurs extérieurs. Parmi les membres de l'équipe, nous remarquons deux collaborateurs (nœuds) qui maintiennent plus de liens que leurs

homologues. Ils sont matérialisés par des nœuds de plus grande taille (proportionnelle au degré du nœud). Ces nœuds maintiennent à la fois des liens intra-organisationnels et inter-organisationnels et font donc apparaître de nombreux motifs hub. Le graphe de collaboration de l'équipe SEMAGRAMME présente une structure tout autre. La majorité des liens inter-organisationnels sont maintenus par un seul chercheur de l'équipe qui n'a (dans ce cas précis) pas publié avec d'autres membres de son équipe. Ces deux graphes illustrent des structures de collaboration orientées vers l'extérieur. De nombreux chercheurs extérieurs sont impliqués dans les activités de publication de ces deux équipes. Nous remarquons cependant des différences significatives entre les deux graphes. Ce qui souligne le fait que la structure de collaboration ne se définit pas uniquement par le type d'acteurs qui la compose. À proportion d'acteurs comparables, nous pouvons observer des structures relativement différentes.



(a) Équipe GAMBLE (b) Équipe SEMAGRAMME

FIGURE 4.8 – Graphes de collaboration de deux équipes. Les nœuds bleus ● représentent les membres de GAMBLE, les nœuds rouges ● représentent les collaborateurs extérieurs. La taille des nœuds est proportionnelle à leur degré. Les liens inter-organisationnels sont colorés en orange ●, et les liens intra-organisationnels sont quant à eux en vert ●. Les liens entre deux nœuds extérieurs ont été retirés, en accord avec le pré-traitement décrit par la Figure 4.3 .

Par ailleurs, nous observons sur les différentes projections de la Figure 4.7 des chevauchements de points ou tout du moins, des points qui sont relativement proches les uns des autres. Les estimateurs des équipes associés à ces points ont donc des valeurs relativement similaires. Ce qui suggère des similitudes de structures au niveau du graphe de collaboration. Notons néanmoins que cette observation se base uniquement sur la valeur des MAPs. Il pourrait être intéressant de vérifier ce constat en nous appuyant sur la distribution échantillonnée *a posteriori*. La forme de la distribution est aussi une source d'information qui renseigne sur les propensions d'occurrence des motifs structurels.

Dans la section suivante, nous proposons une méthode permettant d'identifier ces motifs récurrents à partir des distributions échantillonnées *a posteriori*.

4.4 Identification de motifs de collaboration communs à plusieurs équipes

Pour un paramètre donné, nous souhaitons comparer les distributions *a posteriori* des différentes équipes entre elles. En comparant les distributions deux à deux, nous voulons estimer la distance qui les sépare l'une de l'autre. Pour ce faire, nous utilisons le test statistique de K-S. Ce test (dans sa configuration deux échantillons) compare les fonctions de répartition empiriques des deux distributions d'entrée. Plus précisément, il calcule la distance maximale entre ces fonctions de répartition. Pour chaque paramètre du modèle, nous obtenons une matrice (en deux dimensions) qui contient les distances entre chaque distribution.

Ces matrices sont représentées sous la forme de cartes de chaleurs par les Figures 4.9, 4.10 et 4.11, respectivement pour les paramètres θ_{11}, θ_{12} et θ_{22}. Chaque case de la carte de chaleur est colorée en fonction de la valeur de la statistique K-S pour les deux distributions associées aux équipes en abscisse et en ordonnée. Comme indiqué par le gradient de couleur situé en haut à gauche de chaque figure, plus la valeur est petite, plus la couleur est sombre et inversement. Des dendrogrammes placés sur les côtés gauche et supérieur permettent d'estimer la distance entre deux sous-ensembles d'équipes. Plus les branches sont courtes, plus les sous-ensembles contenus par ses branches sont proches selon la statistique K-S. En résumé, les zones sombres de la carte de chaleur, représentés par des sous-arbres peu profonds dans les dendrogrammes représentent des groupes d'équipes (*clusters*) dont les distributions sont semblables (au sens de la statistique K-S).

Une première lecture de ces trois figures nous fait remarquer que les zones très sombres qui apparaissent sur les cartes de chaleur impliquent un (très) petit nombre d'équipes. Cela signifie que les clusters associés sont de très petites tailles, se réduisant pour la plupart à deux équipes. Les distributions échantillonnées diffèrent pour la plupart les unes des autres. Nous avons identifié les clusters à partir des dendrogrammes et selectionnées tous les sous arbres dont la profondeur est inférieure à un seuil fixé ($threshold = 0.5$).

Avant de nous concentrer plus en détail sur les clusters et leurs potentielles significations, il est intéressant de vérifier si les zones sombres des cartes de chaleur associent des équipes travaillant sur les mêmes thématiques. Si tel est le cas, il se peut que les thématiques de recherche aient une influence sur la façon dont les chercheurs interagissent entre eux. Pour rappel, le laboratoire est organisé en 5 départements. Ces départements définissent les principaux axes de recherche que couvrent les équipes qui les composent. Nous considérons que le département auquel appartient une équipe définit la principale thématique de recherche de cette dernière. Dans les trois figures, les labels des équipes sont colorés en fonction du département auquel elles appartiennent : Algorithmique, calcul, image et géométrie ●, Méthodes formelles ●, Réseaux, systèmes et services ●, Traitement automatique des langues et des connaissances ●, Systèmes complexes, intelligence artificielle et robotique ●. Une rapide lecture des trois figures

nous montre qu'il n'y a aucune évidence visuelle quant au fait que les équipes d'un cluster appartiennent au même département. Les thématiques de recherche ne semblent pas avoir d'influence sur les structures prises en compte par le modèle. Cette question sur le lien entre la structure de collaboration et les thématiques de recherche traitées par les équipes mérite sûrement d'être approfondie. Nous pourrions par exemple objecter que l'intitulé des départements ne définit pas assez clairement les questions de recherche sous-jacentes. D'autant plus que les équipes ne travaillent généralement pas sur une seule problématique. L'étude de cette question dépasse le cadre de nos travaux. Nous ne cherchons pas à y répondre, mais à présenter les perspectives que notre approche laisse entrevoir.

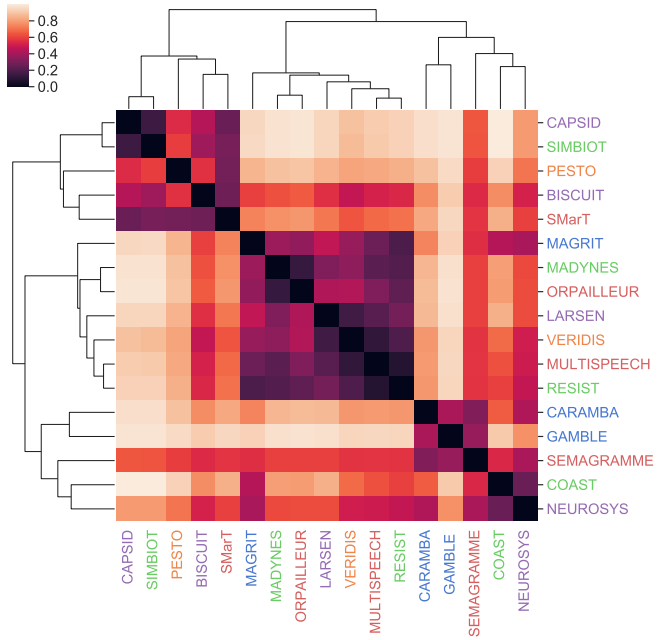


FIGURE 4.9 – Résultats de la classification hiérarchique calculée à partir de la distance de Kolmogorov-Smirnov entre les distributions (pour le paramètre θ_{11}). Le nom des équipes est coloré en fonction de la thématique de recherche.

La Figure 4.9 nous permet d'identifier trois clusters. Le premier, qui se situe dans la partie supérieure gauche de la carte de chaleur se compose seulement de deux équipes : CAPSID et SIMBIOT. Le second cluster se situe dans la zone centrale de la carte de chaleur est quant à lui constitué des équipes : MADYNES et ORPAILLEUR. Enfin, le dernier cluster se trouve légèrement en dessous du second et regroupe trois équipes : MULTISPEECH, RESIST et VERIDIS. Si nous observons la Figure 4.4, nous remarquons que les distributions échantillonnées à partir de ces équipes sont très proches les unes des autres. Ces distributions se situent sur une plage de valeurs très réduite. Dans ces conditions, il semble difficile d'identifier dans chaque cluster des caractéristiques remarquables. Si on se concentre sur l'interaction $1 \leftrightarrow 1$, il n'y a pas de différences sensibles entre les graphes de collaboration d'équipes appartenant à des clusters distincts. Notons néanmoins que la comparaison des distributions suivant la statistique K-S nous permet de détecter des clusters que nous ne pourrions pas identifier en nous appuyant uniquement sur

des statistiques telles que la moyenne ou la médiane. Ce constat est flagrant pour les deux derniers clusters. En effet, nous aurions tendance à penser qu'il y a une séparation nette dans la Figure 4.4 entre les distributions du second cluster et celles du troisième. En réalité, nous remarquons que les distributions des deux clusters s'entremêlent. Cela est dû au fait que la Figure 4.4 ordonne les boîtes à moustaches suivant la moyenne des distributions. Nous pouvons d'ailleurs remarquer que les boîtes à moustache illustrant les distributions des équipes ORPAILLEUR, MADYNES sont visuellement semblables. Il en est de même pour les équipes MULTISPEECH, RESIST et VERIDIS. La distribution *a posteriori* définit l'ensemble des configurations du modèle en mesure d'expliquer l'observation. En l'occurrence, les distributions portent ici sur l'ensemble des valeurs du paramètre θ_{11}. La forme de la distribution c'est-à-dire les valeurs échantillonnées et leur occurrence est bien plus informative que la moyenne ou la médiane. La comparaison des distributions dans leur ensemble est donc pertinente puisqu'elle permet d'extraire des informations qui ne sont pas forcément identifiables à la lecture de la Table 4.3 ou de la Figure 4.7.

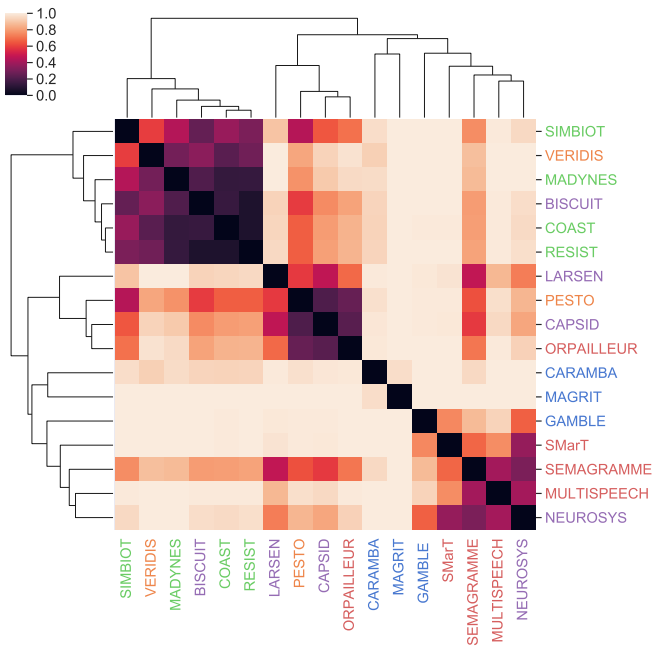


FIGURE 4.10 – Résultats de la classification hiérarchique calculée à partir de la distance de Kolmogorov-Smirnov entre les distributions (pour le paramètre θ_{12}). Le nom des équipes est coloré en fonction de la thématique de recherche.

Nous pouvons identifier deux autres clusters dans la Figure 4.10. Ces deux clusters impliquent respectivement les équipes : RESIST, COAST, BISCUIT et MADYNES ; et les équipes : CAPSID et ORPAILLEUR (pour le second cluseter). Nous remarquons que la carte de chaleur est beaucoup plus contrastée (alternance de couleurs sombres et claires), ce qui laisse entendre que les distributions pour le paramètre θ_{12} sont de formes variées. Ce constat peut d'ailleurs être entendu à la Figure 4.11 qui met également en lumière deux autres clusters : (GAMBLE, CAPSID, PESTO) et (BISCUIT, RESIST). Tout comme pour la première figure (Figure 4.9), il n'y a pas de caractéristiques visuellement

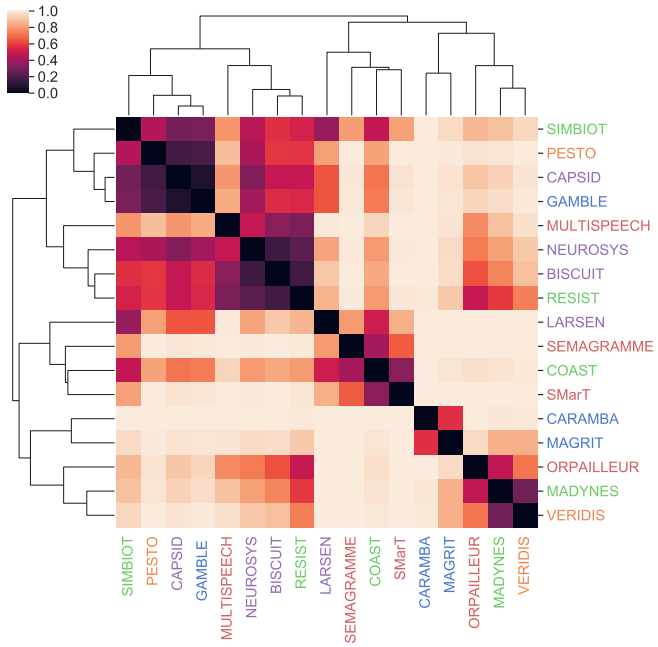


FIGURE 4.11 – Résultats de la classification hiérarchique calculée à partir de la distance de Kolmogorov-Smirnov entre les distributions (pour le paramètre θ_{22}). Le nom des équipes est coloré en fonction de la thématique de recherche.

remarquables entre les différents clusters. Que ce soit pour le paramètre θ_{12} ou θ_{22}, les distributions échantillonnées sont comprises pour la plupart dans un intervalle de valeurs réduit. Comme nous avons pu remarquer dans la section précédente les motifs $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$ ont une relative faible probabilité d'occurrence. C'est une caractéristique commune à la majorité des équipes. Il semble difficile dans ces conditions d'identifier des caractéristiques propres à un cluster.

Nous avons pu remarquer dans la section précédente que certains points de la Figure 4.7 se chevauchaient. Les valeurs des MAPs de certaines équipes sont donc très proches les unes des autres. C'est particulièrement le cas des équipes PESTO et CAPSID dans le plan du milieu $(\theta_{22}, \theta_{12})$. À la lecture de la Figure 4.11 nous avons remarqué que ces deux équipes faisaient partie du même cluster confirmant ainsi la proximité suivant les distributions du paramètre θ_{22}. Même si elle ne font pas partie du même cluster pour le paramètre θ_{12}, nous pouvons remarquer que la distance entre les deux distributions est relativement réduite suggérant ainsi qu'elle sont très proches l'une de l'autre. Ce qui nous permet de conclure que les deux équipes partagent des caractéristiques communes suivant les interactions $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$.

Les distributions échantillonnées par ABC Shadow dans la section précédente représentent l'ensemble des configurations du modèle capables d'expliquer les observations associées. La pertinence de ces configurations est pondérée par leur occurrence au sein de la distribution. La forme de ces distributions apporte donc une information sur les propriétés mesurées par le modèle. Pour un paramètre donné, si deux distributions sont proches l'une de l'autre, cela signifie que le phénomène contrôlé par le paramètre (en l'occurrence des interactions dans notre cas) est présent dans une même mesure dans

les deux observations. Dans cette section, nous avons calculé les distances entre les distributions échantillonnées en nous appuyant sur la statistique de Kolmogorov-Smirnov. À partir de ces mesures, nous avons appliqué une méthode de classification hiérarchique pour identifier les groupes de distributions semblables. De cette manière, il est possible d'identifier des groupes d'équipes partageant des propriétés sur les interactions communes. Dans notre cas, les distributions échantillonnées évoluent sur un faible intervalle de valeurs. Dans ces conditions, il est difficile d'identifier des caractéristiques suffisamment marquées d'un groupe à l'autre. Cette section a principalement pour objectif de montrer une application sur les distributions *a posteriori* échantillonnées. La méthode que nous proposons peut être réinvestie avec d'autres modèles. Elle peut également être utilisée pour étudier des évolutions dans le temps à partir d'observations temporelles.

4.5 Synthèse

Les collaborations entre les chercheurs peuvent s'illustrer au travers des publications scientifiques qui ont été produites. Plus précisément, nous pouvons considérer que les auteurs d'un même article ont tous collaboré pour mener à bien le travail de recherche et de rédaction. Ainsi, à partir d'une collection d'articles scientifiques, nous pouvons tracer un graphe illustrant les liens de collaboration tissés pas les auteurs. Ce type de collaborations dépasse très souvent les frontières institutionnelles et constitue en cela un très bon exemple de collaboration inter-organisationnelle. Ces collaborations peuvent à la fois être vues comme un moyen de partager des connaissances, mais également comme une occasion d'accroître la visibilité des résultats et du travail de recherche. Elles peuvent alors être perçues comme un double enjeu au sein des communautés scientifiques. L'objet de notre travail porte sur les interactions initiées par les membres de plusieurs équipes de recherche. Nous cherchons à caractériser l'occurrence des interactions au sein de ces équipes (lien intra-organisationnel) et avec l'extérieur (lien inter-organisationnel). À cette fin, nous mettons en œuvre l'approche décrite dans le Chapitre 3 pour chaque équipe considérée et exploitons les distributions échantillonnées à travers des mesures statistiques.

Nous nous sommes intéressés aux collaborations initiées par les chercheurs d'un laboratoire d'informatique, le Loria. Les données ont été collectées à partir de la base de connaissance mise à disposition par l'archive ouverte HAL. Nous avons sélectionné tous les articles de journaux ou de conférence déposés en 2018 impliquant au moins un chercheur affilié au Loria. Grâce aux métadonnées recueillies, nous avons pu tracer les graphes de collaboration de chaque équipe du Loria. Ces graphes sont composés des membres de l'équipe en question et des collaborateurs extérieurs, c'est-à-dire non affiliés au Loria.

