



HAL
open science

Etude structurelle des réseaux : modèles aléatoires, motifs et cycles.

Etienne Birmele

► **To cite this version:**

Etienne Birmele. Etude structurelle des réseaux : modèles aléatoires, motifs et cycles.. Bio-informatique [q-bio.QM]. Université d'Evry-Val d'Essonne, 2011. tel-00750375

HAL Id: tel-00750375

<https://theses.hal.science/tel-00750375>

Submitted on 9 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Etude structurelle des réseaux : modèles aléatoires, motifs et cycles.

Etienne Birmelé

*Laboratoire Statistique et Génome (UMR CNRS 8071),
Tour Évry 2, 523 pl. des Terrasses de l'Agora,
91000 Évry, France*
e-mail: etienne.birmele@genopole.cnrs.fr
url: stat.genopole.cnrs.fr/~ebirmele

Habilitation à diriger des recherches

Soutenue le 3 novembre 2011 à l'Université d'Évry-Val d'Essonne

devant un jury composé de

Pr.	Christophe	AMBROISE	
Pr.	Gilles	BLANCHARD	
Pr.	Adrian	BONDY	
Dr.	Matthieu	LATAPY	<i>Rapporteur</i>
Pr.	Gábor	LUGOSI	<i>Rapporteur</i>
Dr.	Marie-France	SAGOT	<i>Présidente</i>
Pr.	Korbinian	STRIMMER	<i>Rapporteur</i>

Table des matières

Remerciements	5
Contexte	6
L'objet graphe	6
Organisation du mémoire	9
I Modèles aléatoires pour les réseaux	11
1 Caractéristiques topologiques des réseaux	11
2 Modèles aléatoires existants	12
2.1 Modèles <i>pas à pas</i> vs modèles <i>à tirage simultané</i>	12
3 Approche bayésienne du modèle de mélange	18
3.1 Contexte	18
3.2 Approche variationnelle bayésienne	19
3.3 Un critère de choix de modèle	20
3.4 Expériences comparatives	21
3.4.1 Données d'affiliation	21
3.4.2 Données d'affiliation avec hubs	21
3.4.3 Réseau métabolique d' <i>Escherichia coli</i>	22
4 Un modèle de mélange à classes chevauchantes	23
4.1 Contexte	23
4.2 Le modèle OSBM	25
4.3 Identifiabilité	26
4.3.1 Correspondance avec SBM	26
4.3.2 Permutations et inversions	27
4.3.3 Identifiabilité générique	28
4.4 Estimation des paramètres	29
4.5 Expériences comparatives	30
4.5.1 Simulations	32
4.5.2 Blogosphère politique française	33
4.5.3 Réseau de régulation de la levure	34
4.6 Un critère de choix de modèle	34
5 Un modèle basé sur les graphes bipartis	35
5.1 Contexte	35
5.2 Le modèle	36
5.3 Propriétés topologiques des graphes générés	37
5.3.1 Distribution des degrés	37
5.3.2 Densités globale et locale	38
5.3.3 Composante géante et diamètre	39

5.4	Conclusion	39
6	Perspectives	39
II Recherche de motifs dans les réseaux biologiques		41
7	Introduction	41
7.1	Motivations	41
7.2	Schémas et occurrences	42
7.3	Le problème du comptage	43
8	Motifs globaux dans le modèle EDD	43
9	Recherche de motifs locaux	45
9.1	Motifs locaux	47
9.2	Le modèle aléatoire	48
9.3	Approximation de Poisson pour la p -valeur	49
9.3.1	Borne locale	49
9.3.2	Borne globale	51
9.3.3	Procédure de filtrage	52
9.3.4	Aspects algorithmiques	52
9.4	Borne inférieure de la p -valeur	54
9.5	Simulations et application	55
9.5.1	Données simulées	55
9.5.2	Influence du modèle choisi	56
9.5.3	Réseaux réels	58
9.6	Perspectives	59
III Cycles dans les réseaux		63
10	Bornes pour la largeur d'arborescence	63
10.1	Largeur d'arborescence	63
10.2	Mineurs interdits	64
10.3	Borne polynômiale pour la largeur d'arborescence des graphes sans longs cycles	66
10.4	Perspectives	68
11	Cycles et réseaux métaboliques	68
11.1	Réseaux métaboliques consistants	68
11.2	Organisations chimiques et complexité de leur énumération	69
11.3	Cycles bloquants	71
11.4	Algorithme	73
11.5	Perspectives	75
	Références	77
	Publications	83

Liste des co-auteurs 85

Remerciements

Je tiens à remercier Matthieu Latapy, Gábor Lugosi et Korbinian Strimmer d'avoir accepté la tâche fastidieuse consistant à écrire un rapport sur ce travail. Merci également à Marie-France Sagot, Christophe Ambroise, Gilles Blanchard et Adrian Bondy pour leur participation au jury.

Parmi toutes les personnes qui ont jalonné mon parcours scientifique, je voudrais tout particulièrement remercier Adrian Bondy, qui m'a définitivement inoculé le virus de la théorie des graphes, si riche en problèmes qui s'énoncent en deux minutes et se résolvent en autant de décennies (ou pas). Un grand merci également à Christophe Ambroise pour l'initiation aux joies de l'encadrement et de la classification et à Marie-France Sagot pour l'ouverture sur le monde de l'informatique théorique appliquée. Enfin, Bernard Prum et son équipe d'alors ont pris le risque de faire confiance à un jeune non-biologiste à peine-probabiliste pour intégrer leur laboratoire. J'espère que la lecture de ce document leur laissera le sentiment d'un pari réussi.

Depuis mes débuts en thèse, j'ai eu l'occasion de travailler avec de nombreux co-auteurs, sans qui mes travaux n'auraient pas vu le jour et que je tiens à saluer. Merci également à toutes celles et ceux dont les remarques ou suggestions m'ont ouvert de nouvelles perspectives.

Je voudrais ensuite remercier tous les membres passés et présents du laboratoire Statistique et Génome pour l'ambiance de travail particulièrement agréable dans laquelle j'ai la chance d'évoluer. S'il est vrai que partager un café et mieux se connaître améliore les collaborations, l'avenir du laboratoire est assuré! Mes remerciements vont également aux membres des groupes SSB et Simbiosi, en espérant avoir l'occasion de partager encore de nombreux piques-niques post-séminaires ou de repas vin/fromage.

Enfin, je tire mon chapeau à Eloïse et Romane qui n'ont jamais mis en doute que mon travail puisse consister à remplir des bloc-notes entiers de triangles et de flèches, ce que même leur petit frère serait capable de faire.

Introduction

Contexte

Suite à mes études de mathématiques, je me suis spécialisé lors de mon DEA puis de ma thèse en combinatoire et plus spécialement en théorie des graphes. J’y ai étudié, sous la direction de J.A. Bondy, des bornes théoriques sur la largeur d’arborescence qui est une caractéristique des graphes associée aux performances algorithmiques des problèmes qu’on cherche à y résoudre.