Les graphes de collaboration générés à l'issue de la collecte des données constituent nos observations sur lesquelles nous avons appliqué notre approche de modélisation et d'inférence. Les valeurs des paramètres échantillonnées nous permettent de caractériser l'occurrence des interactions prises en compte par le modèle. Pour la majeure partie des équipes, les interactions considérées par le modèle ont une faible probabilité de se produire. Ce constat est plus marqué pour les interactions $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$ qui impliquent des liens inter-organisationnels. Si les collaborations inter-organisationnelles représentent un enjeu pour les équipes de recherche, la faible présence de hub au sein de

ces dernières peut représenter une menace pour la pérennité des collaborations. Si quelques acteurs jouant le rôle de hub venaient à quitter l'équipe, certains liens, notamment avec l'extérieur, pourraient être rompus sans que d'autres membres puissent prendre le relais. Ce type de problème se retrouve également dans la dynamique des réseaux pair-à-pair. Les membres d'un réseau pair-à-pair se connectent et se déconnectent plus ou moins régulièrement. Ce phénomène de *churn* peut alors affecter la disponibilité des données hébergées par un utilisateur d'autant plus lorsque ces déconnexions sont accidentelles et donc, non prévues. Ce problème a notamment était étudié dans le domaine des *Distributed Online Social Networks*s [40, 88]. Ces solutions reposent essentiellement sur une stratégie de sélections de pairs de confiance sur lesquels seront répliquées les données d'intérêt. Les technologies pair-à-pair telles que les DOSNs reposant sur des topologies *friend-to-friend* (F2F) (littéralement : "ami à ami") sont des candidats intéressants pour supporter (technologiquement) de telles collaborations. Dans ce cas de figure, la topologie du réseau représente les liens sociaux et permet de ce fait de préserver le caractère privé des informations partagées au sein des différents cercles d'amis. En extrapolant, nous pourrions envisager un tel dispositif pour ces collaborations et considérer la question de la réplication, non pas uniquement des données, mais aussi des liens inter-organisationnels dans le sens où leur disponibilité constituent un enjeu pour la collaboration.

Enfin, nous avons comparé les distributions échantillonnées entre elles de manière à identifier des caractéristiques structurelles communes à plusieurs équipes. Nous avons tout d'abord constaté qu'il n'y avait pas de lien explicite entre les thématiques de recherche et les structures étudiées. Nous avons pu identifier quelques clusters à partir des mesures de distance entre les distributions. Les valeurs échantillonnées sont définies sur un intervalle de valeurs trop restreint pour que des différences soient visuellement perceptibles entre les différents clusters. Néanmoins, nous avons pu montrer que l'exploitation des distributions dans leur intégralité permettait de compléter les observations réalisées sur les statistiques telles que la moyenne ou le MAP.

Dans ce chapitre nous avons eu l'occasion de d'utiliser une méthode bayésienne pour étudier la structure des liens de collaboration. Ce travail soulève plusieurs questions. La première porte sur la taille des équipes et des collaborations, qui comme nous avons pu le remarquer sont relativement disparates. Nous pourrions approfondir cette réflexion en proposant une procédure de normalisation des données. D'autres perspectives sont envisageables portant notamment sur le choix du modèle. En effet, nous pourrions considérer d'autres classes de modèles [68] et les mettre en perspective avec le modèle que nous proposons.

Dans le chapitre suivant (Chapitre 5), nous appliquons notre approche à un autre jeu de données représentant les interactions entre les élèves d'une école primaire. Le contexte et la thématique de cette application ne se rapportent plus à l'étude des collaborations. Nous souhaitons ainsi montrer que l'approche présentée dans ce travail de thèse peut s'appliquer à l'étude d'autres structures relationnelles. Nous présentons également dans ce chapitre une tentative d'amélioration de l'algorithme de simulation. En effet, l'échantillonnage de la variable auxiliaire se révèle être l'étape la plus coûteuse d'ABC Shadow. Nous proposons une implantation parallèle de l'échantillonneur de Gibbs.

Chapitre 5

Extension de l'approche

Sommaire

5.1 Motivation

Jusqu'à présent, nous avons démontré l'approche de modélisation et d'inférence (décrite dans le Chapitre 3) sur un cas d'étude particulier : les collaborations entretenues par les chercheurs dans le cadre de leur activité de publication (Chapitre 4). Cette approche peut aussi s'appliquer à d'autres contextes impliquant d'autres types de relations. Nous proposons ici d'étudier une structure relationnelle qui représente les interactions entre les élèves d'une école primaire. Des chercheurs ont enregistré pendant deux jours tous les contacts qui se sont produits entre les élèves d'une école primaire. Cette collecte de données s'est déroulée dans le cadre d'une étude [94] visant à mesurer et étudier les contacts entre les enfants en milieu scolaire pouvant favoriser la propagation de maladies respiratoires. Nous considérons les classes comme structures organisationnelles et comptabilisons les interactions au sein de chaque classe (liens intra-organisationnels) et entre les classes (liens inter-organisationnels). À partir de notre modèle d'interactions (section 3.3), nous cherchons à déterminer s'il existe des motifs communs entre les différentes classes notamment celles de même niveau. Nous souhaitons également observer si la méthode permet de déterminer des motifs communs entre les classes et selon quels facteurs. La mise en œuvre de notre approche sur ce nouveau cas d'étude est décrite dans la section 5.2.

Nous avons également travaillé sur une perspective d'amélioration de l'algorithme d'inférence permettant de traiter des graphes de plus grande taille dans un temps raisonnable. L'échantillonnage de la variable auxiliaire se trouve être l'étape la plus coûteuse en temps dans l'algorithme ABC Shadow. L'échantillonneur de Gibbs (Algorithme 2) qui est utilisé pour la simulation de la variable auxiliaire évalue séquentiellement le voisinage de tous les nœuds du graphe. Pour chaque nœud, l'échantillonneur détermine quel est le label du nœud le plus vraisemblable d'après les paramètres du modèle sachant l'état du voisinage. Le temps de simulation de la variable auxiliaire dépend donc linéairement du nombre de nœuds dans le graphe. Pour atténuer ce problème, nous proposons une version parallélisée de l'échantillonneur de Gibbs qui s'inspire des travaux de Bauer, Fürlinger et Kauermann [4]. Il est possible d'évaluer simultanément les nœuds indépendants, c'est-à-dire tous les nœuds n'étant pas directement voisins les uns des autres. Une présentation détaillée de l'algorithme ainsi que les résultats observés sont présentés dans la section 5.3.

5.2 Étude des interactions entre les élèves d'une école primaire

5.2.1 Contexte et travaux initiaux

Dans cette section, nous proposons d'étudier un autre type de graphe social qui porte sur les contacts (face à face) qui se sont produits au cours de deux journées consécutives entre les élèves d'une école primaire. Nous cherchons à vérifier si à travers le modèle que nous proposons (section 3.3), il existe des motifs de contacts récurrents entre les classes et dans le temps.

Pour ce faire, nous nous appuyons sur les données recueillies dans le cadre du projet : *The SocioPatterns project* [93]. Les données ont été collectées durant deux journées consécutives, le 1^{er} et 2 octobre 2009 dans une école primaire de Lyon. L'école est organisée en 10 classes, deux classes par niveau, allant du CP au CM2. La collecte de données visait à identifier les contacts physiques face à face à l'aide de dispositifs RFID placés sur chaque élève et enseignant(e) participant à l'expérience. Les badges RFID émettaient un signal toutes les 20 secondes. Pour qu'un contact entre deux personnes soit détecté, il fallait nécessairement que les badges RFID aient échangé au moins un paquet durant ces 20 secondes. Le dispositif mesurait également la durée de chaque contact. À partir de cette mesure il est possible de calculer le temps de contact cumulé : le temps total pendant lequel un individu est resté en contact avec un autre individu. Les contacts ont été enregistrés uniquement au sein de l'école, entre 8h45 et 17h20 pour la première journée et entre 8h30 et 17h05 pour la seconde journée. La plupart des enfants (96%) et tous les enseignant(e)s (100%) ont pris part à l'expérience.

Ces données ont permis à Stehlé *et al.* [94] de dresser une analyse précise des contacts en milieu scolaire afin d'identifier les motifs d'interactions qui pourraient favoriser la propagation de maladies respiratoires. Cette étude a notamment montré que les contacts entre les enfants se produisaient principalement entre les élèves d'une même classe. En considérant le temps de contact cumulé, nous remarquons qu'une partie non négligeable des contacts implique des élèves de même niveau. Il a également été montré que les temps de contact étaient assez brefs, moins d'une minute pour 88% des contacts. Les temps de contacts sont relativement disparates s'étalant sur une échelle de temps pouvant aller de 20 secondes à plus d'une vingtaine de minutes. Ce phénomène a également été constaté pour les temps de contact cumulés. Enfin, les auteurs ont également cherché à observer l'évolution de ces contacts entre la première et la seconde journée. Cette comparaison a mis en lumière certains comportements récurrents qui concernent le nombre de contacts, le temps passé en contact avec d'autres individus et le nombre d'individus distincts avec qui il y a eu un contact. Il existe néanmoins une part non négligeable des contacts qui ont été renouvelés pendant la seconde journée. Toutes ces observations permettent de faciliter la prise de décisions pour contenir une épidémie. Par exemple en cas d'épidémie, plutôt que de fermer toute l'école, il pourrait être envisageable de cibler les classes vecteur de la maladie, puisque l'essentiel des contacts se produisent au sein de la classe. Certains résultats ne sont pas mentionnés ici et nous encourageons les lecteurs et lectrices qui le souhaitent à se référer à la publication originale [94] afin d'apprécier l'étendue du travail réalisé par les auteurs.

Nous utilisons ici le jeu de données pour un autre propos. Nous ne considérons pas les contacts entre les enfants uniquement comme vecteur d'une maladie, mais plus largement comme une représentation d'un comportement social. En adoptant une approche similaire à celle décrite dans le Chapitre 4, nous souhaitons évaluer dans quelle mesure les élèves d'une classe interagissent au sein de leur propre classe et avec des élèves d'autres classes. En reprenant la terminologie évoquée dans le chapitre précédent, nous associons les classes à des organisations. Nous définissons les liens entre les élèves d'une même classe comme des liens intra-organisationnels (ou intra-classes). Réciproquement, les liens entre des élèves n'appartenant pas à la même classe sont dits inter-organisationnels (ou inter-classes). Dans ce cadre, nous pouvons reprendre le modèle d'interactions décrit dans la section 3.3.

Le motif $1 \leftrightarrow 1$ désigne dans ce contexte le fait pour un élève d'interagir avec deux de ses camarades appartenant à la même classe. Le motif $1 \leftrightarrow 2$ représente cette fois-ci un élève qui interagit à la fois avec un élève de sa classe et un second élève d'une autre classe. Enfin, le motif $2 \leftrightarrow 2$ représente un élève qui interagit avec deux autres élèves n'étant pas dans la même classe que lui. Les paramètres θ_{11}, θ_{12} et θ_{22} associés aux statistiques t_{11}, t_{12} et t_{22} contrôlent respectivement la prépondérance des motifs $1 \leftrightarrow 1$, $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$ dans le graphe (3.1).

Nous distinguons les élèves de chaque classe et observons les interactions initiées par ses élèves. En appliquant le modèle d'interactions au graphe de contacts de chaque classe, nous cherchons à observer des motifs récurrents entre les classes notamment celles de même niveau. Nous voulons également observer l'évolution de l'occurrence de ces motifs entre la première et la seconde journée afin de vérifier s'il existe une constance ou non dans la façon dont les élèves interagissent. Enfin, nous souhaitons mettre en perspective ces observations avec les résultats proposés par Stehlé *et al.* [94]. Un des résultats avancés porte par exemple sur la tendance pour un élève à avoir des contacts préférentiellement avec des membres de sa classe plutôt que des élèves d'autres classes. Nous voulons retrouver ce résultat en comparant la valeur estimée du paramètre θ_{11} avec celles des paramètres θ_{12} et θ_{22} pour toutes les classes.

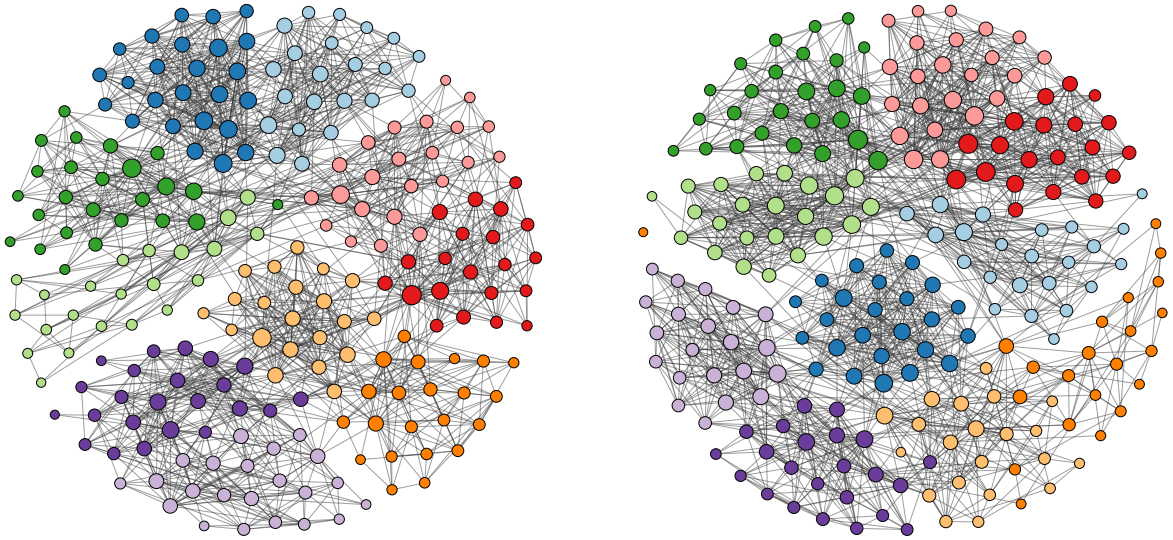
5.2.2 Présentation des données

Les données se trouvent sous la forme de graphes qui recensent tous les contacts qui se sont produits au cours d'une journée. Deux graphes composent ce jeu de données, un pour chaque journée. Les données permettant de construire ces graphes sont disponibles sur la page de l'article original [11]. Les nœuds des graphes représentent les élèves ou les enseignant(e)s. Deux nœuds sont connectés par une arête si et seulement si au moins un contact s'est produit entre les deux individus correspondants. Un graphe donne un aperçu des contacts cumulés dans une journée. Cela signifie que pour une arête donnée il a pu se produire plusieurs contacts. Pour compléter la représentation graphique, deux attributs caractérisent chaque arête. L'attribut `duration` correspond à la durée totale des contacts qui se sont produits entre les deux nœuds placés à chaque extrémité. L'attribut `count` renseigne quant à lui sur le nombre de contacts qui se sont produits entre les deux nœuds. Ce sont en tout 37351 contacts qui ont été enregistrés et encodés dans le graphe de la première journée et 40064 contacts dans celui de la seconde journée. Le graphe qui agrège les contacts de la première journée compte 236 nœuds (dont 10 enseignant(e)s et 226 élèves) et 5899 arêtes. Pour celui de la seconde journée, 238 nœuds (dont 10 enseignant(e)s et 228 élèves) et 5539 arêtes sont encodés.

Nous avons choisi de conserver les arêtes représentant des contacts assez "forts", des contacts dont la durée cumulée est d'au moins deux minutes. Ce seuil de deux minutes a initialement été choisi par les auteurs du papier original. Il permet de supprimer tous les contacts éphémères qui ne reflètent pas réellement un lien social. À l'issue de cette première opération, 3943 arêtes ont été supprimées dans le premier graphe et 3362 dans le second. Nous avons aussi retiré les nœuds représentant les enseignant(e)s. Stehlé *et al.*

11. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023176

[94] montrent que les enseignant(e)s n'interagissent pas de la même manière que les élèves. Ils interagissent principalement avec leurs élèves et beaucoup moins avec leurs collègues. Les motifs d'interactions des enseignant(e)s ne semblent pas comparables avec ceux des enfants. La Figure 5.1 représente les graphes résultants. Nous remarquons visuellement que les classes de même niveau semblent proches les unes des autres dans le sens où elles tissent de nombreux liens entre elles. Ce qui semble corroborer les propos des auteurs de [94].



(a) Jour 1 (b) Jour 2

FIGURE 5.1 – Graphes de contacts cumulés pendant deux journées dans une école primaire. Les Figures 5.1(a) et 5.1(b) représentent respectivement les contacts agrégés pendant la première et seconde journée. Les nœuds représentent les enfants ayant pris part à l'expérience. Ils sont colorés en fonction de leur classe : CP A ●, CP B ●, CE1 A ●, CE1 B ●, CE2 A ●, CE2 B ●, CM1 A ●, CM1 B ●, CM2 A ●, CM2 B ●. Les nœuds symbolisant les enseignant(e)s, de même que les arêtes représentant des contacts dont la durée cumulée est inférieure à 2 minutes ont été retirés des graphes. La taille des nœuds est proportionnelle à leur degré.

À partir des deux graphes illustrés par la Figure 5.1, nous avons calculé le graphe de contacts de chaque classe. Pour cela, nous avons suivi le même pré-traitement que celui mis en œuvre pour calculer les graphes de collaboration des équipes de recherche dans la section 4.3. Pour chaque classe :

Étape 1 : Nous distinguons les élèves appartenant à la classe courante et ceux issus des autres classes. Les arêtes reliant deux nœuds affiliés à la classe courante sont dites intra-organisationnelles. Les autres arêtes sont inter-organisationnelles.

Étape 2 : Toutes les arêtes qui relient exclusivement des élèves n'appartenant pas à la classe courante sont supprimées. Ce qui signifie que seules les arêtes dont au moins un membre de la classe courante est placé à une des extrémités sont conservées.

TABLE 5.1 – Statistiques suffisantes des graphes de contacts des 10 classes de l'école primaire collectés durant deux journées.

classe	période	t_{11}	t_{12}	t_{22}	Élève de la classe	Contacts extérieurs à la classe
CP A	jour 1	2021	1394	431	22	45
	jour 2	1657	876	290	23	43
CP B	jour 1	4134	1947	474	25	36
	jour 2	3694	1682	344	25	50
CE1 A	jour 1	543	648	352	22	37
	jour 2	3556	2050	672	23	45
CE1 B	jour 1	1739	1165	344	25	45
	jour 2	2672	1825	601	26	41
CE2 A	jour 1	1324	1207	464	23	42
	jour 2	2472	2488	1050	23	40
CE2 B	jour 1	1645	1230	381	21	38
	jour 2	2916	2433	1018	21	46
CM1 A	jour 1	2201	1138	292	21	35
	jour 2	904	1003	475	21	37
CM1 B	jour 1	1079	572	202	22	25
	jour 2	396	284	191	21	23
CM2 A	jour 1	1287	950	331	22	26
	jour 2	2301	1108	271	21	26
CM2 B	jour 1	1655	1455	510	23	41
	jour 2	1733	1141	336	23	37

Étape 3 : Nous calculons les statistiques suffisantes (t_{11}, t_{12} et t_{22}) en comptabilisant les interactions $1 \leftrightarrow 1$, $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$ uniquement au niveau des nœuds affiliés à la classe courante.