Mon intégration au sein du laboratoire *Statistique et Génome* en 2005 m’a amené à changer totalement de thèmes de recherche. En effet, les années 2000 ont été marquées par une très grande augmentation des données disponibles en biologie. Une des conséquences de cette abondance a été l’organisation de ces données en réseaux, c’est-à-dire en données structurées sous forme de graphes afin d’en avoir une vue d’ensemble. Ces réseaux biologiques contenant plusieurs centaines ou milliers de noeuds, les statistiques sont naturellement devenues un outil indispensable de leur étude.

Historiquement, le laboratoire *Statistique et Génome* était surtout spécialisé dans l’étude statistique des séquences. La volonté de créer une équipe de recherche centrée autour de l’étude des réseaux s’est traduite par le recrutement ou la conversion thématique de plusieurs chercheurs ayant des formations en statistique, théorie des graphes et/ou informatique. Cette équipe de recherche est élargie à des chercheurs de l’INRA issu de l’école d’Agronomie de Paris et du centre de Jouy-en-Josas pour former le groupe *SSBNet*. Notre but principal est la recherche des motifs sur-représentés dans les réseaux, c’est-à-dire des petites structures dont la présence importante indique potentiellement une sélection positive au cours de l’évolution. Notre interaction avec le *Laboratoire de Biométrie et de Biologie Evolutive* a donné lieu au projet ANR NeMo.

Depuis deux ans, je participe également à un autre groupe de recherche (équipe associée *SIMBIOSI* de l’INRIA) s’intéressant à la propagation des flux dans les réseaux métaboliques, dans le but de mieux comprendre les mécanismes de la symbiose. Ces travaux portent pour le moment sur des problèmes algorithmiques d’énumération, une partie statistique étant à prévoir dans le futur.

L’objet graphe

Mes travaux scientifiques ont été réalisés dans des cadres assez différents, mêlant les mathématiques discrètes, les statistiques, l’application à la biologie et l’informatique. L’ensemble a cependant un élément commun qui est l’objet graphe.

Cet objet mathématique a l’avantage d’être à la fois simple à appréhender et capable de receler des structures complexes. Ainsi, un graphe G est la donnée d’un ensemble V de sommets et d’un ensemble E d’arêtes les reliant. Il peut être représenté à l’aide d’une matrice d’adjacence ou d’une liste d’arêtes. Lorsqu’il a une taille qui le permet, il est également très simple à représenter graphiquement.

Cependant, le fait qu’il introduise les relations entre sommets en fait un objet permettant de décrire de nombreux phénomènes réels. Parmi les principaux champs d’application des graphes, citons la sociologie,

l'informatique et la biologie. Dans chacun de ces domaines, les applications sont elles-mêmes très variées. Dans le cas de la biologie cellulaire, on peut par exemple citer l'étude des interactions entre protéines, de la régulation entre gènes ou de l'enchaînement de réactions dans le métabolisme.

Types de graphes

Tout au long de ce mémoire, je désignerai par *réseau* les interactions observées et par *graphe* les objets mathématiques manipulés pour étudier le réseau d'intérêt. En fonction du type de réseau étudié, les graphes peuvent être de plusieurs types afin de correspondre au mieux à la réalité :

graphe orienté ou non-orienté : les arêtes d'un graphe non-orienté sont des relations symétriques, au contraire des arêtes d'un graphe orienté. Des interactions physiques entre protéines seront modélisées par des graphes non orientés alors que des processus de régulation entre gènes le seront à l'aide de graphes orientés.

graphe avec ou sans auto-arêtes : une *auto-arête*, ou *boucle*, est une arête reliant un sommet à lui-même. Les boucles sont par exemple présentes dans les réseaux de régulation entre gènes, un gène pouvant s'auto-réguler.

graphe biparti : un graphe biparti est un graphe dont les sommets peuvent être séparés en deux groupes tels que toute arête du graphe relie deux sommets de groupes différents.

graphe coloré : un *graphe coloré* est un graphe dont les sommets ou les arêtes ont une couleur. Il est alors possible de modéliser des contraintes du problème étudié via des contraintes sur les couleurs du graphe. Les réactions d'un graphe métaboliques peuvent être par exemple colorées suivant un code correspondant à la classification des enzymes qui les gouvernent.

graphe valué : un *graphe valué* est un graphe dont les arêtes, qui sont une relation binaire, sont remplacés par une fonction de poids, permettant à nouveau d'introduire de nouvelles données. La similarité de séquences entre deux protéines est ainsi plus finement codée par une arête portant un score obtenu par alignement de séquences que par une arête simple.

graphe dynamique : un *graphe dynamique* est un graphe évoluant en fonction du temps. Par opposition, un graphe n'évoluant pas est dit *statique*.

La liste précédente permet de voir la grande diversité des problèmes pouvant être modélisés à l'aide de graphes. Dans le suite de ce mémoire, je ne considérerai cependant que des graphes non colorés, non valués et statiques. Les différents résultats seront énoncés soit dans le cadre orienté, soit non-orienté mais sauf mention contraire, ils peuvent être étendus à chacune des deux situations.

Notations liées aux graphes

Un graphe non orienté est un couple $G = (V, E)$ où V est un ensemble de sommets et $E \subset V \times V$ un ensemble d'arêtes. Une arête entre les sommets u et v est notée (u, v) . $n(G)$ et $e(G)$ désignent

respectivement le nombre de sommets et le nombre d'arêtes du graphe, en omettant la référence à G quand il n'y a pas d'ambiguïté. Sauf précision contraire, la lettre n est réservée à cet usage.

La *densité* $D(G)$ du graphe est le rapport du nombre d'arêtes présentes sur le nombre d'arêtes possibles, c'est-à-dire, dans le cas des graphes sans boucles,

$$D(G) = \frac{2e(G)}{n(G)(n(G) - 1)}$$

Dans le cas des graphes avec boucles, le dénominateur est remplacé par $n(G)(n(G) + 1)$.

Un graphe H est un *sous-graphe* d'un graphe G si H peut être obtenu à partir de G par suppression de sommets et d'arêtes (la suppression d'un sommet entraînant automatiquement la suppression de toutes les arêtes qui lui sont incidentes). H est un *sous-graphe induit* s'il est obtenu à partir de G en supprimant uniquement des sommets. Un exemple de comparaison des deux définitions est donné Figure 1. Pour tout sous-ensemble U de $V(G)$, la notation $G[U]$ désigne le sous-graphe de G induit par les sommets de U , c'est-à-dire le sous graphe induit de G obtenu en ne gardant que les sommets de U et toutes les arêtes les reliant.

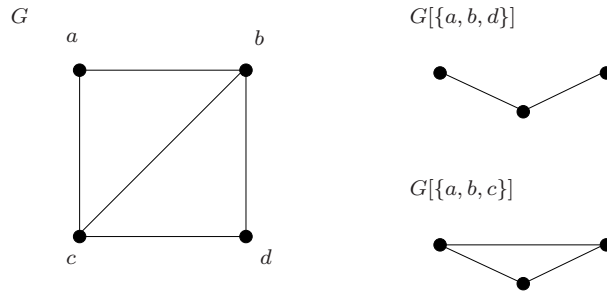


FIGURE 1. Exemples de sous-graphes induits d'un graphe sur deux ensembles de sommets. Le chemin composé de trois arêtes est un sous-graphe de G (il est présent six fois) mais n'en est pas un sous-graphe induit.

Pour un sommet v donné, on note $N(v)$ le voisinage de v , c'est-à-dire l'ensemble des sommets reliés à v par une arête. Le cardinal de $N(v)$ est noté d_v et est appelé le *degré* de v .

Dans le cas d'un graphe orienté, une arête est un couple ordonné de sommets. On note \vec{uv} l'arête de u vers v . Toutes les définitions précédentes se généralisent sans mal aux graphes orientés. Les seules différences notables sont le facteur 2 de la densité qui disparaît et le fait que le voisinage d'un sommet est séparé en un voisinage entrant $N^-(v)$ regroupant les sommets d'où sortent des arêtes dirigées vers v et un voisinage sortant $N^+(v)$ regroupant les sommets atteints par les arêtes partant de v . Cette séparation induit l'existence de deux degrés d_v^- et d_v^+ , appelés degré entrant et degré sortant de v .

Dans le cadre des graphes aléatoires, la variable X_{ij} désigne, pour tous sommets i et j , la variable indicatrice de la présence de l'arête entre i et j . On note alors $p_{ij} = \mathbb{P}(X_{ij} = 1) = \mathbb{E}(X_{ij})$ la probabilité d'apparition d'une arête entre i et j .

Organisation du mémoire

Le présent mémoire reprend les principaux travaux auxquels j'ai participé en les organisant autour de trois thèmes principaux.

La première partie est consacrée à la mise au point de modèles de graphes aléatoires. Le principe de tels modèles est de permettre de déduire les caractéristiques topologiques des réseaux réels en cherchant les différences entre les réseaux réels et les familles aléatoires de graphes. Afin de déterminer des caractéristiques topologiques de plus en plus fines, il faut être capable de générer des graphes aléatoires ayant des topologies aussi proches que possible des réseaux observés. Une seconde application, liée à l'utilisation de modèles de mélanges, est la classification des sommets d'un réseau en groupes ayant un comportement topologique similaire.

La seconde partie est dédiée à une caractéristique topologique particulière qui est le comptage et la répartition des petits sous-graphes. Plus précisément, elle traite de la recherche de motifs dans les réseaux, c'est-à-dire de petits sous-graphes dont le nombre est significativement plus élevé que dans un modèle aléatoire donné. L'hypothèse biologique sous-jacente à cette recherche est que si une telle structure apparaît plus souvent dans les réseaux réels que dans des réseaux similaires mais aléatoires, sa présence est due à un mécanisme de sélection positive au cours de l'évolution. Ces structures ont donc potentiellement un intérêt biologique. L'apport principal de mon travail est la définition et la mise au point de motifs locaux, c'est-à-dire la prise en compte de la répartition des petits sous-graphes et pas seulement de leur comptage.

La troisième partie s'intéresse à une structure particulière des graphes qui est le cycle. Elle se différencie des deux premières parties dans la mesure où elle ne fait pas appel aux probabilités ou aux statistiques. Elle regroupe une partie de mon travail de thèse et des travaux récents qui portent respectivement sur la largeur d'arborescence, c'est-à-dire un outil d'informatique théorique lié à la complexité des algorithmes sur certaines familles de graphes, et sur l'énumération de structures chimiques stables dans les réseaux métaboliques. Ces deux thèmes à priori très différents ont en commun le fait de voir leur complexité directement liée à l'organisation des cycles dans les graphes traités, et plus précisément à la recherche d'un ensemble transversal pour les cycles.

Quelques-unes de mes publications, ne s'inscrivant pas dans les trois thèmes précédents, ne sont pas développées dans ce mémoire. Certaines correspondant à des problèmes de théorie des graphes non motivés par l'analyse des réseaux biologiques. Elles traitent respectivement de coloration des arêtes d'un graphe dirigé [B3], d'existence de cordes dans les plus long cycles des graphes planaires 3-connexes [B5] et de performances en moyenne d'algorithmes correspondant au problème du *Vertex Cover* [B8,B18]. [B12] porte au contraire sur une procédure de classification des gènes dans les réseaux de régulations inférés par [41] et à l'étude de sa pertinence à l'aide d'une analyse des annotations *Gene Ontology* des gènes regroupés.

Première partie

Modèles aléatoires pour les réseaux

1. Caractéristiques topologiques des réseaux

Il est important pour un modèle aléatoire de respecter un maximum de caractéristiques topologiques de réseaux biologiques. Guimera *et al.* [51] suggèrent de comparer chaque réseau à une suite de modèles de complexité croissante. A chaque étape, il s'agit d'imaginer un modèle aléatoire respectant à la fois les propriétés du modèle précédent et les caractéristiques du réseau mises à jour lors de l'étape précédente. Les propriétés du réseau décrites à l'aide du nouveau modèle sont ainsi indépendantes de celles mises en lumière précédemment.

Certaines caractéristiques topologiques sont communes à tous les réseaux dit *réels*, c'est-à-dire issus essentiellement de la biologie, des sciences sociales ou de l'informatique [4]. Elles sont les suivantes, classées suivant ce qui me semble être l'ordre dans lequel elles doivent être prises en compte :

1. les réseaux réels ont une *composante connexe géante*. On entend par ce terme que si quelques sommets peuvent être isolés ou former de petites composantes connexes, l'essentiel du réseau est constitué d'une seule composante connexe.
2. les réseaux réels sont creux, c'est-à-dire qu'ils ont un nombre faible d'arêtes par rapport au nombre d'arêtes possibles. La croissance du nombre d'arêtes présentes se fait de façon linéaire en le nombre de sommets, alors que le nombre d'arêtes possibles est quadratique.
3. les réseaux contiennent des sommets de grand degré, communément appelés *hubs* par analogie avec le réseau de transport aérien. En d'autres termes, la distribution des degrés a une queue de distribution lourde.