La Table 5.1 recense les statistiques suffisantes et la taille des graphes de chacune des classes pour la première et la seconde journée. Ces statistiques suffisantes donnent une représentation des graphes de contacts observés pour chaque classe. Elles sont fournies en entrée de l'algorithme d'inférence ABC Shadow (paramètre y de l'Algorithme 7). Dans la sous-section suivante (sous-section 5.2.3), nous présentons et analysons les estimations fournies par ABC Shadow. Les paramètres estimés définissent les proportions d'occurrence des interactions. À l'aide d'un algorithme de simulation comme Metropolis-Hastings ou un échantillonneur de Gibbs nous pouvons avec les paramètres estimés reproduire les observations, c'est-à-dire générer des graphes dont les statistiques suffisantes sont identiques à celles observées.

5.2.3 Analyse et résultats

L'algorithme ABC Shadow (Algorithmes 6 et 7) a été exécuté pour chaque statistique suffisante (détaillée dans la Table 5.1). La paramétrisation est semblable à celle utilisée pour l'étude des collaborations présentée dans le chapitre précédent (Chapitre 4). Les paramètres initiaux θ_0 sont fixés à $\theta_0 = [0, 0, 0]$. La distribution de proposition est une loi uniforme de volume $\Delta = [0.001, 0.005, 0.005]$. La longueur de la *shadow chain* est de $m = 200$ itérations. La variable auxiliaire est simulée avec un échantillonneur de Gibbs qui effectue 500 itérations. ABC Shadow échantillonne $iters = 1.002 \times 10^6$ échantillons. Les 2×10^3 premiers échantillons sont évincés. À cette première opération, s'ajoute une procédure de sous-échantillonnage qui conserve tous les 10^3 échantillons. La distribution

a posteriori qui est traitée est une chaîne de paramètres qui contient 10^3 échantillons. Les distributions sont représentées sous la forme de boites à moustache. Les Figures 5.2, 5.3 et 5.4 représentent respectivement les distributions échantillonnées des paramètres θ_{11}, θ_{12} et θ_{22}. Les distributions sont regroupées par classe et les deux périodes (jour 1 et jour 2) sont distinguées par deux teintes différentes. En complément de ces figures, la Table 5.1 présente la moyenne, la médiane et le MAP des distributions échantillonnées. Nous considérons le MAPs comme un estimateur de vraisemblance et avons calculé l'écart-type asymptotique ($\hat{\sigma}_\theta$) et l'écart-type de Monte-Carlo ($\hat{\sigma}_\theta^{MC}$). Les valeurs numériques sont présentées en annexe dans la Table B.1. Tout comme pour la précédente application (section 4.3), ces calculs se basent sur 10^7 échantillons simulés avec un échantillonneur de Gibbs dont seuls tous les 10^3 échantillons ont été conservés (soit une chaîne résultante de 10^4 échantillons). Les intervalles de confiance sont également disponibles dans l'Annexe B, présentés par la Table B.2.

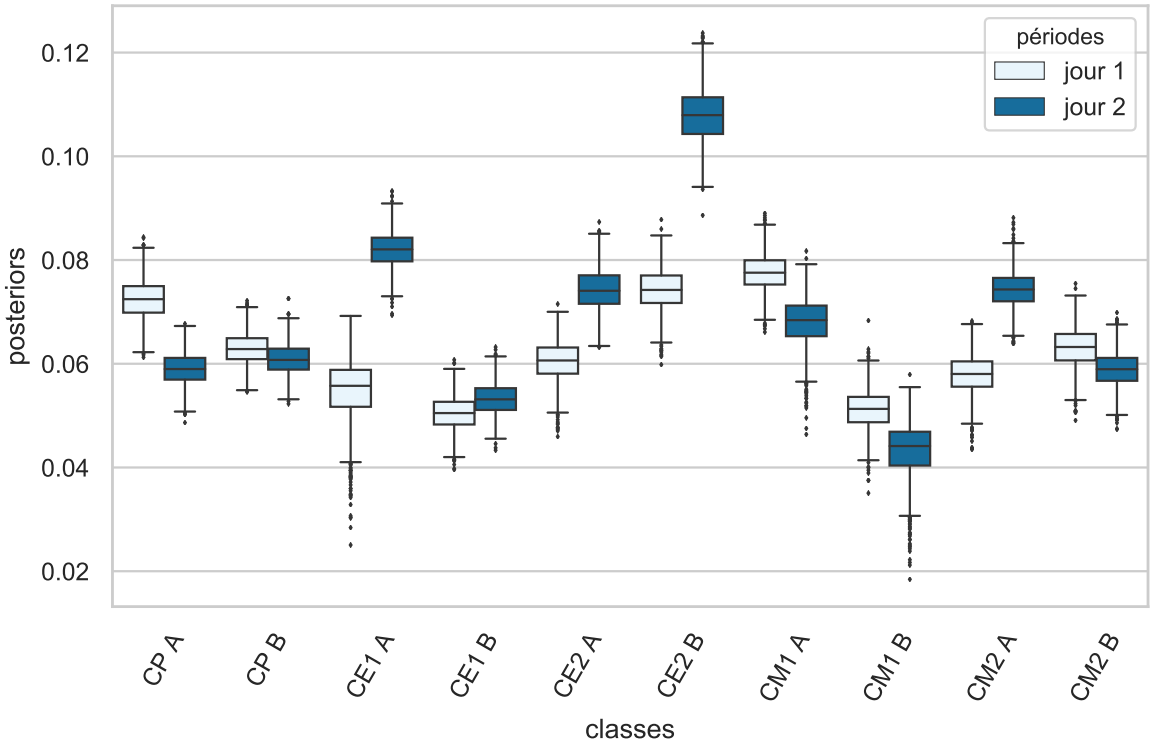


FIGURE 5.2 – Distributions *a posteriori* du paramètre θ_{11} pour chaque classe (représentées par des diagrammes en boîte)

Nous pouvons tout d'abord remarquer que les paramètres θ_{11} sont échantillonnés sur un intervalle de valeurs positives (Figure 5.2), alors que les paramètres θ_{12} et θ_{22} se trouvent majoritairement en dessous de 0. Les interactions $1 \leftrightarrow 1$ ont donc une propension d'occurrence plus importante que celles des interactions $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$. Ce qui corrobore les propos de STEHLÉ *et al.* [94] selon lesquels les enfants interagissent en premier lieu avec les camarades de leur classe. Inversement, les interactions impliquant des liens inter-classes (ou inter-organisationnels) ont une faible tendance d'apparaître dans les graphes.

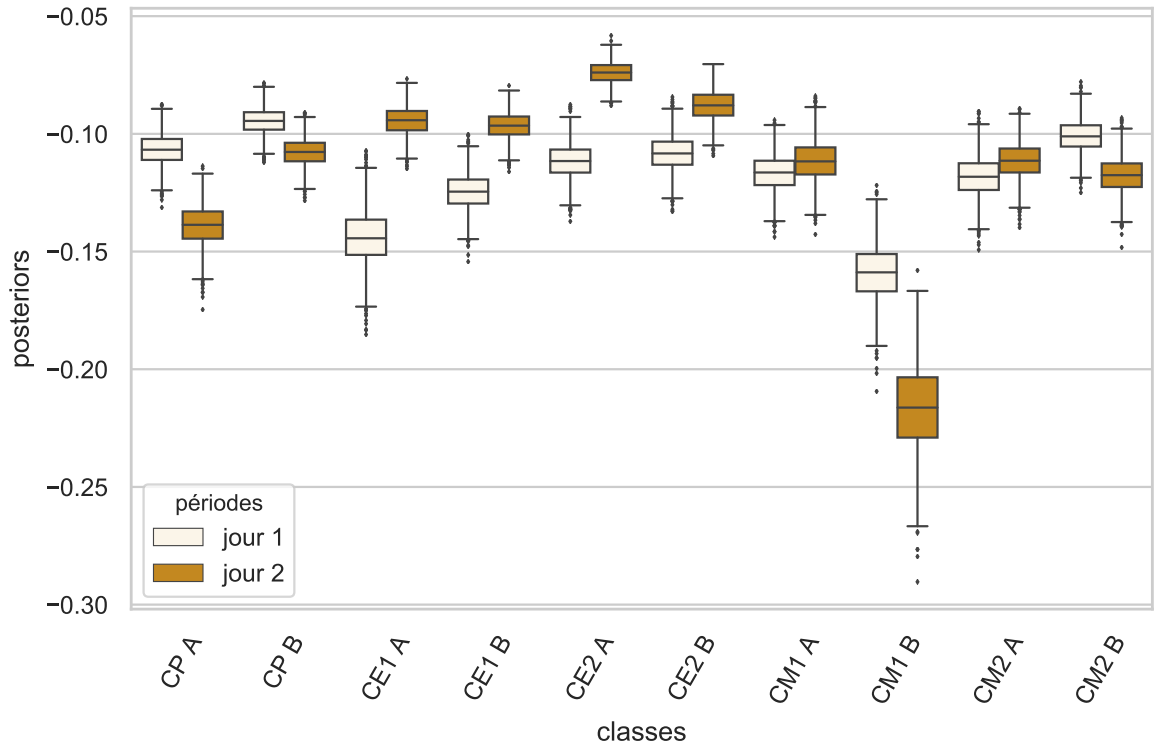


FIGURE 5.3 – Distributions *a posteriori* du paramètre θ_{12} pour chaque classe (représentées par des diagrammes en boîte)

TABLE 5.2 – Moyennes, médianes et estimateurs (MAP) calculés à partir des distributions échantillonnées par l'algorithme ABC Shadow

classe	période	$\overline{\theta_{11}}$	Q_{50} θ_{11}	MAP θ_{11}	$\overline{\theta_{12}}$	Q_{50} θ_{12}	MAP θ_{12}	$\overline{\theta_{22}}$	Q_{50} θ_{22}	MAP θ_{22}
CP A	jour 1	0.073	0.073	0.073	-0.107	-0.106	-0.106	-0.103	-0.102	-0.099
	jour 2	0.059	0.059	0.060	-0.139	-0.138	-0.137	-0.159	-0.158	-0.158
CP B	jour 1	0.063	0.063	0.063	-0.094	-0.094	-0.095	-0.032	-0.032	-0.031
	jour 2	0.061	0.061	0.060	-0.108	-0.108	-0.107	-0.116	-0.115	-0.115
CE1 A	jour 1	0.055	0.056	0.057	-0.144	-0.144	-0.145	-0.159	-0.158	-0.158
	jour 2	0.082	0.082	0.081	-0.094	-0.094	-0.093	-0.039	-0.040	-0.040
CE1 B	jour 1	0.051	0.051	0.050	-0.124	-0.124	-0.124	-0.155	-0.154	-0.154
	jour 2	0.053	0.053	0.052	-0.096	-0.096	-0.096	-0.075	-0.075	-0.074
CE2 A	jour 1	0.061	0.061	0.061	-0.111	-0.111	-0.111	-0.116	-0.115	-0.112
	jour 2	0.074	0.074	0.073	-0.074	-0.074	-0.075	-0.028	-0.027	-0.026
CE2 B	jour 1	0.074	0.075	0.075	-0.108	-0.108	-0.109	-0.089	-0.088	-0.085
	jour 2	0.108	0.108	0.110	-0.088	-0.088	-0.087	-0.011	-0.011	-0.010
CM1 A	jour 1	0.078	0.078	0.078	-0.116	-0.116	-0.115	-0.070	-0.069	-0.064
	jour 2	0.068	0.069	0.070	-0.111	-0.111	-0.111	-0.098	-0.097	-0.094
CM1 B	jour 1	0.051	0.052	0.052	-0.159	-0.159	-0.158	-0.102	-0.101	-0.102
	jour 2	0.043	0.044	0.044	-0.216	-0.216	-0.215	-0.152	-0.151	-0.149
CM2 A	jour 1	0.058	0.058	0.059	-0.118	-0.118	-0.118	-0.053	-0.052	-0.050
	jour 2	0.075	0.075	0.075	-0.111	-0.111	-0.109	-0.015	-0.014	-0.011
CM2 B	jour 1	0.063	0.064	0.064	-0.101	-0.101	-0.100	-0.093	-0.092	-0.089
	jour 2	0.059	0.059	0.059	-0.117	-0.117	-0.117	-0.108	-0.107	-0.107

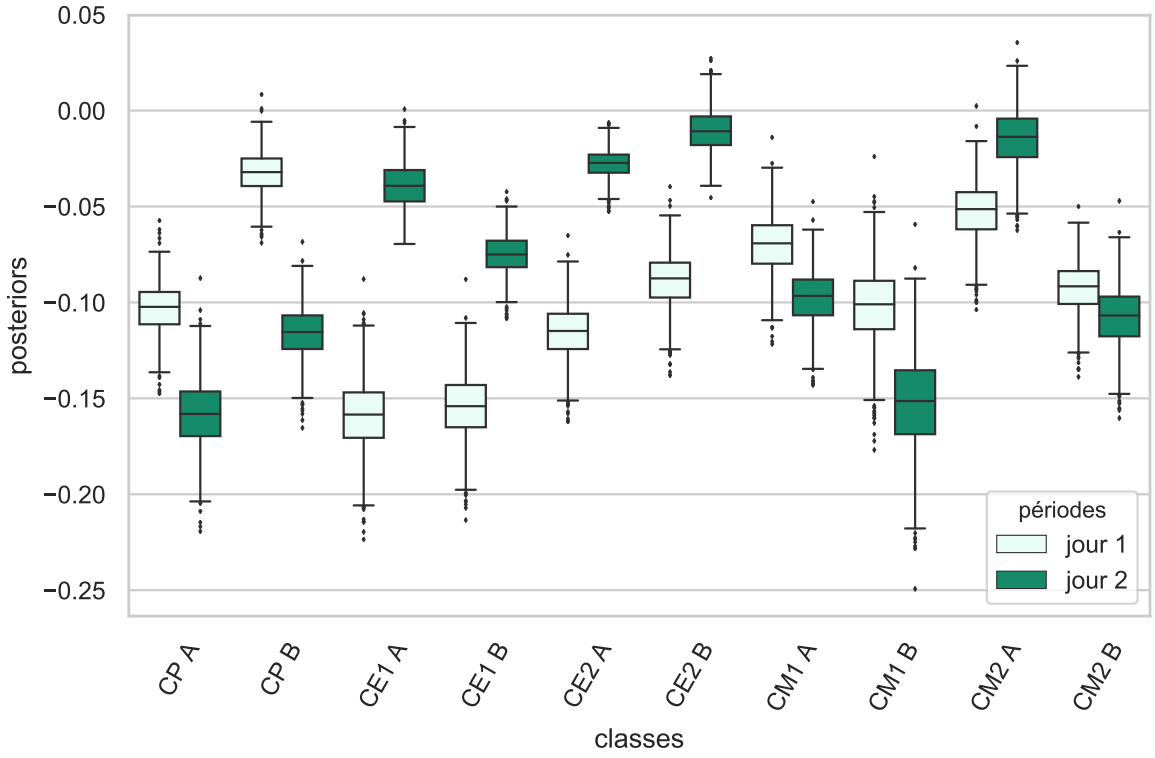


FIGURE 5.4 – Distributions *a posteriori* du paramètre θ_{22} pour chaque classe (représentées par des diagrammes en boîte)

Nous pouvons également observer l'évolution des distributions entre la première et la seconde journée. Pour les distributions concernant le paramètre θ_{11} (Figure 5.2), il n'y a pas visuellement de similitudes entre les distributions de la première journée et celles de la seconde. Mis à part certaines exceptions, la Figure 5.3 ne montre pas non plus de similitude entre les deux périodes. Enfin, la Figure 5.4 montre une variabilité encore plus importante entre les deux périodes. Ce qui peut s'expliquer par le fait qu'il y a un renouvellement important des liens inter-classes comme le soulignent les auteurs du papier original. Notons néanmoins que certaines classes de même niveau ont des distributions similaires pour une même période. C'est le cas par exemple des classes de CP A et CP B durant la première journée pour la distribution du paramètre θ_{12}. Cette observation se limite à quelques classes et ne permet pas de dire que les classes de même niveau partagent des caractéristiques communes. Pour compléter ces observations, nous avons réutilisé les mesures de distance de Kolmogorov-Smirnov que nous avons déjà mises en œuvre dans la section 4.4. Nous l'avons ici appliqué aux distributions échantillonnées pour chaque classe et sur les deux périodes. Les distances sont présentées en Annexe B sous la forme de cartes de chaleur par les Figures B.1, B.2 et B.3.

Ces observations sont limitées à deux journées. Elles ne sont pas suffisantes pour observer des récurrences dans l'apparition des motifs qui pourraient être extrapolés. D'autant que les interactions entre les enfants sont intrinsèquement liées à leur emploi du temps et à leur présence ou non le midi à la cantine.

En résumé, cette application nous permet d'illustrer l'approche dans un contexte totalement différent des collaborations entre chercheurs. Les structures organisationnelles se retrouvent dans de nombreux cas de figure et l'étude des interactions au sein de ces organisations et à leurs frontières se révèle pertinente pour de nombreuses questions. Dans ce cas précis, nous avons considéré les classes comme des organisations. Les interactions se déroulent soit entre les élèves de la même classe, soit avec des élèves d'autres classes. À travers le modèle d'interactions, nous avons cherché à observer des motifs récurrents entre les différentes classes et au cours du temps. Nous avons notamment observé pour toutes les classes une inclination positive pour un élève à interagir avec les autres élèves de sa classe. Pour les motifs impliquant des liens interclasses, la tendance est inverse. Nous n'avons pas observé de récurrence temporelle avec les données disponibles qui portent sur deux journées. Pour observer de telles récurrences, il serait nécessaire de travailler sur une échelle de temps plus importante. Enfin ici, nous avons fait abstraction du contexte épidémiologique. Néanmoins, les motifs que nous prenons en compte semblent totalement adaptés à la question. Chaque motif ($1 \leftrightarrow 1$, $1 \leftrightarrow 2$ et $2 \leftrightarrow 2$) représente un individu connecté à deux autres, jouant ainsi le rôle de relais. Ce relais peut être alors vecteur de la propagation d'un virus. Une perspective envisageable porterait sur l'application du modèle d'interactions à des questions en lien avec la propagation des virus.

5.3 Parallélisation de l'algorithme de simulation

Les algorithmes de simulations, tels que Metropolis-Hastings ou l'échantillonneur de Gibbs, jouent un rôle prépondérant dans les méthodes d'inférence. Dans le cas de l'algorithme ABC Shadow, elles permettent d'échantillonner la variable auxiliaire. Cet échantillonnage constitue l'étape la plus coûteuse en temps de l'algorithme. D'autant que le temps de calcul de cette opération dépend de la taille du graphe, et limite par conséquent la capacité de l'algorithme ABC Shadow à traiter des graphes de taille importante dans un temps raisonnable. Nous proposons dans cette section (sous-section 5.3.1) une stratégie de parallélisation de l'échantillonneur de Gibbs qui s'inspire des travaux de Bauer, Fürlinger et Kauermann [4]. L'objectif est donc de réduire le temps de calcul de cette étape pour des graphes de taille importante. Cette stratégie de parallélisation a fait l'objet d'une implantation. Dans la sous-section 5.3.2, nous présentons une étude empirique qui porte sur l'efficacité de cette implantation ainsi qu'une discussion sur les limites d'une telle stratégie de parallélisation.

5.3.1 Présentation de l'algorithme

Les procédures de simulation telles que le Metropolis-Hasitings ou l'échantillonneur de Gibbs sont utilisés dans le calcul du Maximum de vraisemblance (sous-section 2.3.2) et dans l'échantillonnage de la distribution *a posteriori* (ABC Shadow, EE algorithm) (section 3.4). Ces procédures interviennent par exemple dans l'algorithme ABC Shadow (Algorithme 6) et servent à échantillonner la variable auxiliaire. De manière générale, les algorithmes de simulation constituent l'étape la plus coûteuse en temps dans les méthodes d'inférence qui y font appel. Elles doivent réaliser un grand nombre d'itérations pour

évaluer l'ensemble des nœuds du graphe des arêtes (ou l'ensemble des arêtes du graphe social) et ce, à plusieurs reprises, de manière à converger. Ce nombre d'itérations est donc directement lié à la taille du graphe observé puisque le nombre d'arêtes équivaut à $n(n-1)/2$ avec n le nombre de nœuds du graphe social observé. Notre capacité à traiter des observations impliquant un grand nombre d'acteurs est de ce fait, fortement limitée. L'échantillonneur de Gibbs et le Metropolis-Hasitings partagent de nombreux points communs. Les deux algorithmes mettent à jour l'état du nœud courant en fonction des autres nœuds suivant les hypothèses de dépendance. Nous considérons ici les hypothèses de dépendance de Markov qui réduisent la dépendance des nœuds à leur voisinage (section 3.3). Cela signifie que les algorithmes de simulation prennent uniquement en compte le voisinage du nœud courant pour mettre à jour son état.

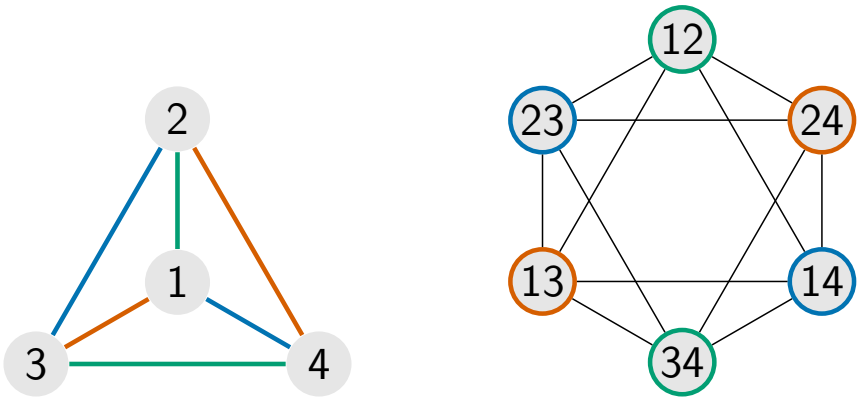
La différence entre l'algorithme Metropolis-Hastings et l'échantillonneur de Gibbs réside dans la façon de choisir le nouvel état d'un nœud. Le Metropolis-Hastings choisi aléatoirement (suivant une loi uniforme) un nouveau label et accepte ou rejette ce dernier selon la valeur du ratio d'acceptation $\alpha_{y \to y'}$ (Algorithme 1). L'échantillonneur de Gibbs détermine le nouveau label à partir de la probabilité $P(Y|y^c)$ (Algorithme 2) où y^c est la configuration actuelle. Contrairement au Metropolis-Hastings, l'échantillonneur de Gibbs accepte chaque nouveau label, ce qui revient à fixer $\alpha_{y \to y'} = 1$ et le rend donc plus efficace. Une description plus détaillée des deux algorithmes est donnée dans la sous-section 2.3.1. Une mise en œuvre de l'échantillonneur de Gibbs sur le graphe des arêtes est présentée dans la section 3.3.

Bauer, Fürlinger et Kauermann [4] proposent une version parallélisée de l'algorithme Metropolis-Hastings dédiée aux ERGMs. Ils utilisent le modèle de Markov pour des graphes non-dirigés tel que présenté dans l'état de l'art (sous-section 2.2.2). La configuration du modèle utilisé prend en compte les arêtes, les étoiles à deux branches et les triangles :

$$p(Y = y|\theta) = \frac{\exp(\theta_L L(y)) + \theta_{S_2} S_2(y) + \theta_T T(y))}{\kappa(\theta)}. \tag{5.1}$$

À la différence de l'approche que nous présentons, les auteurs considèrent directement les arêtes du graphe observé et non les nœuds du graphe des arêtes. Le vocabulaire est différent, mais les deux objets sont strictement équivalents : un nœud du graphe des arêtes représente une arête du graphe social et inversement. La Figure 5.5 illustre cette équivalence en présentant d'une part un graphe non-dirigé G (Figure 5.5(a)) et d'autre part la structure du graphe des arêtes correspondante (Figure 5.5(b)). Ainsi, chaque arête du graphe G est associée à un nœud du graph $L(G)$. Par ailleurs, la structure du graphe des arêtes est en tout point identique à celle du graphe de dépendance de Markov. Le graphe des arêtes permet de représenter directement les hypothèses de dépendance au travers du voisinage de ses nœuds. Pour faciliter la lecture, nous tâcherons autant que possible d'utiliser le terme de nœud qui désigne un nœud du graphe des arêtes (ou du graphe de dépendance).

L'idée qui sous-tend la parallélisation des algorithmes de simulation repose sur le fait que des nœuds indépendants peuvent être évalués et mis à jour simultanément. Autrement dit, l'évaluation de chacun de ces nœuds peut s'effectuer en parallèle sur des *threads* différents. Les dépendances de Markov nous permettent de paralléliser l'évaluation de



(a) Graphe non-dirigé G

(b) Graphe de dépendance de Markov ou graphe des arêtes $L(G)$ correspondant

FIGURE 5.5 – Structure de Markov. Les nœuds du graphe des arêtes $L(G)$ sont identifiés de la même manière que les arêtes du graphe G. Chaque couple de nœuds/arêtes indépendant(e)s conditionnellement au reste du graphe est associé à une couleur.

nœuds n'étant pas directement connectés les uns aux autres. Pour rappel, deux nœuds sont connectés si les arêtes correspondantes dans le graphe G sont reliées à un même nœud. Dans la Figure 5.5(b), les nœuds `(1,2)` et `(2,3)` sont voisins l'un de l'autre. Ils sont donc mutuellement dépendants et ne peuvent être évalués simultanément. Inversement, les nœuds `(1,2)` et `(3,4)` ne sont pas directement connectés et sont donc mutuellement indépendants conditionnellement au reste du graphe. Dans la Figure 5.5(b), chaque couple de nœuds indépendants est associé à une couleur.

Plus formellement deux nœuds représentés par les variables aléatoires $Y_{i,j}$ et $Y_{k,l}$ sont indépendants si les indices (représentant les nœuds aux extrémités des arêtes dans G) i, j, k et l sont distincts. Par abus de notation, nous considérerons dans la suite de ce chapitre que N correspond au nombre de nœuds du graphe social (le nombre d'individus). D'apparence, la stratégie de parallélisation semble simple, voire évidente. Elle nécessite néanmoins de déterminer des ensembles de nœuds indépendants. Ce problème est plus généralement connu sous le nom de problème du *k-ensemble* stable. Un stable est une autre manière de désigner un ensemble de nœuds indépendant. Ce problème, sous sa forme décisionnelle, revient à vérifier l'existence d'un ensemble de k nœuds indépendants. C'est un problème NP-complet[53], sa complexité est donc exponentielle. BAUER, FÜRLINGER et KAUERMANN [4] proposent une stratégie assez intuitive permettant de déterminer les ensembles de nœuds indépendants. Cette stratégie tire parti du fait que le graphe dans lequel les ensembles de nœuds indépendants sont recherchés est le graphe des arêtes. Les auteurs considèrent un graphe social G dont les nœuds sont numérotés de 1 à N avec N pair. Ils proposent alors de décomposer en tuples l'ensemble $\{1, \ldots, N\}$:

$$\mathcal{D}(N|2) = \{(1,2), (3,4), \ldots, (N-1, N)\}$$

De cette manière, chaque tuple qui compose $\mathcal{D}(N|2)$ contient des indices distincts. Par conséquent, tous les nœuds $Y_{i,j}, (i,j) \in \mathcal{D}(N|2)$ sont mutuellement indépendants

par rapport au reste du graphe. L'ensemble de tuples $\mathcal{D}(N|2)$ peut être construit en mélangeant préalablement l'ordre des indices $\{1, \ldots, N\}$. Tous les $Y_{i,j}, (i,j) \in \mathcal{D}(N|2)$ peuvent être évalués en parallèle.

Si nous reprenons l'exemple illustré par la Figure 5.5, l'algorithme de simulation peut s'exécuter de la manière suivante :

Itération 1 Les nœuds $Y_{1,2}$ et $Y_{3,4}$ (en vert) sont évalués en parallèle

Itération 2 Les nœuds $Y_{1,3}$ et $Y_{2,4}$ (en orange) sont évalués en parallèle

Itération 3 Les nœuds $Y_{2,3}$ et $Y_{1,4}$ (en bleu) sont évalués en parallèle

L'ordre des itérations importe peu. Il est seulement nécessaire que les nœuds évalués en parallèle soient mutuellement indépendants par rapport au reste du graphe. L'implantation proposée par les auteurs repose sur l'algorithme Metropolis-Hastings. L'échantillonneur de Gibbs n'est qu'un cas particulier du Metropolis-Hastings. Comme évoqué en préambule de cette section, la différence entre ces deux algorithmes réside dans la manière de proposer un nouvel état. L'échantillonneur de Gibbs parallélisé que nous proposons est donc proche de l'implantation de Bauer, Fürlinger et Kauermann [4] et reprend les mêmes concepts que nous venons de développer.

Une description de l'échantillonneur de Gibbs parallélisé est donnée par l'Algorithme 8. La première étape de cet algorithme (ligne 2) consiste à calculer les indices des nœuds indépendants et à les placer dans des conteneurs que l'on nomme ici *chunks*. Les *chunks* correspondent aux ensembles de tuples $\mathcal{D}(N|2)$ précédemment mentionnés. Il est nécessaire de générer autant de *chunks* que d'itérations *sim_iter*. Les *chunks* peuvent intuitivement être représentés sous la forme d'un tableau à deux dimensions où $chunks[i]$ permet d'accéder au ième *chunk* et $chunks[i][n]$ au nième couple d'indices du ième *chunk*. La Figure 5.6 illustre cette organisation pour l'exemple représenté par la Figure 5.5. Les *chunks* ont tous la même taille que l'on associe à la variable *chunk_size* (ligne 3). Un chunk contient les indices d'un nombre réduit de nœuds : $N/2$ [12]. Cela signifie qu'une étape de la boucle principale (lignes 5 à 12 de l'Algorithme 8), ne permettra pas d'évaluer l'ensemble des nœuds du graphe des arêtes. Il convient d'augmenter le nombre d'itérations *sim_iter* pour évaluer autant de nœuds que dans la version en série de l'échantillonneur (Algorithme 2). Dans l'exemple présenté par la Figure 5.6, nous avons représenté trois *chunks* qui nous permettent de couvrir tous les nœuds du graphe des arêtes. Dans ce cas précis, la version parallélisée de l'échantillonneur de Gibbs doit réaliser trois itérations pour évaluer autant de nœuds qu'une itération de la version en série. À chaque itération, l'échantillonneur parallélisé peut évaluer et mettre à jour indépendamment tous les nœuds contenu dans un *chunk*. En l'occurrence, dans l'exemple de la Figure 5.6, l'Algorithme 8 peut évaluer et mettre à jour en parallèle seulement deux nœuds.

5.3.2 Résultats empiriques

L'algorithme 8 a fait l'objet d'une implantation en C++. Nous utilisons la bibliothèque OpenMP (Open Multi-Processing) [76] pour assurer la parallélisation des tâches. Cette

12. en faisant l'hypothèse que N est pair

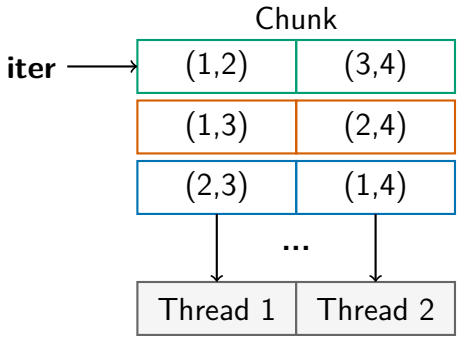


FIGURE 5.6 – Organisation des nœuds indépendants en *chunks*

Algorithme 8 Échantillonneur de Gibbs parallélisé

function PARALLELGIBBSSAMPLER(y_0, sim_iter)
  $chunks \leftarrow getIndependentNodes(y, sim\_iter)$ ▷ Calcul des nœuds indépendants
  $chunk\_size \leftarrow chunks.chunk\_size$ ▷ Taille d'un chunk
  $y \leftarrow y_0$
  for $iter = 1 \dots sim\_iter$ **do**
    parfor $n = 1 \dots chunk\_size$ **do** ▷ Évaluation en parallèle de chaque nœud
      $(i, j) \leftarrow chunks[iter][n]$ ▷ Récupération de l'indice du nœud
      $y'_{i,j} \sim P(\cdot|y_{V_{i,j}})$ ▷ Proposition d'un nouveau label
      $y_{i,j} \leftarrow y'_{i,j}$ ▷ Mise à jour du nœud
    end parfor ▷ Barrière
    $y^{(iter)} \leftarrow y$ ▷ Sauvegarde du graphe
  end for
  return $\{y^{(i)}\}_{i=1\dots sim\_iter}$
end function

bibliothèque à l'avantage de s'intégrer facilement au code source existant, sans l'altérer. Elle simplifie la gestion des *threads* et de la mémoire partagée à l'aide de directives `pragma`.

L'implantation reprend le formalisme des graphes tel que ne l'avons présenté dans la section 3.2. Nous considérons un graphe social qui représente une organisation et les individus extérieurs qui interagissent avec les membres de cette organisation. C'est une représentation que nous avons déjà eu l'occasion d'utiliser auparavant. Dans l'application sur les collaborations entre les chercheurs, nous considérions une équipe de chercheurs et les collaborateurs extérieurs qui avaient corédigé une publication scientifique avec au moins un membre de cette équipe. L'étude des interactions entre les élèves d'une école primaire, nous a conduits à considérer chaque classe et à observer les intéractions que chacune d'elles pouvait avoir avec les autres élèves. Nous définissons la taille du graphe par le tuple `(inSize, outSize)` :

`inSize` : Le nombre d'individus appartenant à l'organisation
`outSize` : Le nombre d'individus extérieurs à l'organisation

Ainsi, un graphe social de taille `(1000, 600)` représente une organisation qui compte 1000 membres et qui interagit potentiellement avec 600 individus extérieurs. Nous

réutilisons également le même conditionnement que celui utilisé dans les applications précédentes. Par conséquent, nous ignorons toutes les arêtes qui relient exclusivement des individus n'appartenant pas à l'organisation. Enfin, l'implantation s'appuie sur le graphe des arêtes. Nous distinguons dans ce graphe deux types de nœuds. Les nœuds `intra` représentent les interactions entre les membres de l'organisation, les liens intra-organisationnels. A contrario, les nœuds `inter` représentent les liens inter-organisationnels, c'est-à-dire, les liens qui connectent un membre de l'organisation et un individu extérieur. La génération des *chunks* est réalisée de telle sorte que la proportion entre les nœuds *inter* et les nœuds *intra* soit conservée. Le paramètre *sim_iter* est adapté de manière à ce que la version parallélisée de l'échantillonneur de Gibbs évalue autant de nœuds que la version en série. De cette manière, nous limitons les biais dus à la génération des nœuds indépendants située à la ligne 2 de l'Algorithme 8.

Afin d'évaluer les performances de la stratégie de parallélisation de l'Algorithme 8, nous avons mesuré le temps d'éxécution de la simulation sur plusieurs configurations de *threads* et pour différentes tailles de graphe. La taille des graphes simulés varie de `(1000, 1000)` à `(5000, 5000)`. Nous avons souhaité comparer les performances de la version de l'échantillonneur de Gibbs en série et celles de la version parallèle pour différents nombres de *threads* : $2, 4, 6, 8, 10, 12, 14$ et 16. Les exécutions ont été réalisées sur un serveur équipé de deux sockets avec pour chacun, un CPU Intel(R) Xeon(R) E5-2640 v4 d'une fréquence de 2,40 GHz (3,40 GHz en turbo). Chaque CPU est équipé de 10 cœurs et 20 *threads* et dispose d'un cache de 25 Mo. L'espace mémoire RAM total est de 110 Go. Pour chaque taille de graphe et nombre de *threads*, la simulation a exécuté l'équivalent de $sim\_iter = 100$ itérations d'un échantillonneur en série. Chaque simulation a été répétée 50 fois afin d'obtenir un nombre suffisant de mesures.

La Figure 5.7 synthétise les temps mesurés pour chaque exécution. Cette durée ne prend pas en compte le temps de génération des nœuds indépendant (ligne 2 de l'Algorithme 8) et ce, pour deux raisons. D'une part, le temps de calcul des nœuds indépendants est constant pour une taille de graphe et un nombre d'itérations donnés. Ce temps de calcul n'affecte pas la comparaison entre différentes configurations de *threads*. D'autre part, la génération des indices des nœuds indépendants peut-être réalisée une seule fois et non à chaque appel de l'échantillonneur. Ce qui génère un coût marginal de l'ordre d'en moyenne 0.5 secondes pour un graphe de taille `(1000,1000)` et 11.60 secondes pour un graphe de taille `(5000,5000)`, et $sim\_iter = 100$. Dans la Figure 5.7, les échelles de temps différent en fonction de la taille du graphe. Pour faciliter la lecture et constater plus aisément les effets de la parallélisation, nous avons représenté les accélérations pour chaque taille de graphe et nombre de *threads* dans la Figure 5.8. Le calcul de l'accélération s'appuie sur la durée moyenne d'exécution de chaque configuration (*i.e.* la taille du graphe et le nombre de *threads*). L'accélération s'exprime comme le rapport du temps de calcul de l'algorithme en série sur le temps de calcul de l'algorithme parallélisé : T_{ser}/T_{par} . Par exemple, une accélération de 2 signifie que l'algorithme parallélisé est deux fois plus "rapide" que la version en série.