Jeong *et al.* [63] décrivent la distribution des degrés dans des réseaux issus de sources diverses comme suivant la distribution d'une loi puissance. Autrement dit,

$$\mathbb{P}(d(v) = k) \sim k^{-\gamma}, \quad \gamma > 0$$

On parle alors de réseau sans échelle (*scale-free graph*). Cette assertion est communément admise, même si elle est de plus en plus critiquée, notamment car les réseaux observés sont des sous-réseaux des réseaux réels et qu'être sans échelle n'est pas une propriété stable par échantillonnage [103]. De plus, la loi puissance n'est en général une bonne approximation que pour les grands degrés. D'autres lois sur les degrés, en particulier les lois de mélange [33], permettent un meilleur ajustement sur les degrés faibles tout en maintenant l'existence de hubs.

La présence des *hubs* dans les réseaux a des conséquences topologiques fortes. La première est appelée effet petit monde et consiste en l'existence de courts chemins entre toute paire de sommets. La seconde est l'existence de nombreux chemins entre différentes parties du réseau, assurant une meilleure résistance du réseau aux attaques aléatoires. Par contre, ces réseaux sont vulnérables aux attaques ciblées sur les hubs.

Ces deux propriétés ont des interprétations évidentes en termes de réseaux sociaux ou de routage mais leur intérêt biologique est moins évident. Cependant, il a été démontré que la létalité de la suppression d'une protéine est considérablement plus forte si cette protéine est un hub d'un PPI [64].

- du fait de leur faible nombre d'arêtes, les réseaux sont de densité globale très faible. Malgré cela, ils sont localement denses, à savoir que le réseau réduit au voisinage d'un sommet est plus riche en arêtes que l'ensemble du réseau.

L'indice communément utilisé pour mettre en valeur ce phénomène est le *coefficient de clustering*. Celui-ci est défini comme la valeur moyenne de la densité des voisinages, à savoir

$$CC(G) = \frac{1}{n} \sum_{v \in V(G)} D(G[N(v)])$$

Le graphe de la Figure 2 montre un exemple de calcul de densité et de coefficient de clustering et illustre le fait que l'organisation en module mène à une densité locale moyenne supérieure à la densité globale.

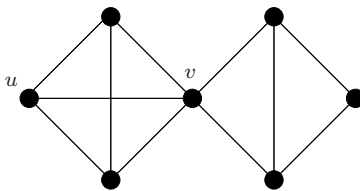


FIGURE 2. La densité de ce graphe est de $\frac{11}{22} = \frac{1}{2}$, mais le coefficient de clustering est de $\frac{172}{210} > \frac{4}{5}$.

- enfin, il existe dans la plupart des réseaux des structures sous-jacentes menant à des connexions plus probables entre certains sommets du réseau. Ces structures sont d'origines différentes suivant les cas : elles sont, par exemple, spatiales dans le cas d'un réseau de neurones où les connexions longue distance sont rares, ou fonctionnelles dans le cadre d'un réseau PPI. Cette structuration des réseaux en groupes de connectivité est au coeur des modèles de mélange détaillés dans les sections 3 et 4.

2. Modèles aléatoires existants

2.1. Modèles pas à pas vs modèles à tirage simultané

Les modèles existants de graphes aléatoires peuvent être classés en deux grandes familles, que j'appellerai respectivement les *modèles pas à pas* et les *modèles à tirage simultané* suivant le mode de tirage aléatoire des arêtes.

Les modèles pas à pas sont des modèles où les graphes aléatoires sont construits ou modifiés arête par arête. Il existe de nombreux modèles de ce type mais qui sont à ma connaissance des adaptations

diverses de trois modèles principaux que sont les modèles de Watts et Strogatz, le modèle d'attachement préférentiel et le modèle de *stub-rewiring*.

Watts-Strogatz : Ce modèle [107] consiste à considérer n sommets numérotés de 1 à n et à relier toute paire de sommets (i, j) vérifiant $|i - j| \leq k$. On obtient ainsi un graphe dont tous les sommets sont de degré $2k$.

Chacune des arêtes (i, j) , $i < j$, est alors redirigée avec probabilité β , la redirection consistant à la remplacer par une arête (i, j') où j' est choisi de façon uniforme parmi les sommets non voisins de i .

Ce modèle appliqué pour $n \gg k$ a pour avantage de créer des graphes de densité faible mais de densité locale forte. La distribution des degrés est cependant trop proche de la distribution uniforme.

Attachement préférentiel : Ce modèle [12], également appelé modèle de Barabasi-Albert, génère des graphes en introduisant des sommets un par un. A l'étape t , le nouveau sommet est connecté à l'un des sommets déjà présents. Ce sommet est choisi aléatoirement suivant une distribution de probabilités proportionnelle à la distribution des degrés après l'étape $t - 1$.

Chaque nouveau sommet ne créant qu'une seule arête, le modèle initial tel que présenté dans [12] ne permet pas de créer de cycles et ne génère donc que des arbres. Il a été adapté de plusieurs façons [39, 69, 71], afin de pouvoir générer des topologies plus complexes. Cependant, l'idée principale reste, à savoir que plus un sommet est de degré important, plus il est probable qu'il soit relié à tout nouveau sommet ou toute nouvelle arête. Ce processus, connu sous le nom de *rich get richer*, permet d'imiter la création de réseaux internet puisque les hyperliens d'une page nouvellement créée auront tendance à pointer vers des pages déjà hautement référencées.

Les modèles de cette famille génèrent des distributions des degrés proches des distributions observées mais la densité locale n'est pas respectée.

stub-rewiring : Ce modèle introduit par Milo *et al.* [80] est de loin le plus utilisé dans les applications en biologie. Il permet en effet de générer des graphes ayant la même distribution de degrés que le réseau observé.

La génération de tels graphes se fait en partant du réseau observé et en appliquant un grand nombre de fois l'opération consistant à choisir quatre sommets u_1, u_2, v_1 et v_2 tels que (u_1, v_1) et (u_2, v_2) sont des arêtes alors que (u_1, v_2) et (u_2, v_1) n'en sont pas. On supprime alors les arêtes (u_1, v_1) et (u_2, v_2) pour les remplacer par (u_1, v_2) et (u_2, v_1) .

Cette technique peut se généraliser sans mal aux graphes orientés en remplaçant $\overrightarrow{u_1v_1}$ et $\overrightarrow{u_2v_2}$ par $\overrightarrow{u_1v_2}$ et $\overrightarrow{u_2v_1}$. Elle permet également de prendre en compte la couleur des noeuds si nécessaire [114].

En raison de son caractère markovien, cette procédure permet de tirer de manière uniforme parmi les graphes ayant la même distribution de degrés que le réseau initial. Cette raison ainsi que son adaptabilité, la simplicité de son implémentation et le fait qu'elle génère des graphes de même distribution de degrés font de ce modèle l'un des plus utilisés pour comparer un réseau réel à de l'aléa. Il sert de référence dans les algorithmes les plus populaires de recherche de motifs [80, 108].