Nous remarquons à la lecture de la Figure 5.7 que pour des tailles de graphe supérieures ou égales à `(2000,2000)`, le temps de calcul de la version parallélisée sur deux *threads* est plus important que les temps de calcul de la version en série. Ce qui signifie que la version parallélisme de l'algorithme induit un surcoût. Ce phénomène n'est pas observé pour des

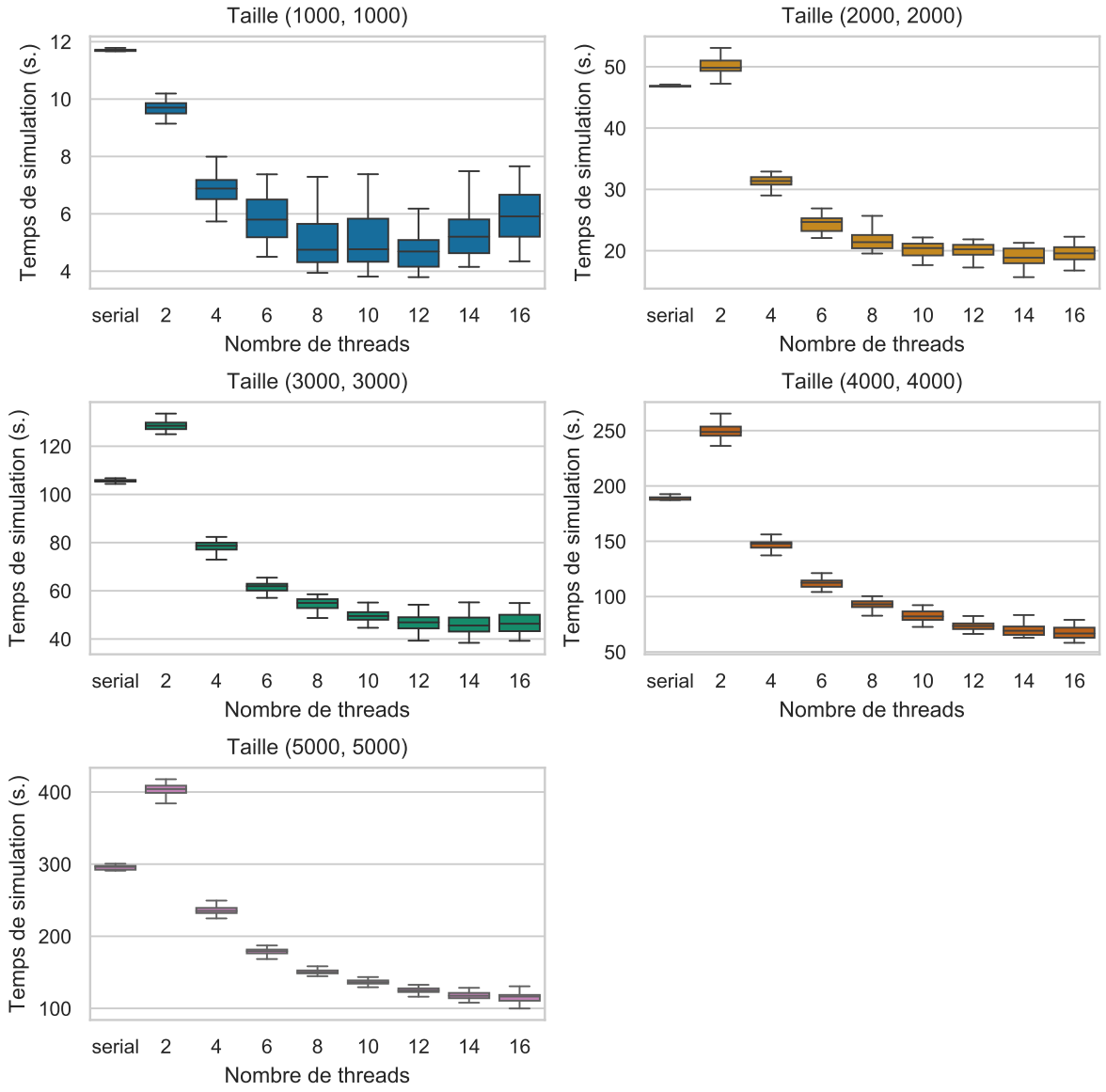


FIGURE 5.7 – Temps d'éxecution de la simulation en fonction de la taille du graphe simulé et du nombre de *threads*. Calcul effectué à partir des temps d'exécution moyens

graphes de taille plus modeste, comme `(1000,1000)`. Il y a donc un lien avec le nombre de nœuds à traiter. Nous pouvons faire plusieurs hypothèses quant à l'origine de ce surcoût. La principale différence entre l'implantation en série et celle en parallèle réside dans la manière de stocker et récupérer l'indice des nœuds. Dans le cas de la version en série, l'algorithme itère simplement sur la liste des nœuds du graphe des arêtes, et ce, autant de fois qu'il y a d'itérations (*sim_iter*) à réaliser. Dans le cas de la version parallélisée, tous les indices de nœuds sont préalablement calculés et stockés dans un tableau de *chunks* (Figure 5.6) et ce, pour toutes les itérations. Même si l'accès à une valeur d'un tableau se fait en temps constant, nous pouvons supposer qu'il y a un coût dû à la mise en cache des valeurs. Nous pouvons également supposer que certains mécanismes de synchronisation

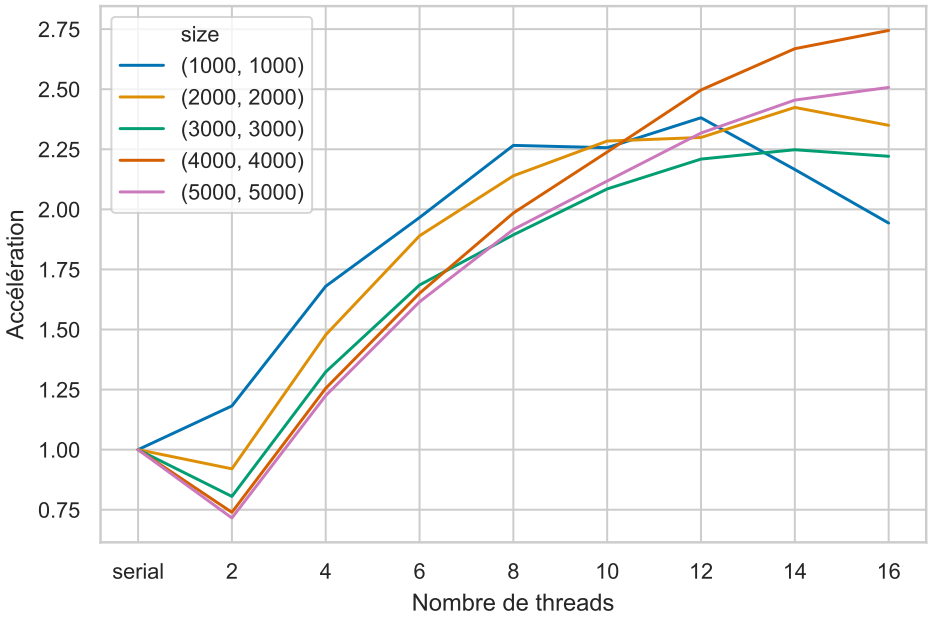


FIGURE 5.8 – Mesure du gain des différentes exécutions en terme d'accélération

d'OpenMP induisent également un surcoût. Malgré ces surcoûts, un gain en performance est observé à partir de 4 *threads* sur toutes les configurations. Ce qui s'illustre par des accélérations supérieures à 1 dans la Figure 5.8.

Le gain en performances est somme toute relativement faible, les accélérations mesurées ne dépassent pas 2.75, avec 16 *threads*. Nous avons identifié plusieurs raisons à cela. Tout d'abord, la tâche parallélisée (de la ligne 6 à 10 dans l'Algorithme 8) s'exécute sur une très petite durée, de l'ordre de 10^{-7} secondes. Dans les travaux de BAUER, FÜRLINGER et KAUERMANN [4], le calcul à paralléliser est plus coûteux, puisqu'il nécessite de compter les configurations en triangle. Ce qui explique la différence entre les résultats obtenus par les auteurs et les résultats que nous présentons.

On remarque par ailleurs que l'Algorithme 8 reste malgré tout très séquentiel. Les *threads* de la boucle `for` parallélisée doivent nécessairement s'attendre à la fin de cette dernière (ligne 10) avant de passer au *chunk* suivant. Les temps de calcul de la section parallélisée se rapporte donc au temps d'exécution du *thread* le plus lent. D'autre part, si le nombre de tâches à paralléliser (*i.e.* la taille du chunk) n'est pas suffisamment élevé, nous sommes confrontés à une situation où les temps induits par OpenMP (lancement des *threads*, synchronisation, accès aux données partagées ...) sont plus importants que le temps de calcul lui-même. Dans ce cas de figure, la parallélisation n'apporte aucun bénéfice. La taille des *chunks* dépend de la taille du graphe considérée, dans notre cas la taille des *chunks* varie de 500 à 2500 nœuds pour des tailles de graphe allant respectivement de `(1000, 1000)` à `(5000, 5000)`. Pour des graphes de taille `(1000,1000)`, nous observons par exemple qu'à partir de 8 *threads*, le temps de calcul cesse de diminuer. Dans ce cas-là, le travail parallélisé sur chaque *thread* n'est pas suffisant pour contrebalancer le coût inhérent à la parallélisation. Nous serions alors tenté de penser que plus la taille des

graphes augmente, plus la taille des *chunks* augmente également et donc, plus le travail parallélisé sur chaque *thread* est important. Ce qui devrait par conséquent, améliorer l'efficacité de l'algorithme. Seulement, à mesure que la taille des graphes augmente, le nombre de chunks augmente également, si nous voulons traiter autant de nœuds que dans la version en série. Le nombre de *chunks* dépend intrinsèquement de la taille du graphe. Il y a environ 5 fois plus de *chunks* générés pour des graphes de taille `(5000, 5000)` que pour des graphes de taille `(1000,1000)`. Nous restons donc confrontés aux limites liées au caractère séquentiel de l'Algorithme 8.

Nous faisons face aux limites de la stratégie de parallélisation. La très faible durée de la tâche parallélisée et la nécessité pour les *threads* de s'attendre à la fin de chaque itération de la boucle principale limitent l'accélération de l'Algorithme 8. Nous pouvons néanmoins mitiger ce constat, notamment, dans le cas d'ABC Shadow et de l'échantillonnage de la variable auxiliaire. En effet, il n'est pas nécessaire de réaliser un grand nombre d'itérations pour échantillonner la variable auxiliaire. Un très petit nombre d'itérations suffit pour fournir suffisamment d'informations pour le calcul du ratio d'acceptation. Il est donc possible avec ABC Shadow de traiter des graphes de taille comparable à celles que nous venons d'utiliser dans cette section, et cela, dans un temps raisonnable.

Par ailleurs, Bauer, Fürlinger et Kauermann [4] soulignent l'existence d'autres stratégies de simulation et d'inférence permettant de traiter des graphes de taille importante. Une de ces stratégies consiste à utiliser une observation partielle, échantillonnée et non le graphe dans sa globalité. Il est alors possible de traiter l'observation échantillonnée en considérant que le reste du graphe est conditionnellement indépendant. L'approche proposée par Pattison *et al.* [78] permet de traiter des graphes dont la taille est importante, voire, inconnue. Rappelons également que Snijders [89] suggère différentes stratégies visant à améliorer la vitesse de convergence de la chaîne résultante.

De manière générale, le problème du passage à l'échelle n'est pas une limitation rédhibitoire qui nous empêcherait définitivement de traiter des graphes de taille importante. Notre approche nous permet facilement de changer d'échelle et de considérer non plus des interactions entre des individus, mais entre des groupes d'individus (organisations) eux-mêmes imbriqués dans une autre échelle organisationnelle ("super groupe").

Chapitre 6

Conclusions et perspectives

La compréhension des réseaux sociaux (au sens graphe des interactions sociales) constitue un enjeu dans de nombreux domaines : études épidémiologiques, lutte contre la propagation d'infox, études sociales pour le développement économique ... Ce travail de thèse a été motivé par la volonté de comprendre les interactions entre les membres de différentes organisations impliqués dans des collaborations. L'objectif *in fine*, est de proposer une topologie pair-à-pair permettant de redonner à l'utilisateur le contrôle sur les données et les échanges tout en assurant techniquement la disponibilité des données et des liaisons pair-à-pair. Cette notion de contrôle est primordiale pour préserver le caractère sensible des données. D'autant que cette question autour du contrôle des données se pose aussi à l'échelle des organisations dans le cadre des collaborations inter-organisationnelles. Le réseau social qui se tisse entre les collaborateurs au fil des rencontres et des échanges dans le monde réel représente un maillage de confiance que nous souhaitons exploiter pour assurer ce contrôle. Nous considérons que le contrôle tant individuel que collectif s'illustre à travers ce réseau. Le manque de connaissance sur les propriétés de ce maillage nous conduit à mener une étude sur ses caractéristiques morphologiques en développant une approche de modélisation et d'inférence.

6.1 Contributions

Nous avons proposé en premier lieu une approche de modélisation des réseaux sociaux impliquant une structure organisationnelle. L'objet d'étude est un graphe social. Nous nous intéressons à la distribution des arêtes dans ce graphe et considérons de cette manière : le graphe des arêtes. C'est une représentation duale du graphe social observé qui met en avant les interactions entre les arêtes. La dimension organisationnelle est représentée par la nature des liens. Nous distinguons ainsi deux types de lien : les liens inter-organisationnels et les liens intra-organisationnels, représentant respectivement des interactions entre deux membres d'organisations différentes et des interactions au sein de la même organisation. L'émergence d'une arête dans le graphe social entraîne l'activation du nœud correspondant dans le graphe des arêtes. Cela revient à considérer le processus social de création d'une arête comme un processus de coloration de graphe. À partir de cette représentation spatiale, nous proposons un modèle stochastique. Ce

modèle est assimilé aux ERGMs et s'inspire du modèle de Potts, utilisé notamment en traitement numérique de l'image. Ce modèle d'interactions comptabilise les connexions entre les différents types de liens (*i.e.* nœuds du graphe des arêtes). Les motifs sont donc élémentaires, mais permettent d'évaluer la propension pour un individu à collaborer avec deux autres individus sous différentes configurations organisationnelles (au sein de la même organisation, entre plusieurs organisations ...). C'est un motif générique où un individu joue le rôle de relais entre deux autres individus, ce qui a du sens dans le cas des réseaux pair-à-pairs et celui des collaborations. L'utilisateur est un relais informationnel dans ce cas de figure. Il est d'autant plus intéressant d'étudier ce motif élémentaire, qu'il peut être repris dans d'autres contextes comme celui de la propagation épidémiologique.

La forme générale du modèle implique une constante de normalisation qui ne permet pas d'estimer directement ses paramètres. Nous proposons d'utiliser un algorithme d'inférence bayésienne pour contourner ce problème. De la même manière que nous nous sommes inspirés de la modélisation de processus spatiaux pour l'élaboration du modèle, nous utilisons un algorithme d'inférence, ABC Shadow, déjà démontré sur l'estimation de modèles spatiaux (modèle de Strauss et Candy). L'algorithme ABC Shadow approxime la distribution *a posteriori* du model, étant donnée une observation. Il s'inspire de deux approches, des méthodes *Approximate Bayesian Computation* (ABC) d'une part et l'approche de Møller *et al.* [67] d'autre part. Comparé aux méthodes ABC, il permet un meilleur contrôle sur l'approximation et tend à proposer des échantillons toujours plus proches de la vraie distribution. Par ailleurs, l'échantillonnage de la variable auxiliaire suggérée par Møller *et al.* [67], ne nécessite pas d'utiliser une méthode de simulation parfaite [47, 81]. En outre, l'inférence bayésienne nous permet d'exploiter la distribution *a posteriori*. À partir de cette distribution, nous pouvons mesurer l'incertitude autour du MAP estimé. La forme de la distribution est également une source d'information. Nous pouvons comparer ces distributions pour identifier des similitudes morphologiques entre différents graphes.

Nous avons démontré l'approche sur un exemple réel de collaboration inter-organisationnelle : les collaborations initiées par les chercheurs d'un laboratoire. Les publications scientifiques sont très souvent corédigées par plusieurs chercheurs. Elles sont le fruit d'une collaboration qui peut dépasser les frontières des équipes de recherche et des laboratoires. Elles sont en cela un très bon exemple de collaboration impliquant des membres de différentes organisations. Nous nous sommes appuyés sur les données bibliographiques publiées par l'archive ouverte HAL pour extraire les liens de coauteurs entre les chercheurs. Ainsi, nous pouvons tracer un graphe social des chercheurs ayant copublié une ou plusieurs productions scientifiques. En étudiant les graphes des coauteurs de chaque équipe, nous avons remarqué qu'il y avait une faible tendance générale à tisser des liens de collaboration avec d'autres laboratoires. Ce qui s'illustre dans le graphe social par un très petit nombre d'acteurs maintenant à la fois des liens intra-organisationnels et des liens inter-organisationnels. Si ces acteurs venaient à partir, les liens inter-organisationnels pourraient être rompus du point de vue de l'équipe de recherche. Ce qui mettrait un terme à ces collaborations. Replacé dans le contexte des réseaux pai-à-pair, ce résultat suggère que le réseau social qui sous-tend la collaboration est très sensible aux départs des pairs. Différentes stratégie de réplication se basant sur des attributs [40, 88] sociaux pourraient être adaptées pour assurer une meilleure disponibilité

des collaborations.

L'approche proposée peut-être appliquée à d'autres exemples d'interactions sociales, ne se rapportant pas nécessairement aux collaborations. Nous prenons comme second exemple d'application les interactions sociales entre des enfants d'une école primaire enregistrées durant deux journées. Les classes représentent la structure organisationnelle. L'approche nous permet donc d'étudier la tendance pour les enfants à interagir au sein de leur classe ou avec des enfants d'autres classes. Nous avons observé pour toutes les classes une tendance positive à interagir avec les enfants de la même classe. Inversement, nous avons remarqué que les motifs impliquant des liens inter-organisationnels ont une faible propension d'occurrence. En somme, les enfants interagissent préférentiellement avec les enfants de leur classe. Ce résultat peut avoir des implications dans la gestion d'une crise sanitaire, comme le soulignent Stehlé *et al.* [94]. Par exemple, si plusieurs cas d'une maladie sont recensés dans une classe, la fermeture de la classe peut suffire à endiguer la propagation de l'épidémie.

Enfin, nous proposons une stratégie de parallélisation de l'échantillonneur de Gibbs visant à réduire le temps de simulation de la variable auxiliaire dans l'algorithme ABC Shadow. Le temps de calcul lié à la simulation de la variable auxiliaire dépend intrinsèquement de la taille du graphe. La stratégie de parallélisation a initialement été proposée par Bauer, Fürlinger et Kauermann [4] pour l'algorithme Metropolis-Hastings. Cette stratégie tire parti des hypothèses de dépendances de Markov. Elle peut se résumer ainsi : au lieu d'évaluer séquentiellement chaque nœud du graphe, nous pouvons évaluer en parallèle les nœuds indépendants suivant les hypothèses de Markov. Nous proposons une évaluation empirique des performances de la version parallélisme de l'échantillonneur de Gibbs. Cette stratégie offre des gains en performance limités qui permettent seulement de diviser le temps de calcul par 2. Dans notre cas, le temps de calcul des tâches parallélisées est très petit et devient rapidement négligeable comparé au temps de synchronisation des *threads*. L'algorithme reste malgré tout très séquentiel, contraignant les *threads* à s'attendre très régulièrement. Pour autant, il ne faut pas y voir une limite infranchissable réduisant notre capacité à traiter des graphes de taille importante. Le modèle que nous proposons permet de changer facilement d'échelle et d'agréger les observations, diminuant ainsi la taille du graphe à traiter.

6.2 Discussions et perspectives

Au terme de ce travail, nous avons relevé plusieurs limites inhérentes à notre approche et identifié des perspectives que nous souhaiterions développer dans de futurs travaux. La première limite que nous avons identifiée concerne le modèle. Ce dernier repose sur les hypothèses de dépendances de Markov. Il est donc sujet à des problèmes de dégénérescence comme nous avons pu l'évoquer dans la sous-section 2.2.2 et le constater dans la section 4.3. De plus, il est relativement sensible, une faible modification des paramètres peut engendrer une importante variation des statistiques suffisantes. Le modèle n'est donc pas en mesure de reproduire toutes les observations avec précision. Un futur travail est donc nécessaire sur une redéfinition du modèle permettant de représenter un plus grand nombre d'observations et moins assujetti aux problèmes de stabilité. Cette redéfinition

devra aussi permettre d'améliorer l'expressivité du modèle, permettant par exemple de prendre en compte différents degrés de centralisation. Un rapprochement avec le modèle proposé par Snijders *et al.* [92] fait très clairement l'objet d'une perspective à étudier. Il faut aussi garder à l'esprit notre intention de départ qui est d'entretenir une proximité avec les structures pair-à-pair. Toujours dans l'optique de s'inspirer des méthodes qui ont trait à la modélisation de processus spatiaux, nous avons identifié une piste de recherche portant sur les *Markov connected component fields* [68]. Il serait intéressant de confronter les composantes connectées de ce formalisme avec la structure organisationnelle et d'étudier les interactions aux frontières des organisations au travers de ce prisme.

Enfin, il nous semble important d'étudier la dynamique des réseaux sociaux. Jusqu'à présent nous avons considéré les graphes observés comme des instantanés. Nous avons remarqué que le maillage social qui sous-tend les collaborations entre les chercheurs était faiblement dense et que les proportions d'occurrences des motifs impliquant des liens inter-organisationnels étaient elles aussi faibles. Les connexions entre différentes organisations sont donc maintenues par un très petit nombre d'individus. Il est à présent nécessaire de comprendre la manière dont ces liens inter-organisationnels évoluent au cours du temps. Est-ce que ces liens sont uniquement maintenus par des membres permanents de l'équipe ou bien aussi par des doctorants ou des postdocs ? Auquel cas, est-ce que les liens inter-organisationnels sont toujours maintenus une fois que la personne a quitté l'équipe ? Comme nous avons pu le constater dans la deuxième application (section 5.2), il est nécessaire de disposer d'un nombre suffisant d'observations longitudinales. La collecte des données est en soi une difficulté à laquelle nous nous sommes déjà heurtés lors de la création du jeu de données sur HAL. Par ailleurs, il est nécessaire de définir le pas de temps que représente chaque observation.

La thèse portait initialement sur l'intégration de mécanismes de fédération dans les systèmes collaboratifs distribués. Nous avons progressivement considéré les interactions sociales comme éléments structurant ces mécanismes. Ce qui nous a amenés à étudier ces interactions à travers l'approche stochastique que nous présentons ici. Cependant, nous n'avons pas pris la juste mesure du temps nécessaire à la réalisation de ce travail. Ce qui devait être un travail préliminaire permettant de justifier notre approche est devenu l'essentiel du travail de thèse, et ce, pour plusieurs raisons. Il a tout d'abord fallu du temps pour s'approprier les outils et concepts que nous développons ici. La collecte d'un jeu de données pertinent illustrant une collaboration et intégrant une structure organisationnelle a également fait l'objet de difficultés que nous avions sous-estimées. Malgré le tournant inattendu qu'a pris cette thèse, nous avons pu élaborer une approche et la mettre en œuvre sur des cas d'étude réels [60].

Annexe A

Description du modèle de Potts : influence des paramètres

L'échantillonneur de Gibbs a été exécuté sur 500 itérations pour chaque configuration du modèle. Nous représentons l'évolution des statistiques suffisantes des échantillons au cours de la simulation.

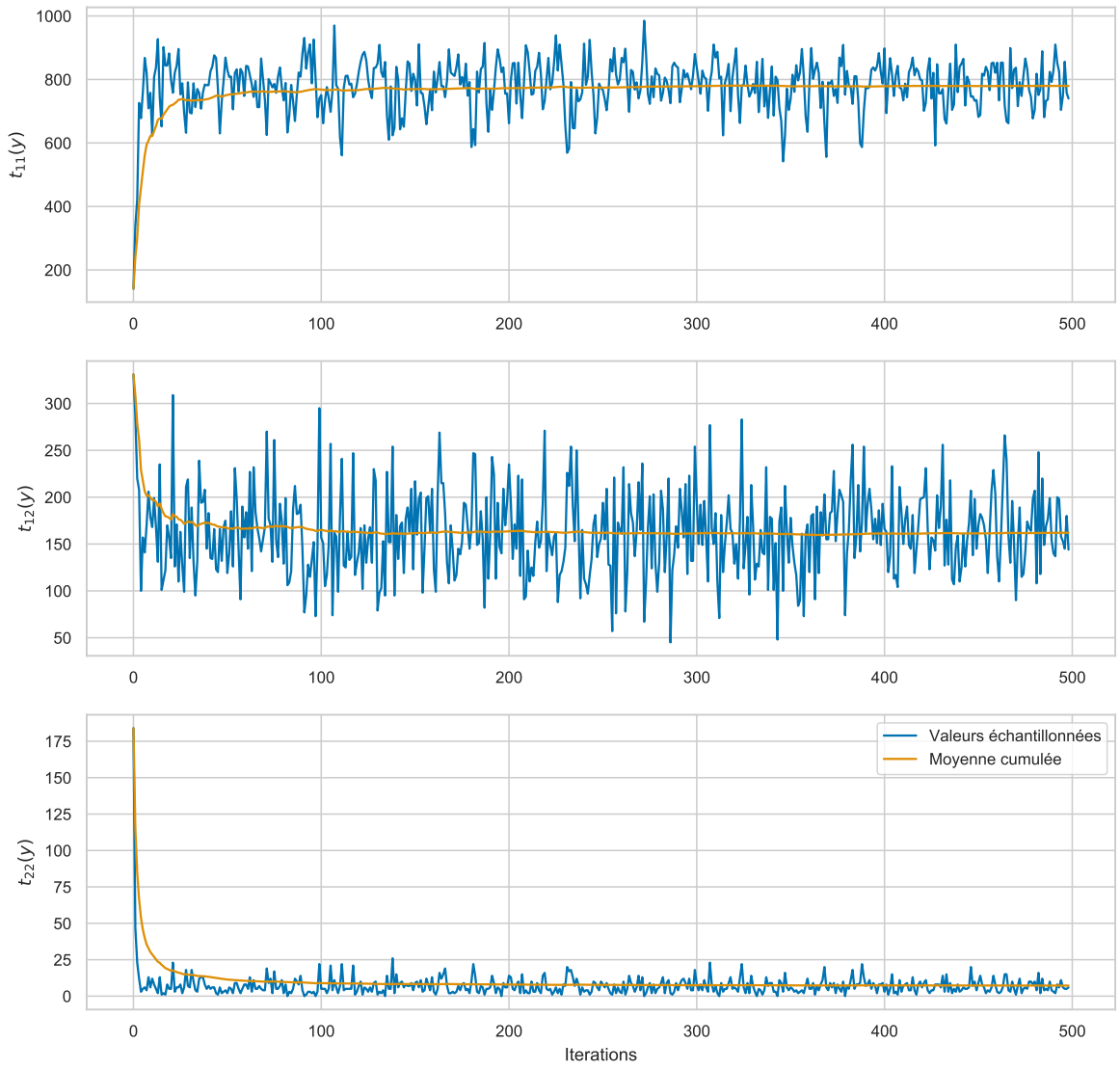


FIGURE A.1 – Statistiques suffisantes des échantillons simulés par l'échantillonneur de Gibbs pour le modèle paramétré avec $\theta = [0.15, -0.2, -0.2]$. La valeur moyenne des statistiques suffisantes est $\bar{t}(y) = [780, 162, 7]$.

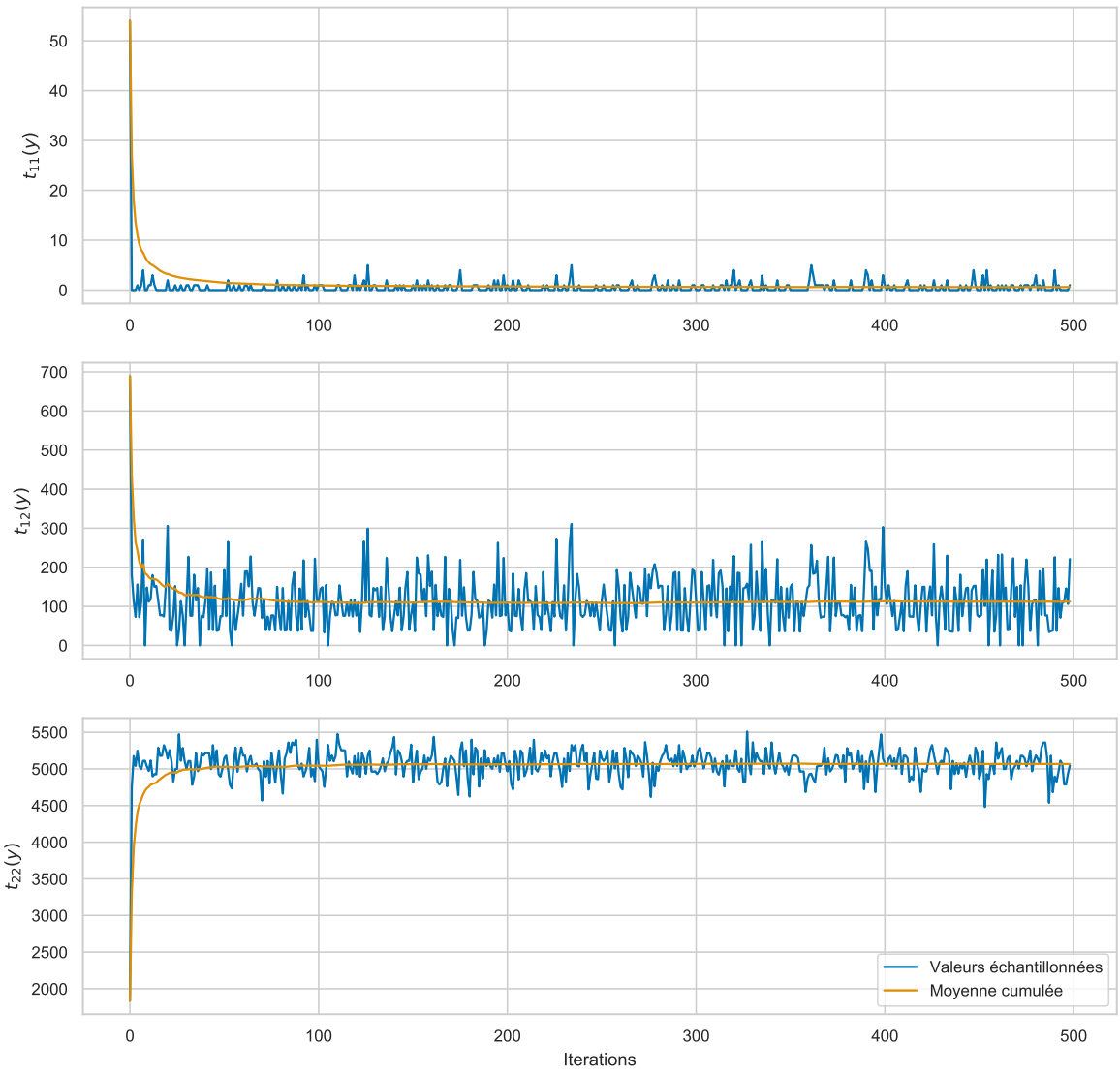


FIGURE A.2 – Statistiques suffisantes des échantillons simulés par l'échantillonneur de Gibbs pour le modèle paramétré avec $\theta = [-0.1, -0.1, 0.08]$. La valeur moyenne des statistiques suffisantes est $\bar{t}(y) = [1, 112, 5066]$.

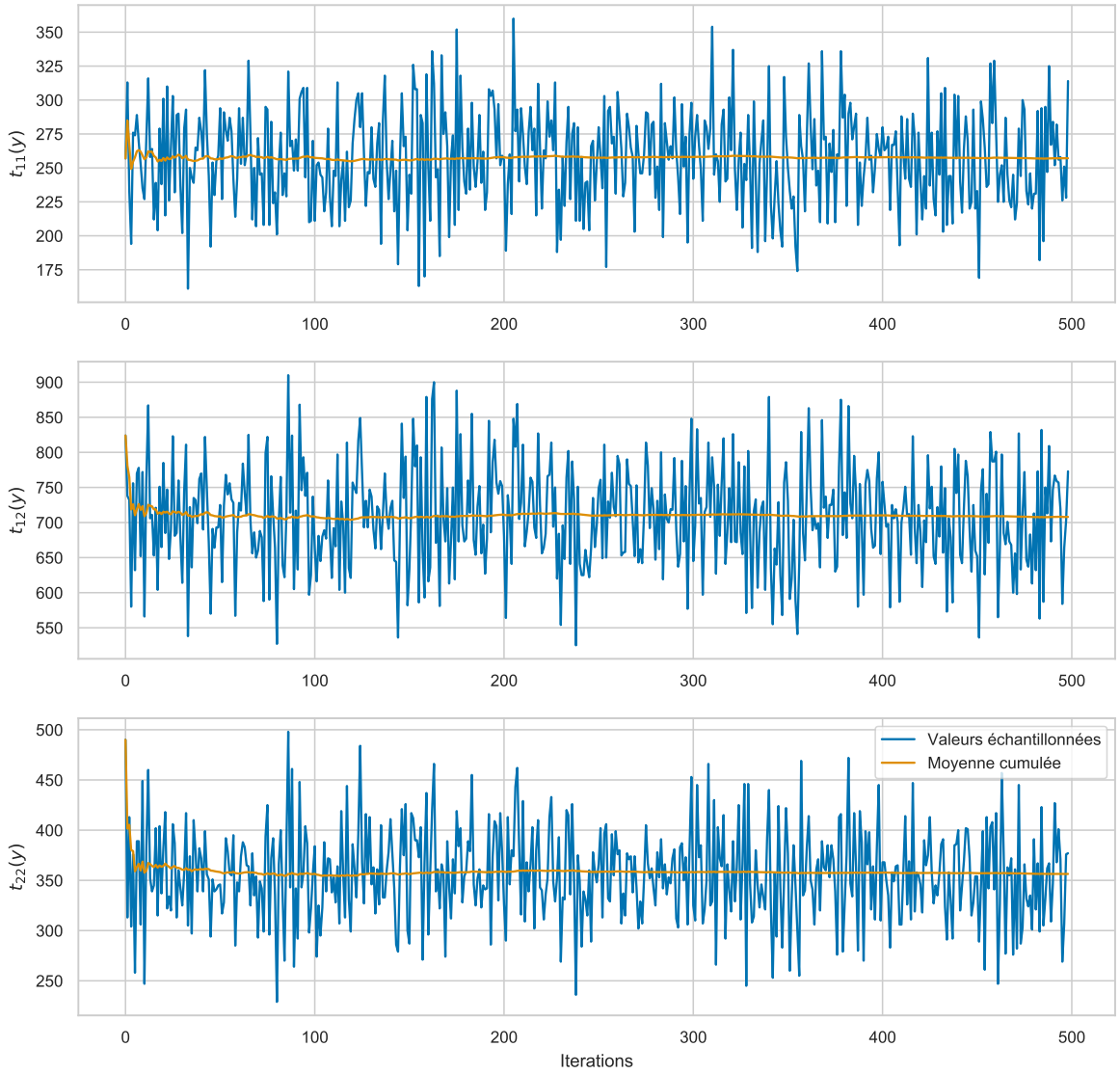
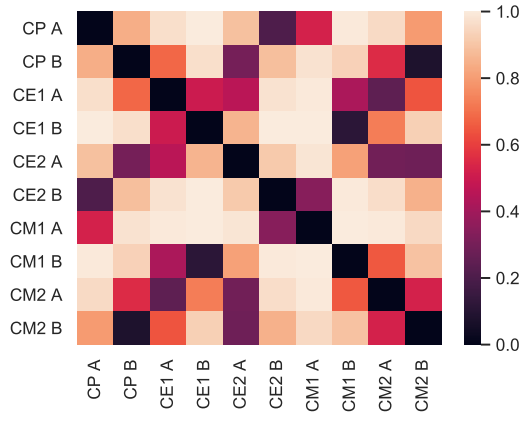


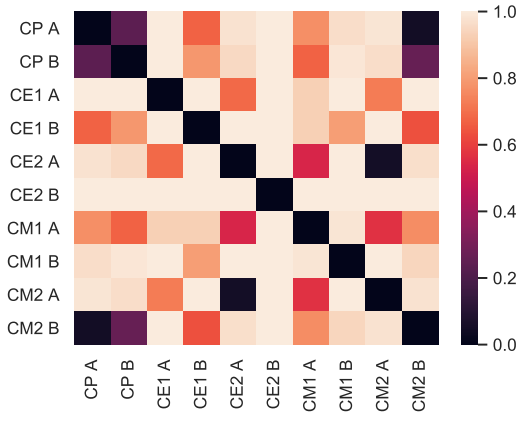
FIGURE A.3 – Statistiques suffisantes des échantillons simulés par l'échantillonneur de Gibbs pour le modèle paramétré avec $\theta = [-0.2, 0.15, -0.25]$. La valeur moyenne des statistiques suffisantes est $\bar{t}(y) = [257, 708, 356]$.

Annexe B

Étude des contacts entre les élèves d'une école primaire



(a) Jour 1



(b) Jour 2

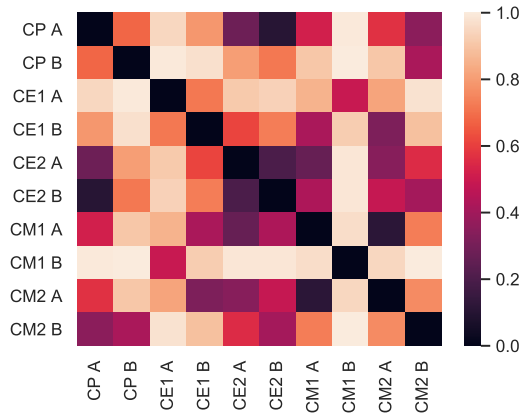
FIGURE B.1 – Cartes de chaleur représentant les valeurs des distances K-S entre les distributions échantillonnées du paramètre θ_{11} pour les deux journées : Figures B.1(a) et B.1(b) .