Cependant, cette méthode reste criticable du point de vue de la densité locale qu'elle néglige totalement. En effet, les arêtes du réseau initial, qui traduisent le plus souvent des relations locales, sont échangées au profit d'arêtes de plus longue portée, éliminant de fait la propriété du coefficient de clustering élevé.

De plus, elle ne permet pas de prendre pas en compte la création préférentielle de liens entre certaines classes de sommets. [8] montrent que cela conduit à une mauvaise modélisation aléatoire dans le cas des réseaux neuronaux.

Au contraire des modèles pas à pas, les modèles à tirage simultané sont définis par le fait que la totalité du graphe aléatoire est générée en même temps. Les arêtes ne sont donc plus dépendantes les unes des autres via une histoire évolutive du graphe. Afin de modéliser la dépendance entre arêtes, il sera donc nécessaire d'introduire des variables latentes.

Voici à nouveau une liste non exhaustive des principaux modèles à tirage simultané :

Erdős-Rényi : Le modèle d'Erdős-Rényi est le plus ancien et le plus simple des modèles de graphes aléatoires. Erdős et Rényi [42] l'ont utilisé pour montrer qu'il existe des graphes ayant un nombre chromatique supérieur à k tout en n'ayant pas de cycle de longueur inférieure à m , et ce pour tous entiers k et m . Pour ce faire, ils ont montré que l'ensemble des graphes vérifiant cette propriété est de mesure non nulle sous ce modèle. La nouveauté de cette approche probabiliste évitant la construction explicite de graphes et la simplicité du modèle ont popularisé ce dernier, au point que le terme *graphes aléatoires* est parfois utilisé dans la littérature sans autre précision pour y faire référence.

Le modèle d'Erdős-Rényi dépend d'un seul paramètre $p \in [0, 1]$ et suppose que les variables $(X_{ij})_{i,j \in V(G)}$ sont indépendantes et identiquement distribuées suivant une loi binomiale de paramètre p . En d'autres termes, toutes les arêtes sont considérées comme indépendantes et la probabilité d'apparition des arêtes est uniforme.

L'extrême simplicité de ce modèle en fait cependant la faiblesse dans la mesure où il ne permet pas de retranscrire la complexité de la structure des réseaux réels. En effet, le seul ajustement possible est de choisir une valeur de p telle que les graphes générés ont en moyenne la bonne densité. Le nombre moyen d'arêtes dans un graphe aléatoire de ce type étant de $\frac{n(n-1)}{2}p$ et le nombre d'arêtes dans les réseaux réels croissant linéairement en le nombre de sommets [29], cela revient à choisir p de la forme $p = \frac{c}{n}$.

Les graphes générés ont ainsi une densité satisfaisante et possèdent bien une composante géante pour $c > 1$ [20]. Cependant, ils sont homogènes et ne satisfont pas les autres critères requis. Ainsi, la distribution des degrés suit une loi de Poisson qui n'est suffisamment lourde ni pour les grands degrés ni pour les très petits degrés. Les graphes générés ne présentent en particulier pas de *hub*. De plus, l'espérance du coefficient de clustering est égale à p et tend donc vers 0 et aucune modularité n'apparaît grâce à ce modèle.

Je me référerai cependant à ce modèle dans le partie II car il est un bon outil pour tester la faisabilité d'une méthode en tant que modèle le plus simple à manipuler d'un point de vue mathématique.

Configuration model : Molloy et Reed [83] ont proposé un modèle qui permet de générer des graphes de distribution de degrés donnée avec un tirage simultané. Il consiste à faire sortir de tout sommet v un nombre d_v de demi-arêtes. Ces demi-arêtes sont alors regroupées par paires selon un tirage uniforme afin de former des arêtes. Le graphe ainsi obtenu est alors un graphe aléatoire ayant la distribution de degrés voulus, à ceci près qu'il peut y avoir plusieurs arêtes reliant le même couple de sommets. Cependant, pour un nombre de demi-arêtes croissant linéairement en n , il est aisé de montrer que ces phénomènes sont asymptotiquement de probabilité nulle.

Ce modèle a été abondamment utilisé et adapté [1, 28, 82, 84, 86]. Cependant, la modularité n'y est pas présente et la relation complexe de dépendance entre les arêtes ne permet pas l'étude de la loi de comptage d'un sous-graphe.

Expected Degree Distribution (EDD) : Le modèle de Molloy et Reed se révèle très difficile à manipuler d'un point de vue théorique car la contrainte de l'égalité stricte entre les degrés des graphes aléatoires et du graphe observé est trop forte. Afin de pouvoir mener des calculs, il est préférable de relâcher cette contrainte en n'imposant qu'une égalité des degrés en espérance.

Ceci permet, via la définition d'une probabilité p_{ij} de connexion pour tout couple de sommets i et j , de supposer les $(X_{ij})_{i,j \in V(G)}$ indépendants tout en préservant l'hétérogénéité de la distribution des degrés. Plus précisément, soit d_i et d_j les degrés respectifs de i et j dans le réseau réel. On pose alors

$$p_{ij} = P(X_{ij} = 1) = \frac{d_i d_j}{C}.$$

La constante C doit être choisie de façon à ce que chacune des probabilités de connexion soit inférieure ou égale à 1. En pratique, le choix $C = \sum_i d_i$ convient. Ce choix est le plus pertinent car il permet de générer des graphes dont la distribution des degrés est proche de celle du graphe réel. En effet, on a alors

$$\mathbb{E}\left(\sum_j X_{ij}\right) = d_i$$

ce qui revient à dire que chaque sommet a en moyenne son degré égal au degré observé.

Il est naturellement possible de modifier le modèle *EDD* en interdisant les auto-arêtes ou en lui faisant générer des graphes orientés. Dans ce dernier cas, la probabilité de connexion devient

$$P(X_{ij} = 1) = \frac{d_i^+ d_j^-}{C},$$

ce qui permet de générer des graphes ayant en moyenne les degrés entrants et sortants correspondants aux degrés observés.

Ce modèle est le plus simple des modèles à tirage simultané permettant de respecter l'hétérogénéité des degrés dans les réseaux réels et de calculer les premiers moments des comptages de motifs en respectant cette contrainte [B2].

Cependant, ce modèle a deux inconvénients majeurs. Le premier est de générer des graphes non localement denses et systématiquement assortatifs, c'est-à-dire pour lesquels les hubs ont une forte

tendance à être reliés les uns aux autres. Le second est que d'un point de vue pratique, la seule amélioration possible des performances est le regroupement des sommets ayant le même degré. Cependant, un graphe réel présentant en général plusieurs dizaines de degrés différents, les calculs réalisables en pratique sont réduits.

Stochastic Block Model (SBM) : Le modèle EDD permet d'introduire des comportements topologiques différents suivant les sommets en fonction d'un paramètre qui est le degré observé. Cependant, les classes correspondant à des comportements topologiques similaires ne correspondent généralement pas aux sommets de même degré.

Il est donc naturel d'introduire cette notion de classes topologiques à l'aide d'un modèle de mélange de graphes d'Erdős-Rényi. Ceci permet de modéliser des structures plus complexes de dépendances entre arêtes grâce à des variables latentes d'appartenance à des classes.

D'un point de vue formel, ce modèle est paramétré par un nombre Q de classes de sommets, un vecteur de répartition $\alpha \in \mathbb{R}^Q$ tel que $\sum_{1 \leq i \leq Q} \alpha_i = 1$ et une matrice Π de dimension $Q \times Q$ constituée de probabilités de connection.

Pour tout sommet i , un vecteur latent \mathbf{Z}_i est tiré suivant une loi multinomiale $\mathcal{M}(1, \alpha)$. Ainsi, tout sommet appartient à une unique classe qui est l'indice q tel que $Z_{iq} = 1$. Les arêtes sont ensuite tirées suivant des lois de Bernoulli indépendantes conditionnellement aux classes :

$$X_{ij} | Z_{iq} Z_{jl} = 1 \sim \mathcal{B}(\pi_{ql}),$$

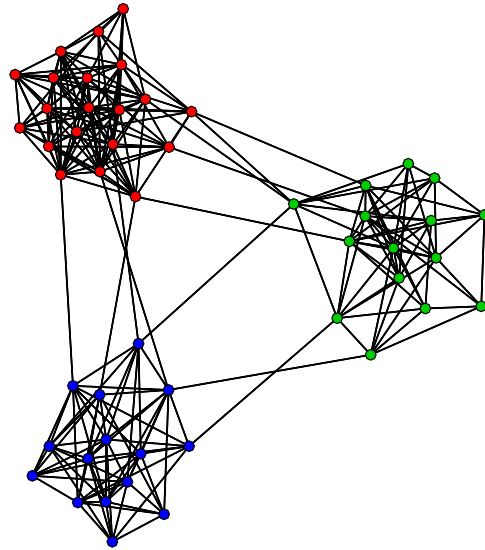


FIGURE 3. Exemple de graphe tiré suivant un modèle SBM à trois classes avec $\alpha = (1/3, 1/3, 1/3)$, $\pi_{i,i} = 0,5$ pour tout $1 \leq i \leq 3$ et $\pi_{i,j} = 0,02$ pour tout $i \neq j$.

Ce modèle, appelé *Stochastic Block Model* à la suite des travaux de Nowicki et Snijders [87], a été introduit en sciences sociales dès les années 80 [44, 46, 57] suite aux travaux de White *et al.* [109].

Sa souplesse en terme de schémas de connectivité permet de générer des graphes de structures très différentes, et notamment contenant des hubs [33]. L'estimation des vecteurs latents permet de plus de disposer d'un outil de classification des sommets d'un réseau suivant leur comportement topologique [56, 85].

Bollobas-Jensen-Riordan : Les auteurs de [21] ont proposé un modèle encore plus complet basé sur un noyau.

En résumé, il s'agit de considérer un espace mesurable \mathcal{S} , une mesure de probabilité μ sur cet espace et une fonction noyau κ mesurable de $\mathcal{S} \times \mathcal{S}$ dans $[0, 1]$. A chaque sommet i est alors associé un élément x_i de \mathcal{S} tiré suivant la loi μ et p_{ij} est défini comme égal à $\kappa(x_i, x_j)$.

Bollobás et ses co-auteurs démontrent alors que pour une large classe de fonctions noyau κ , les graphes obtenus ont une composante géante, un petit diamètre et une distribution des degrés correspondant à un mélange de loi de Poisson. De plus, suivant le choix de la mesure μ , ce mélange peut correspondre à une loi puissance.

Ce modèle est une généralisation du précédent dans la mesure où le choix $\mathcal{S} = \{1, \dots, Q\}$, $\mu(q) = \alpha_q$ pour tout $1 \leq q \leq Q$ et $\kappa(q, l) = \pi_{ql}$ correspond au modèle de mélange. Cependant, le problème de l'identifiabilité du modèle et l'estimation de κ n'est à ma connaissance pas traité. De plus, en sortant du cadre du mélange, on perd la notion de classes de sommets, qui est un élément important de l'interprétation des résultats.

Du point de vue de la loi des comptages des sous-graphes, les modèles pas à pas se révèlent impossibles à gérer de par le fait que les changements en terme de comptages impliqués par chaque étape de création ou de remplacement d'un sommet ou d'une arête ne sont pas maîtrisables. Les modèles à tirage simultané sont par contre plus abordables, ne serait-ce qu'au niveau du calcul des moments des comptages. Il est ainsi aisé de se convaincre que dans un modèle d'Erdős-Rényi de paramètre p , l'espérance du nombre T de triangles sera

$$\mathbb{E}(T) = \binom{n}{3} p^3$$

Le fil directeur de mon travail sur les modèles aléatoires étant la recherche de motifs dans les réseaux à l'aide des lois de comptage, la suite de ce travail n'abordera que des modèles à tirage simultané.

Dans la suite de ce chapitre, je présente les travaux auxquels j'ai participé concernant trois types de modèles simultanés. Le chapitre 3 concerne l'estimation des paramètres du modèle SBM dans un cadre bayésien. La partie 4 généralise ce modèle au cas de groupes pouvant avoir des sommets communs. Il est à noter que ces deux parties correspondent à des travaux en commun avec Pierre Latouche et Christophe Ambroise et ont fait l'objet de la thèse de Pierre [75]. Le dernier de ces travaux, présenté au chapitre 5, consiste à élaborer un modèle tirant parti de la structure bipartie sous-jacente de nombreux graphes réels.

3. Approche bayésienne du modèle de mélange

3.1. Contexte

Le modèle SBM permet de décrire des types de graphes très différents [33]. Pourtant, la simplicité de sa définition permet d'écrire la vraisemblance d'une observation. En effet, dans le cas de graphes non dirigés et sans auto-arêtes :

$$\begin{aligned}
 p(\mathbf{X} | \mathbf{\Pi}, \boldsymbol{\alpha}) &= \sum_{\mathbf{Z}} p(\mathbf{X} | \mathbf{Z}, \mathbf{\Pi}) p(\mathbf{Z} | \boldsymbol{\alpha}) \\
 &= \sum_{\mathbf{Z}} \prod_{i < j} p(X_{ij} | \mathbf{Z}_i, \mathbf{Z}_j, \mathbf{\Pi}) \prod_{i=1}^N \mathcal{M}(\mathbf{Z}_i; 1, \boldsymbol{\alpha}) \\
 &= \sum_{\mathbf{Z}} \prod_{i < j} \prod_{q, l} \left(\pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \right)^{Z_{iq} Z_{jl}} \prod_{i=1}^N \prod_{q=1}^Q \alpha_q^{Z_{iq}}.
 \end{aligned}$$

Le même calcul peut être mené pour les graphes dirigés et/ou avec auto-arête, en adaptant l'indice du premier produit. Dans tous les cas, on aboutit à une somme de Q^N termes qu'on ne sait pas optimiser directement. L'algorithme EM [35], classiquement utilisé pour résoudre ce type de problèmes dans les modèles de mélange, ne peut être utilisé directement dans ce cas. En effet, il nécessite une écriture sous la forme factorisée de $p(\mathbf{Z} | \mathbf{X}, \mathbf{\Pi}, \boldsymbol{\alpha})$, ce qui n'est pas possible ici [33].