TABLE B.1 – Erreurs sur les estimations : l'écart-type asymptotique ($\hat{\sigma}_\theta$) et l'écart-type de Monte-Carlo ($\hat{\sigma}_\theta^{MC}$)

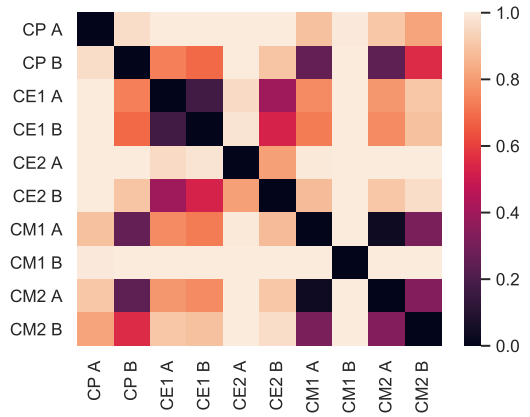
classe	période	$\hat{\sigma}_{\theta_{11}}$	$\hat{\sigma}_{\theta_{12}}$	$\hat{\sigma}_{\theta_{22}}$	$\hat{\sigma}^{MC}_{\theta_{11}}$	$\hat{\sigma}^{MC}_{\theta_{12}}$	$\hat{\sigma}^{MC}_{\theta_{22}}$
CP A	jour 1	3.075e-03	6.042e-03	1.250e-02	2.662e-06	2.290e-06	1.139e-05
	jour 2	2.733e-03	7.940e-03	2.195e-02	4.68e-06	5.968e-06	4.276e-05
CP B	jour 1	2.214e-03	4.968e-03	6.825e-03	9.211e-07	1.160e-06	2.214e-06
	jour 2	3.152e-03	8.161e-03	3.045e-02	6.901e-06	1.279e-05	8.13e-05
CE1 A	jour 1	6.241e-03	1.209e-02	1.813e-02	7.127e-06	1.059e-05	2.559e-05
	jour 2	3.212e-03	6.222e-03	5.209e-03	1.074e-06	1.861e-06	1.012e-06
CE1 B	jour 1	2.985e-03	8.091e-03	1.643e-02	3.805e-06	5.331e-06	2.277e-05
	jour 2	2.728e-03	5.697e-03	8.647e-03	1.538e-06	1.983e-06	5.146e-06
CE2 A	jour 1	3.691e-03	7.667e-03	1.252e-02	3.139e-06	3.8e-06	1.165e-05
	jour 2	2.789e-03	4.107e-03	5.360e-03	6.848e-07	6.305e-07	1.321e-06
CE2 B	jour 1	3.639e-03	6.859e-03	1.473e-02	3.857e-06	3.834e-06	1.718e-05
	jour 2	9.753e-03	1.707e-02	4.284e-02	1.954e-05	3.968e-05	1.087e-04
CM1 A	jour 1	3.465e-03	6.714e-03	1.182e-02	2.585e-06	2.463e-06	8.440e-06
	jour 2	4.736e-03	8.36e-03	1.248e-02	3.691e-06	4.279e-06	1.094e-05
CM1 B	jour 1	3.035e-03	1.145e-02	2.253e-02	5.185e-06	1.311e-05	4.44e-05
	jour 2	8.666e-03	1.970e-02	2.368e-02	1.075e-05	3.102e-05	4.180e-05
CM2 A	jour 1	3.469e-03	8.524e-03	1.404e-02	3.533e-06	6.116e-06	1.543e-05
	jour 2	3.568e-03	6.548e-03	1.095e-02	2.321e-06	2.925e-06	7.229e-06
CM2 B	jour 1	3.403e-03	6.835e-03	1.548e-02	4.172e-06	5.112e-06	2.094e-05
	jour 2	3.054e-03	7.76e-03	1.693e-02	4.061e-06	5.593e-06	2.466e-05

TABLE B.2 – Intervalles de confiances (95%) sur les MAPs estimés, calculés à partir de l'écart-type de Monte-Carlo (Table B.1)

classe	période	CI 95% θ_{11}	CI 95% θ_{12}	CI 95% θ_{22}
CP A	jour 1	0.0726302 ± 5.325e-06	-0.10609111 ± 4.580e-06	-0.09856755 ± 2.278e-05
	jour 2	0.06002834 ± 9.36e-06	-0.13710483 ± 1.194e-05	-0.1584851 ± 8.551e-05
CP B	jour 1	0.06277702 ± 1.842e-06	-0.0952834 ± 2.321e-06	-0.03113339 ± 4.429e-06
	jour 2	0.060075 ± 1.380e-05	-0.1068226 ± 2.558e-05	-0.11482637 ± 1.626e-04
CE1 A	jour 1	0.05698039 ± 1.425e-05	-0.14531345 ± 2.119e-05	-0.15753439 ± 5.117e-05
	jour 2	0.08130508 ± 2.149e-06	-0.09278148 ± 3.721e-06	-0.03950996 ± 2.023e-06
CE1 B	jour 1	0.05032459 ± 7.610e-06	-0.12408186 ± 1.066e-05	-0.15396297 ± 4.554e-05
	jour 2	0.05240562 ± 3.076e-06	-0.09599088 ± 3.965e-06	-0.07369044 ± 1.029e-05
CE2 A	jour 1	0.06103879 ± 6.279e-06	-0.11144471 ± 7.599e-06	-0.11211426 ± 2.331e-05
	jour 2	0.07297123 ± 1.37e-06	-0.07508761 ± 1.261e-06	-0.02593955 ± 2.641e-06
CE2 B	jour 1	0.07499943 ± 7.714e-06	-0.10855389 ± 7.669e-06	-0.08473211 ± 3.436e-05
	jour 2	0.11002625 ± 3.907e-05	-0.0865914 ± 7.936e-05	-0.00977771 ± 2.175e-04
CM1 A	jour 1	0.07785562 ± 5.17e-06	-0.11496529 ± 4.925e-06	-0.06419063 ± 1.688e-05
	jour 2	0.06976402 ± 7.383e-06	-0.11142119 ± 8.558e-06	-0.0937617 ± 2.187e-05
CM1 B	jour 1	0.05181835 ± 1.037e-05	-0.15846503 ± 2.621e-05	-0.10176366 ± 8.879e-05
	jour 2	0.04426831 ± 2.15e-05	-0.21535911 ± 6.205e-05	-0.14928851 ± 8.361e-05
CM2 A	jour 1	0.05853673 ± 7.065e-06	-0.1177495 ± 1.223e-05	-0.04987178 ± 3.085e-05
	jour 2	0.07485534 ± 4.642e-06	-0.10897868 ± 5.851e-06	-0.01073477 ± 1.446e-05
CM2 B	jour 1	0.06430615 ± 8.344e-06	-0.10017104 ± 1.022e-05	-0.08897915 ± 4.188e-05
	jour 2	0.05926046 ± 8.122e-06	-0.11668834 ± 1.119e-05	-0.10710676 ± 4.933e-05

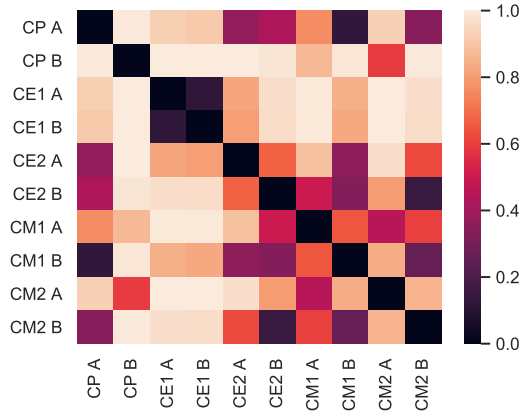


(a) Jour 1

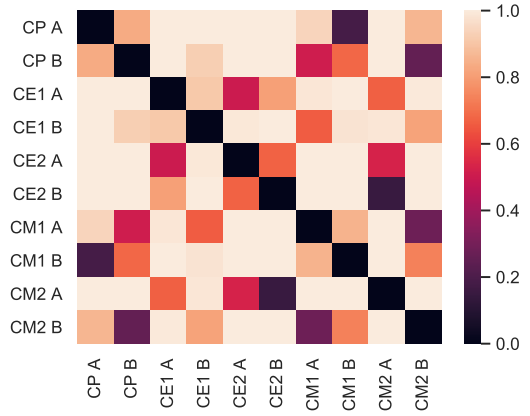


(b) Jour 2

FIGURE B.2 – Cartes de chaleur représentant les valeurs des distances K-S entre les distributions échantillonnées du paramètre θ_{12} pour les deux journées : Figures B.2(a) et B.2(b) .



(a) Jour 1



(b) Jour 2

FIGURE B.3 – Cartes de chaleur représentant les valeurs des distances K-S entre les distributions échantillonnées du paramètre θ_{22} pour les deux journées : Figures B.3(a) et B.3(b) .

Bibliographie

[1] Yves F. Atchadé, Nicolas Lartillot et Christian Robert. « Bayesian Computation for Statistical Models with Intractable Normalizing Constants ». In : *Brazilian Journal of Probability and Statistics* 27.4 (nov. 2013), p. 416–436. ISSN : 0103-0752. DOI : 10.1214/11-BJPS174.

[2] Adrian Baddeley et Jesper Møller. « Nearest-Neighbour Markov Point Processes and Random Sets ». In : *International Statistical Review / Revue Internationale de Statistique* 57.2 (1989), p. 89–121. ISSN : 0306-7734. DOI : 10.2307/1403381.

[3] Albert-László Barabási et Márton Pósfai. *Network Science*. Cambridge University Press, juil. 2016. ISBN : 978-1-107-07626-6.

[4] Verena Bauer, Karl Fürlinger et Göran Kauermann. « A Note on Parallel Sampling in Markov Graphs ». In : *Computational Statistics* 34.3 (sept. 2019), p. 1087–1107. ISSN : 0943-4062, 1613-9658. DOI : 10.1007/s00180-019-00880-4.

[5] Thomas Bayes et Richard Price. « LII. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S ». In : *Philosophical Transactions of the Royal Society of London* 53 (1er jan. 1763), p. 370–418. DOI : 10.1098/rstl.1763.0053. URL : https://royalsocietypublishing.org/doi/10.1098/rstl.1763.0053 (visité le 18/01/2021).

[6] Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin et Christian P. Robert. « Adaptive Approximate Bayesian Computation ». In : *Biometrika* 96.4 (déc. 2009), p. 983–990. ISSN : 0006-3444. DOI : 10.1093/biomet/asp052.

[7] Julian Besag. « Nearest-Neighbour Systems and the Auto-Logistic Model for Binary Data ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 34.1 (1972), p. 75–83. ISSN : 0035-9246.

[8] Julian Besag. « Spatial Interaction and the Statistical Analysis of Lattice Systems ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), p. 192–236. ISSN : 0035-9246.

[9] Gérard Biau, Frédéric Cérou et Arnaud Guyader. « New Insights into Approximate Bayesian Computation ». In : *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 51.1 (fév. 2015), p. 376–403. ISSN : 0246-0203. DOI : 10.1214/13-AIHP590.

[10] Bela Bollobas. *Modern Graph Theory*. Springer Science & Business Media, déc. 2013. ISBN : 978-1-4612-0619-4.

[11] Stephen Borgatti et Pacey Foster. « The Network Paradigm in Organizational Research : A Review and Typology ». In : *Journal of Management* 29 (déc. 2003), p. 991–1013. DOI : 10.1016/S0149-2063_03_00087-4.

[12] Janet M. Box-Steffensmeier, Benjamin W. Campbell, Dino P. Christenson et Zachary Navabi. « Role Analysis Using the Ego-ERGM : A Look at Environmental Interest Group Coalitions ». In : *Social Networks* 52 (jan. 2018), p. 213–227. ISSN : 0378-8733. DOI : 10.1016/j.socnet.2017.08.004.

[13] Julien Brailly et Emmanuel Lazega. « Diversité des approches de modélisation statistique en analyse de réseaux sociaux multiniveaux ». In : *Mathématiques et sciences humaines* 198 (juil. 2012), p. 5–32. ISSN : 0987-6936, 1950-6821. DOI : 10.4000/msh.12226.

[14] Daniel J. Brass, Joseph Galaskiewicz, Henrich R. Greve et Wenpin Tsai. « Taking Stock of Networks and Organizations : A Multilevel Perspective ». In : *Academy of Management Journal* 47.6 (déc. 2004), p. 795–817. ISSN : 0001-4273. DOI : 10.5465/20159624.

[15] Alberto Caimo et Nial Friel. « Bayesian Inference for Exponential Random Graph Models ». In : *Social Networks* 33.1 (jan. 2011), p. 41–55. ISSN : 0378-8733. DOI : 10.1016/j.socnet.2010.09.004.

[16] Alberto Caimo et Nial Friel. « Bayesian Model Selection for Exponential Random Graph Models ». In : *arXiv :1201.2337 [stat]* (jan. 2013). arXiv : 1201.2337 [stat].

[17] Alberto Caimo et Nial Friel. « Bergm : Bayesian Exponential Random Graphs in R ». In : *arXiv :1201.2770 [stat]* (jan. 2012). arXiv : 1201.2770 [stat].

[18] Peter J Carrington, John Scott et Stanley Wasserman. *Models and Methods in Social : Network Analysis*. 2005.

[19] Christopher J. Conselice, Aaron Wilkinson, Kenneth Duncan et Alice Mortlock. « The Evolution of Galaxy Number Density at z $<$ 8 and Its Implications ». In : *The Astrophysical Journal* 830.2 (oct. 2016), p. 83. ISSN : 1538-4357. DOI : 10.3847/0004-637X/830/2/83. arXiv : 1607.03909.

[20] Esteve Corbera, Laura Calvet-Mir, Hannah Hughes et Matthew Paterson. « Patterns of Authorship in the IPCC Working Group III Report ». In : *Nature Climate Change* 6.1 (jan. 2016), p. 94–99. ISSN : 1758-678X, 1758-6798. DOI : 10.1038/nclimate2782.

[21] Quang-Vinh Dang et Claudia-Lavinia Ignat. « Performance of Real-Time Collaborative Editors at Large Scale : User Perspective ». In : *2016 IFIP Networking Conference (IFIP Networking) and Workshops*. Mai 2016, p. 548–553. DOI : 10.1109/IFIPNetworking.2016.7497258.

[22] Galina Daraganova et Garry Robins. « Autologistic Actor Attribute Models ». In : *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. isbn : 978-0-521-19356-6.

[23] Anwitaman Datta, Sonja Buchegger, Le-Hung Vu, Thorsten Strufe et Krzysztof Rzadca. « Decentralized Online Social Networks ». In : *Handbook of Social Network Technologies and Applications*. Sous la dir. de Borko Furht. Boston, MA : Springer US, 2010, p. 349–378. isbn : 978-1-4419-7142-5. doi : 10.1007/978-1-4419-7142-5_17.

[24] Marijtje A. J. Duijn, Tom A. B. Snijders et Bonne J. H. Zijlstra. « P2 : A Random Effects Model with Covariates for Directed Graphs ». In : *Statistica Neerlandica* 58.2 (mai 2004), p. 234–254. issn : 0039-0402, 1467-9574. doi : 10.1046/j.0039-0402.2003.00258.x.

[25] Paul Erdős et Alfréd Rényi. « On Random Graphs ». In : *Publicationes Mathematicae* 6 (1959), p. 290–297.

[26] Goran Eriksson, Christer Holmberg et Stefan Hakansson. *Web Real-Time Communication Use Cases and Requirements*. https ://tools.ietf.org/html/rfc7478.

[27] Tim S. Evans et Renaud Lambiotte. « Line Graphs, Link Partitions, and Overlapping Communities ». In : *Physical Review E* 80.1 (juil. 2009), p. 016105. doi : 10.1103/PhysRevE.80.016105.

[28] Ove Frank et David Strauss. « Markov Graphs ». In : *Journal of the American Statistical Association* 81.395 (sept. 1986), p. 832–842. issn : 0162-1459. doi : 10.1080/01621459.1986.10478342.

[29] Carlo Gaetan et Xavier Guyon. *Modélisation et Statistique Spatiales*. Mathématiques et Applications. Berlin Heidelberg : Springer-Verlag, 2008. isbn : 978-3-540-79225-3.

[30] Carlo Gaetan et Xavier Guyon. « Simulation Des Modèles Spatiaux ». In : *Modélisation et Statistique Spatiales*. Mathématiques et Applications. Berlin Heidelberg : Springer-Verlag, 2008. isbn : 978-3-540-79225-3.

[31] Wojciech Galuba. *Friend-to-Friend Computing : Building the Social Web at the Internet Edges*. Rapp. tech. REP_WORK. Ecole Polytechnique Fédérale de Lausanne (EPFL), 2008.

[32] Stuart Geman et Donald Geman. « Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (nov. 1984), p. 721–741. issn : 1939-3539. doi : 10.1109/TPAMI.1984.4767596.

[33] Charles J. Geyer. « Likelihood Inference for Spatial Point Processes ». In : *Stochastic Geometry : Likelihood and Computation*. Sous la dir. de 0. E. Barndorff-Nielsen, W. S. Kendall et M. N. M. Van Lieshout. Chapman and Hall, 1999.

[34] Charles J. Geyer. « On the Convergence of Monte Carlo Maximum Likelihood Calculations ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 56.1 (jan. 1994), p. 261–274. issn : 0035-9246. doi : 10.1111/j.2517-6161.1994.tb01976.x.

[35] Charles J. Geyer et Elizabeth A. Thompson. « Constrained Monte Carlo Maximum Likelihood for Dependent Data ». In : *Journal of the Royal Statistical Society. Series B (Methodological)* (1992), p. 657–699.

[36] Wally R. Gilks, Gareth O. Roberts et Edward I. George. « Adaptive Direction Sampling ». In : *Journal of the Royal Statistical Society : Series D (The Statistician)* 43.1 (1994), p. 179–189. issn : 1467-9884. doi : 10.2307/2348942.

[37] Aude Grelaud, Christian P. Robert, Jean-Michel Marin, François Rodolphe et Jean-François Taly. « ABC Likelihood-Free Methods for Model Choice in Gibbs Random Fields ». In : *Bayesian Analysis* 4.2 (juin 2009), p. 317–335. issn : 1936-0975, 1931-6690. doi : 10.1214/09-BA412.

[38] Angela M. Guerrero, Örjan Bodin, Ryan R. J. McAllister et Kerrie A. Wilson. « Achieving Social-Ecological Fit through Bottom-up Collaborative Governance : An Empirical Investigation ». In : *Ecology and Society* 20.4 (2015). issn : 1708-3087. doi : 10.5751/ES-08035-200441.

[39] Angela M. Guerrero, Ryan R.J. Mcallister et Kerrie A. Wilson. « Achieving Cross-Scale Collaboration for Large Scale Conservation Initiatives ». In : *Conservation Letters* 8.2 (mar. 2015), p. 107–117. issn : 1755-263X. doi : 10.1111/conl.12112.