Nowicki et Snijders [87] ont proposé une approche bayésienne dont l'estimation des hyper-paramètres est réalisée à l'aide d'échantillonnages de Gibbs. L'algorithme BLOCKS correspondant, disponible dans le package StoCNET [19], est cependant limité à des réseaux de moins de 200 noeuds. La méthode, également bayésienne, de Hofman et Wiggins [56] permet de traiter des réseaux beaucoup plus grands mais se limite au modèle d'affiliation, qui revient à imposer deux probabilités λ et ϵ telles que une $\mathbf{\Pi}$ est égale à λ sur la diagonale et à ϵ en-dehors. Daudin *et al.* [33] ont développé une méthode fréquentiste basée sur une approche variationnelle. Enfin, Vazquez [105] a développé un modèle de mélange basé sur la structure d'hypergraphe qu'il applique entre autres à des graphes pour en retrouver les groupes de sommets topologiquement similaires. Son approche nécessite cependant de faire une hypothèse d'indépendance entre les lignes de la matrice d'adjacence, ce qui est faux dans les cas des graphes.

Nous proposons d'adapter pour SBM l'approche variationnelle dans un cadre bayésien. Les détails des calculs sont développés dans [B9]. Le but est de développer un nouveau critère de choix de modèle. En effet, en raison du nombre de termes dans $p(\mathbf{X} | \boldsymbol{\alpha}, \mathbf{\Pi})$, les critères basés sur cette quantité tels que BIC ou AIC ne peuvent être utilisés. Daudin *et al.* [33] et Mariadassou *et al.* [76] utilisent un critère nommé ICL, introduit par Biernacki [15] dans le cadre des mélanges gaussiens. Il est basé sur une approximation asymptotique de $p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\alpha}, \mathbf{\Pi})$. Cependant, Biernacki *et al.* [16] a montré qu'en raison de son caractère asymptotique, ce critère a tendance à sous-estimer le nombre de classes pour les données de taille trop faible, ce qui est confirmé par des expériences menées dans le cas du modèle SBM [76].

3.2. Approche variationnelle bayésienne

Dans un cadre bayésien les paramètres α et $\mathbf{\Pi}$ de SBM deviennent des variables aléatoires, dont les distributions à priori sont choisies conjuguées aux distributions de SBM et non-informatives [61].

Ainsi, comme $p(\mathbf{Z}_i | \alpha)$ est une multinomiale, α est modélisé par une distribution de Dirichlet

$$p(\alpha | \mathbf{n}^0 = \{n_1^0, \dots, n_Q^0\}) = \text{Dir}(\alpha; \mathbf{n}^0),$$

où $n_q^0 = 1/2, \forall q$.

De même, comme $p(X_{ij} | \mathbf{Z}_i, \mathbf{Z}_j, \mathbf{\Pi})$ est une distribution de Bernoulli, la matrice de connectivité $\mathbf{\Pi}$ est modélisée à partir de lois Beta

$$p(\mathbf{\Pi} | \boldsymbol{\eta}^0 = (\eta_{ql}^0), \boldsymbol{\zeta}^0 = (\zeta_{ql}^0)) = \prod_{q \leq l} \text{Beta}(\pi_{ql}; \eta_{ql}^0, \zeta_{ql}^0),$$

avec $\eta_{ql}^0 = \zeta_{ql}^0 = 1/2, \forall q$.

Afin d'estimer les lois à posteriori, nous appliquons l'approche variationnelle [52], qui consiste à décomposer la vraisemblance des données observées en

$$\log p(\mathbf{X}) = \mathcal{L}(q(\mathbf{Z}, \alpha, \mathbf{\Pi})) + \text{KL}(q(\mathbf{Z}, \alpha, \mathbf{\Pi}) || p(\mathbf{Z}, \alpha, \mathbf{\Pi} | \mathbf{X})), \quad (1)$$

avec

$$\mathcal{L}(q) = \sum_{\mathbf{Z}} \int \int q(\mathbf{Z}, \alpha, \mathbf{\Pi}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \alpha, \mathbf{\Pi})}{q(\mathbf{Z}, \alpha, \mathbf{\Pi})} \right\} d\alpha d\mathbf{\Pi}, \quad (2)$$

et

$$\begin{aligned} \text{KL}(q(\mathbf{Z}, \alpha, \mathbf{\Pi}) || p(\mathbf{Z}, \alpha, \mathbf{\Pi} | \mathbf{X})) \\ = - \sum_{\mathbf{Z}} \int \int q(\mathbf{Z}, \alpha, \mathbf{\Pi}) \log \left\{ \frac{p(\mathbf{Z}, \alpha, \mathbf{\Pi} | \mathbf{X})}{q(\mathbf{Z}, \alpha, \mathbf{\Pi})} \right\} d\alpha d\mathbf{\Pi}. \end{aligned} \quad (3)$$

où KL désigne la distance de Kullback-Leibler.

Trouver la distribution $q(\mathbf{Z}, \alpha, \mathbf{\Pi})$ minimisant (3) est alors équivalent à maximiser la borne inférieure de la vraisemblance donnée par (2). Afin de pouvoir mener les calculs à bien, l'espace des distributions est restreint à celles pouvant se factoriser sous la forme

$$q(\mathbf{Z}, \alpha, \mathbf{\Pi}) = q(\alpha)q(\mathbf{\Pi})q(\mathbf{Z}).$$

avec

$$q(\mathbf{Z}) = \prod_{i=1}^N q(\mathbf{Z}_i) = \prod_{i=1}^N \mathcal{M}(\mathbf{Z}_i; 1, \boldsymbol{\tau}_i),$$

Les $\boldsymbol{\tau}_i$ correspondent alors aux probabilités d'appartenance à posteriori des sommets du réseau aux différentes classes.

Chacune des distributions $q(\boldsymbol{\alpha})$, $q(\boldsymbol{\Pi})$ et $q(\mathbf{Z}_i)$ pour $1 \leq i \leq N$ est alors optimisée à tour de rôle en considérant toutes les autres comme fixes. Les choix énoncés précédemment des formes initiales de ces distributions assurent que la forme de la distribution optimale trouvée à chaque étape est la même que celle de départ. La remise à jour des distributions revient donc uniquement à une remise à jour des paramètres.