[40] Barbara Guidi, Tobias Amft, Andrea De Salve, Kalman Graffi et Laura Ricci. « DiDuSoNet : A P2P Architecture for Distributed Dunbar-Based Social Networks ». In : *Peer-to-Peer Networking and Applications* 9.6 (nov. 2016), p. 1177–1194. issn : 1936-6442, 1936-6450. doi : 10.1007/s12083-015-0366-7.

[41] John M. Hammersley et Peter Clifford. « Markov Field on Finite Graphs and Lattices ». 1971.

[42] Mark S. Handcock. « Assessing Degeneracy in Statistical Models of Social Networks ». In : *Journal of the American Statistical Association* 76 (2003), p. 33–50.

[43] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau et Martina Morris. « Statnet : Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data ». In : *Journal of statistical software* 24.1 (2008), p. 1548–7660. issn : 1548-7660.

[44] Frank Harary. « Line Graphs ». In : *Graph Theory*. Addison-Wesley Series in Mathematics. Addison-Wesley, 1969.

[45] Wilfred Keith Hastings. « Monte Carlo Sampling Methods Using Markov Chains and Their Applications ». In : *Biometrika* 57.1 (1970), p. 13.

[46] Paul W. Holland et Samuel Leinhardt. « An Exponential Family of Probability Distributions for Directed Graphs ». In : *Journal of the American Statistical Association* 373 (1981), p. 54–57.

[47] Mark L. Huber. *Perfect Simulation*. CRC Press, jan. 2016. isbn : 978-1-4822-3245-5.

[48] Ruth M. Hummel, David R. Hunter et Mark S. Handcock. « Improving Simulation-Based Algorithms for Fitting ERGMs ». In : *Journal of Computational and Graphical Statistics : A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 21.4 (déc. 2012), p. 920–939. issn : 1061-8600. doi : 10.1080/10618600.2012.679224.

[49] David R. Hunter et Mark S. Handcock. « Inference in Curved Exponential Family Models for Networks ». In : *Journal of Computational and Graphical Statistics* 15.3 (sept. 2006), p. 565–583. issn : 1061-8600. doi : 10.1198/106186006X133069.

[50] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau et Martina Morris. « Ergm : A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks ». In : *Journal of statistical software* 24.3 (mai 2008), nihpa54860. issn : 1548-7660.

[51] David R. Hunter, Pavel N. Krivitsky et Michael Schweinberger. « Computational Statistical Methods for Social Network Models ». In : *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 21.4 (déc. 2012), p. 856–882. issn : 1061-8600.

[52] International Consortium of Investigative Journalists. *The Panama Papers : About the Investigation*. https ://www.icij.org/investigations/panama-papers/pages/panama-papers-about-the-investigation/.

[53] Richard M. Karp. « Reducibility among Combinatorial Problems ». In : *Complexity of Computer Computations*. Sous la dir. de Raymond E. Miller, James W. Thatcher et Jean D. Bohlinger. The IBM Research Symposia Series. Boston, MA : Springer US, 1972, p. 85–103. isbn : 978-1-4684-2001-2. doi : 10.1007/978-1-4684-2001-2_9.

[54] Dale V. Kent. *The Rise of the Medici : Faction in Florence 1426-1434*. Oxford university press. Oxford, 1978.

[55] Johan Koskinen. *Bayesian Analysis of Exponential Random Graphs-Estimation of Parameters and Model Selection*. Rapp. tech. Research Report 2004 : 2, Department of Statistics, Stockholm University, 2004.

[56] Johan Koskinen et Galina Daraganova. « Dependence Graphs and Sufficient Statistics ». In : *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. isbn : 978-0-521-19356-6.

[57] Johan Koskinen et Galina Daraganova. « Exponential Random Graph Model Fundamentals ». In : *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. isbn : 978-0-521-19356-6.

[58] Johan Koskinen et Tom A. B. Snijders. « Simulation, Estimation, and Goodness of Fit ». In : *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. isbn : 978-0-521-19356-6.

[59] Quentin Laporte-Chabasse, Radu S. Stoica, Marianne Clausel, François Charoy et Gérald Oster. *Co-Authoring Graphs of Research Teams in a Laboratory in Computer Science*. Oct. 2020. doi : 10.5281/zenodo.4088670.

[60] Quentin Laporte-Chabasse, Radu Stoica, Marianne Clausel, François Charoy et Gérald Oster. « Morpho-statistical description of networks through graph modelling and Bayesian inference ». working paper or preprint. Déc. 2020. url : https://hal.archives-ouvertes.fr/hal-02421787.

[61] Emmanuel Lazega et Tom A. B. Snijders. *Multilevel Network Analysis for the Social Sciences : Theory, Methods and Applications*. Springer, déc. 2015. isbn : 978-3-319-24520-1. doi : 10.1007/978-3-319-24520-1.

[62] Youngmi Lee, In Won Lee et Richard C. Feiock. « Interorganizational Collaboration Networks in Economic Development Policy : An Exponential Random Graph Model Analysis* ». In : *Policy Studies Journal* 40.3 (août 2012), p. 547–573. issn : 1541-0072. doi : 10.1111/j.1541-0072.2012.00464.x.

[63] Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Science & Business Media, 2009.

[64] Dean Lusher, Johan Koskinen et Garry Robins, éds. *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. isbn : 978-0-521-19356-6.

[65] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert et Robin J. Ryder. « Approximate Bayesian Computational Methods ». In : *Statistics and Computing* 22.6 (nov. 2012), p. 1167–1180. issn : 1573-1375. doi : 10.1007/s11222-011-9288-2.

[66] Sean P. Meyn et Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering. London : Springer-Verlag, 1993. isbn : 978-1-4471-3269-1.

[67] J. Møller, A. N. Pettitt, R. Reeves et K. K. Berthelsen. « An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants ». In : *Biometrika* 93.2 (2006), p. 451–458. issn : 0006-3444.

[68] Jesper Møller et Rasmus Plenge Waagepetersen. « Markov Connected Component Fields ». In : *Advances in Applied Probability* 30.1 (1998), p. 1–35. ISSN : 0001-8678.

[69] Martina Morris, Mark S. Handcock et David R. Hunter. « Specification of Exponential-Family Random Graph Models : Terms and Computational Aspects ». In : *Journal of statistical software* 24.4 (2008), p. 1548–7660. ISSN : 1548-7660.

[70] Martina Morris et Mirjam Kretzschmar. « Concurrent Partnerships and the Spread of HIV : » in : *AIDS* 11.5 (avr. 1997), p. 641–648. ISSN : 0269-9370. DOI : 10.1097/00002030-199705000-00012.

[71] Iain Murray. « Advances in Markov Chain Monte Carlo Methods ». Thèse de doct. Gatsby computational neuroscience unit, University College London, 2007.

[72] Iain Murray, Zoubin Ghahramani et David J. C. MacKay. « MCMC for Doubly-Intractable Distributions ». In : *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'06. Arlington, Virginia, USA : AUAI Press, juil. 2006, p. 359–366. ISBN : 978-0-9749039-2-7.

[73] Brice Nédelec, Julian Tanke, Davide Frey, Pascal Molli et Achour Mostéfaoui. « An Adaptive Peer-Sampling Protocol for Building Networks of Browsers ». In : *World Wide Web* (août 2017), p. 1–33. ISSN : 1386-145X, 1573-1413. DOI : 10.1007/s11280-017-0478-5.

[74] Mark E. J. Newman. « The Structure of Scientific Collaboration Networks ». In : *Proceedings of the National Academy of Sciences* 98.2 (jan. 2001), p. 404–409. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.98.2.404.

[75] Matthieu Nicolas, Victorien Elvinger, Gérald Oster, Claudia-Lavinia Ignat et François Charoy. « MUTE : A Peer-to-Peer Web-Based Real-Time Collaborative Editor ». In : *ECSCW 2017 - 15th European Conference on Computer-Supported Cooperative Work*. T. 1. EUSSET, août 2017, p. 1–4. DOI : 10.18420/ecscw2017_p5.

[76] OpenMP Architecture Review Board. *OpenMP Application Program Interface*. 2008.

[77] Srikanth Paruchuri, Martin C. Goossen et Corey C. Phelps. *Conceptual Foundations of Multilevel Social Networks*. SSRN Scholarly Paper ID 3323503. Rochester, NY : Social Science Research Network, jan. 2019.

[78] Philippa E. Pattison, Garry L. Robins, Tom A. B. Snijders et Peng Wang. « Conditional Estimation of Exponential Random Graph Models from Snowball Sampling Designs ». In : *Journal of Mathematical Psychology*. Social Networks 57.6 (déc. 2013), p. 284–296. ISSN : 0022-2496. DOI : 10.1016/j.jmp.2013.05.004.

[79] Corey Phelps, Ralph Heidl et Anu Wadhwa. « Knowledge, Networks, and Knowledge Networks : A Review and Research Agenda ». In : *Journal of Management* 38.4 (juil. 2012), p. 1115–1166. ISSN : 0149-2063. DOI : 10.1177/0149206311432640.

[80] Derek J. De Solla Price. « Networks of Scientific Papers ». In : *Science* 149.3683 (1965), p. 510–515. ISSN : 0036-8075.

[81] James Gary Propp et David Bruce Wilson. « Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics ». In : *Random Structures & Algorithms* 9.1-2 (1996), p. 223–252. ISSN : 1098-2418. DOI : 10.1002/(SICI)1098-2418(199608/09)9:1/2<223::AID-RSA14>3.0.CO;2-O.

[82] Herbert Robbins et Sutton Monro. « A Stochastic Approximation Method ». In : *The Annals of Mathematical Statistics* 22.3 (1951), p. 400–407. ISSN : 0003-4851.

[83] Garry Robins et Galina Daraganova. « Social Selection, Dyadic Covariates, and Geospatial Effects ». In : *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. ISBN : 978-0-521-19356-6.

[84] Garry Robins, Peter Elliott et Philippa Pattison. « Network Models for Social Selection Processes ». In : *Social Networks* 23.1 (jan. 2001), p. 1–30. ISSN : 0378-8733. DOI : 10.1016/S0378-8733(01)00029-6.

[85] Garry Robins et Philippa Pattison. « Interdependencies and Social Processes : Dependence Graphs and Generalized Dependence Structures ». In : *Models and Methods in Social Network Analysis*. T. 28. 2005. DOI : 10.1017/CBO9780511811395.010.

[86] Garry Robins, Philippa Pattison et Peter Elliott. « Network Models for Social Influence Processes ». In : *Psychometrika* 66.2 (juin 2001), p. 161–189. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02294834.

[87] Samuel Franklin Sampson. « A Novitiate in a Period of Change. An Experimental and Case Study of Social Relationships ». Thèse de doct. Cornell University, 1968.

[88] Rajesh Sharma, Anwitaman Datta, Matteo DeH'Amico et Pietro Michiardi. « An Empirical Study of Availability in Friend-to-Friend Storage Systems ». In : *2011 IEEE International Conference on Peer-to-Peer Computing*. Août 2011, p. 348–351. DOI : 10.1109/P2P.2011.6038754.

[89] Tom A. B. Snijders. « Markov Chain Monte Carlo Estimation of Exponential Random Graph Models ». In : *Journal of Social Structure* 3.2 (2002), p. 1–40.

[90] Tom A. B. Snijders et Chris Baerveldt. « A Multilevel Network Study of the Effects of Delinquent Behavior on Friendship Evolution ». In : *The Journal of Mathematical Sociology* 27.2-3 (avr. 2003), p. 123–151. ISSN : 0022-250X. DOI : 10.1080/00222500305892.

[91] Tom A. B. Snijders et Roel J. Bosker. *Multilevel Analysis : An Introduction to Basic and Advanced Multilevel Modeling*. SAGE, oct. 2011. ISBN : 978-1-4462-5433-2.

[92] Tom A. B. Snijders, Philippa E. Pattison, Garry L. Robins et Mark S. Handcock. « New Specifications for Exponential Random Graph Models ». In : *Sociological Methodology* 36.1 (août 2006), p. 99–153. issn : 0081-1750. doi : 10.1111/j.1467-9531.2006.00176.x.

[93] *SocioPatterns.Org*. http ://www.sociopatterns.org/.

[94] Juliette Stehlé, Nicolas Voirin, Alain Barrat, Ciro Cattuto, Lorenzo Isella, Jean-François Pinton, Marco Quaggiotto, Wouter Van den Broeck, Corinne Régis, Bruno Lina et Philippe Vanhems. « High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School ». In : *PLOS ONE* 6.8 (août 2011), e23176. issn : 1932-6203. doi : 10.1371/journal.pone.0023176.

[95] Radu S. Stoica, Madalina Deaconu, Anne Philippe et Lluis Hurtado-Gil. « Shadow Simulated Annealing Algorithm : A New Tool for Global Optimisation and Statistical Inference ». Juil. 2019.

[96] Radu S. Stoica, Anne Philippe, Pablo Gregori et Jorge Mateu. « ABC Shadow Algorithm : A Tool for Statistical Analysis of Spatial Patterns ». In : *Statistics and Computing* 27.5 (sept. 2017), p. 1225–1238. issn : 0960-3174, 1573-1375. doi : 10.1007/s11222-016-9682-x.

[97] David Strauss. « On a General Class of Models for Interaction ». In : *SIAM Review* 28.4 (déc. 1986), p. 513–527. issn : 0036-1445, 1095-7200. doi : 10.1137/1028156.

[98] David Strauss et Michael Ikeda. « Pseudolikelihood Estimation for Social Networks ». In : *Journal of the American Statistical Association* 85.409 (1990), p. 204–212. issn : 0162-1459. doi : 10.2307/2289546.

[99] Luke Tierney. « Introduction to General State-Space Markov Chain Theory ». In : *Markov Chain Monte Carlo in Practice*. Sous la dir. de W. R. Gilks, S. Richardson et D. J. Spiegelhalter. Boca Raton, Fla : Chapman & Hall, 1998. isbn : 978-0-412-05551-5.

[100] Pen Wang. « Exponential Random Graph Model Extensions : Models for Multiple Networks and Bipartite Networks ». In : *Exponential Random Graph Models for Social Networks : Theory, Methods, and Applications*. New York, N.Y., Royaume-Uni de Grande-Bretagne et d'Irlande du Nord : Cambridge University Press, 2013. isbn : 978-0-521-19356-6.

[101] Peng Wang, Garry Robins, Philippa Pattison et Emmanuel Lazega. « Exponential Random Graph Models for Multilevel Networks ». In : *Social Networks* 35.1 (jan. 2013), p. 96–115. issn : 0378-8733. doi : 10.1016/j.socnet.2013.01.004.

[102] Peng Wang, Ken Sharpe, Garry L. Robins et Philippa E. Pattison. « Exponential Random Graph (P_*) Models for Affiliation Networks ». In : *Social Networks* 31.1 (jan. 2009), p. 12–25. issn : 0378-8733. doi : 10.1016/j.socnet.2008.08.002.

[103] Stanley Wasserman et Katherine Faust. *Social Network Analysis : Methods and Applications*. Cambridge University Press, nov. 1994. isbn : 978-0-521-38707-1.

[104] Stanley Wasserman et Philippa Pattison. « Logit Models and Logistic Regressions for Social Networks : I. An Introduction to Markov Graphs Andp ». In : *Psychometrika* 61.3 (sept. 1996), p. 401–425. issn : 1860-0980. doi : 10.1007/BF02294547.

[105] Harrison C. White. *Identity and Control : How Social Formations Emerge*. Princeton university press, 2008.

[106] Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2003. isbn : 978-3-642-62911-2. doi : 10.1007/978-3-642-55760-6.

[107] B. Beverly Yang et Hector Garcia-Molina. « Designing a Super-Peer Network ». In : *Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405)*. IEEE, 2003, p. 49–60.

[108] Paola Zappa et Alessandro Lomi. « Knowledge Sharing in Organizations : A Multilevel Network Analysis ». In : *Multilevel Network Analysis for the Social Sciences*. Springer, 2016, p. 333–353. doi : 10.1007/978-3-319-24520-1_14.

Résumé

Les applications collaboratives décentralisées permettent de répondre aux problèmes de confidentialité, de disponibilité et de sécurité inhérents aux plateformes collaboratives centralisées. Elles reposent sur un paradigme de communication pair-à-pair selon lequel tous les utilisateurs sont directement connectés les uns aux autres. Les collaborations ayant tendance à s'élargir et dépasser les frontières des organisations, il est nécessaire de garantir aux utilisateurs le contrôle sur leurs données tout en assurant la disponibilité de la collaboration. Pour ce faire, il est possible d'utiliser comme topologie le réseau social qui s'est tissé entre les collaborateurs. Le manque d'information sur ce maillage de confiance nous amène à développer une approche pour étudier ses propriétés morphologiques.

Dans cette thèse, nous développons et mettons en œuvre une approche permettant d'étudier la structure sociale des interactions dans le cadre de collaborations inter-organisationnelles. Nous proposons une approche stochastique qui s'inspire des *Exponential Random Graph Model*s et des modèles spatiaux. Nous définissons un formalisme qui met en avant la structure des interactions et intègre la dimension organisationnelle. Nous proposons d'utiliser une méthode d'inférence bayésienne, ABC Shdadow, pour contourner les difficultés liées à l'estimation de ce modèle. Cette approche est mise en œuvre sur un exemple réel : les collaborations initiées par les chercheurs d'un laboratoire. Elle permet notamment de montrer la faible propension, pour un chercheur, à tisser des liens avec d'autres laboratoires. Nous montrons que cette approche peut être appliquée à d'autres types d'interactions sociales, comme les interactions entre les enfants d'une école primaire. Enfin, nous présentons une stratégie de parallélisation de l'échantillonneur de Gibbs visant à traiter des graphes de plus grande taille dans un temps raisonnable.

Mots-clés: Collaborations inter-organisationnelles, Analyse des réseaux sociaux, *Exponential Random Graph Model* (ERGM), Inférence bayésienne.

Abstract

Decentralised collaborative applications address privacy, availability and security issues related to centralised collaborative platforms. Such applications are based on a peer-to-peer communication paradigm according to which all users are directly connected to one another. Collaborations tend to widen and spread beyond the borders of organisations. Under these circumstances, it is necessary to guarantee to users the control over their data, while keeping collaboration available. To that end, the social network that has built between collaborators may be used as topology. Lack of information on this trusted network leads us to develop an approach to study its morphological properties.

In this thesis, we develop and implement an approach to study the social structure of interactions in the context of inter-organisational collaborations. We propose a stochastic approach based on *Exponential Random Graph Model*s and spatial models. We define a formalism that highlights the structure of interactions and integrates the organisational dimension. We propose to use a Bayesian inference method, ABC Shadow, to overcome the issues related to the parameters estimation. This approach is applied to a real case study : the collaborations initiated by researchers in a laboratory. In particular, it highlights the low tendency for a researcher to create collaborative links with other laboratories. We show that this approach can be applied to other kinds of social interactions, such as interactions between pupils of a primary school. Finally, we present a parallelisation strategy of the Gibbs sampler aimed at processing larger graphs in a reasonable time.

Keywords: Inter-organisational collaborations, *Social network analysis* (SNA), *Exponential Random Graph Model* (ERGM), Bayesian inference.