Ce procédé est appelé *Variational Bayes EM* par analogie avec l'algorithme EM classique, la mise à jour des $(\boldsymbol{\tau}_i)_{1 \leq i \leq N}$ correspondant à l'étape E alors que celles des paramètres des lois de $\boldsymbol{\alpha}$ et $\boldsymbol{\Pi}$ correspondent à l'étape M.

Optimiser l'un de ces paramètres en fixant les autres revient à calculer l'espérance de la vraisemblance $p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi})$ par rapport à toutes les variables sauf celle gouvernée par le paramètre d'intérêt [17, 75], puis à renormaliser. Ceci amène à des équations de mise à jour qui sont détaillées dans [B9].

L'algorithme complet permettant d'estimer les paramètres du modèle pour une valeur fixée de Q commence par la création d'une partition initiale des sommets suivant un algorithme de classification hiérarchique. La distance prise en compte est $d(i, j) = \sum_{k=1}^N (X_{ik} - X_{jk})^2$, ce qui consiste à considérer deux sommets comme d'autant plus proches que leurs voisinages se ressemblent. L'algorithme *Variational Bayes EM* est ensuite appliqué jusqu'à ce que la croissance de $\mathcal{L}(q)$ entre deux cycles soit inférieure à un seuil prédéfini. Le coût en terme de temps de calcul de cet algorithme est en $\mathcal{O}(Q^2 N^2)$, ce qui est du même ordre de grandeur que le coût de l'approche fréquentiste de [33].

3.3. Un critère de choix de modèle

L'algorithme du paragraphe précédent permet de résoudre le problème de l'estimation pour un nombre de classes Q fixé. Afin de déterminer le nombre de classes optimal, nous proposons de chercher la valeur Q^* pour laquelle la log-vraisemblance marginale $\log p(\mathbf{X} | Q)$ est maximale. Celle-ci ne peut s'optimiser directement car elle s'écrit

$$\log p(\mathbf{X} | Q) = \log \left\{ \sum_{\mathbf{Z}} \int \int p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\Pi} | Q) d\boldsymbol{\alpha} d\boldsymbol{\Pi} \right\}.$$

Cependant, l'égalité (1) assure que $\log p(\mathbf{X} | Q)$ est minorée par la fonction $\mathcal{L}(q(\cdot))$. Cette fonction est maximisée en $q(\cdot)$ par l'algorithme *Variational Bayes EM* et la différence entre la valeur obtenue et la vraisemblance marginale est la distance de Kullback-Leibler donnée par (3). Il n'y a pas a priori de raison pour que cette distance soit faible ou indépendante de Q , mais nous proposons de faire cette approximation. Ceci mène à l'introduction du critère non asymptotique ILvb correspondant à la valeur de $\mathcal{L}(q(\cdot))$ après maximisation :

$$IL_{vb} = \log \left\{ \frac{\Gamma(\sum_{q=1}^Q n_q^0) \prod_{q=1}^Q \Gamma(n_q)}{\Gamma(\sum_{q=1}^Q n_q) \prod_{q=1}^Q \Gamma(n_q^0)} \right\} + \sum_{q \leq l}^Q \log \left\{ \frac{\Gamma(\eta_{ql}^0 + \zeta_{ql}^0) \Gamma(\eta_{ql}) \Gamma(\zeta_{ql})}{\Gamma(\eta_{ql} + \zeta_{ql}) \Gamma(\eta_{ql}^0) \Gamma(\zeta_{ql}^0)} \right\} - \sum_{i=1}^N \sum_{q=1}^Q \tau_{iq} \log \tau_{iq}, \quad (4)$$

où τ_{iq} est la probabilité à posteriori pour i d'appartenir à la classe q et $(n_q)_q$, $(\eta_{ql})_{ql}$, $(\zeta_{ql})_{ql}$ sont les hyper-paramètres remis à jour.

En pratique, l'algorithme du *Variational Bayes EM* est exécuté pour toutes les valeurs de Q d'un intervalle d'intérêt et Q^* est déterminé à l'aide du critère ILvb.

3.4. Expériences comparatives

Des expériences de comparaison ont été menées entre le critère ILvb introduit ci-dessus, le critère ICL de Daudin *et al.* [33] basé sur une estimation fréquentiste et le critère VBMOD de Hofman et Wiggins [56] développé dans le cadre des modèles d'affiliation.

3.4.1. Données d'affiliation

Le premier type de données auxquels les trois algorithmes d'estimation ont été appliqués sont des graphes d'affiliation, c'est-à-dire des graphes générés par un modèle *SBM* dont la matrice est égale à λ sur sa diagonale et à ϵ ailleurs :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon \\ \epsilon & \lambda & & \vdots \\ \vdots & & \ddots & \epsilon \\ \epsilon & \dots & \epsilon & \lambda \end{pmatrix}.$$

Pour tout $Q \in \{3 \dots 7\}$, 100 graphes de 50 sommets ont été générés suivant ce modèle avec $\alpha_1 = \dots = \alpha_Q = \frac{1}{Q}$, $\lambda = 0.9$ et $\epsilon = 0.1$. Pour chacun de ces graphes, cinq initialisations différentes ont été déterminées et les trois méthodes ont été appliquées. La meilleure des cinq estimations pour chaque méthode a ensuite été retenue suivant le critère correspondant (VBMOD, ICL ou ILvb).

La Table 1 montre les résultats obtenus. On y voit que les critères ont tendance à sous-estimer le nombre de classes quand celui-ci grandit. Ce phénomène s'explique par la petite taille des classes pour un graphe à 50 noeuds et 6 ou 7 classes. Le meilleur comportement de VBMOD s'explique par le fait qu'il est destiné aux graphes d'affiliation et estime directement les deux paramètres λ et ϵ quand les deux autres algorithmes estiment une matrice $Q \times Q$. Cependant, on observe que le phénomène de sous-estimation des classes est moins prononcé pour ILvb, en raison de son caractère non asymptotique.

3.4.2. Données d'affiliation avec hubs

Afin de complexifier la structure des graphes, la procédure précédente a été répétée pour des graphes générés à l'aide de la matrice $\mathbf{\Pi}$ suivante :

$$\mathbf{\Pi} = \begin{pmatrix} \lambda & \epsilon & \dots & \epsilon & \lambda \\ \epsilon & \lambda & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \lambda & \dots & \dots & \dots & \lambda \end{pmatrix},$$

		2	3	4	5	6	7
$Q_{True} \setminus Q_{VBMOD}$	3	0	100	0	0	0	0
	4	0	0	100	0	0	0
	5	0	0	0	100	0	0
	6	0	0	0	0	97	3
	7	0	0	0	2	14	84
		2	3	4	5	6	7
$Q_{True} \setminus Q_{ICL}$	3	0	100	0	0	0	0
	4	0	0	100	0	0	0
	5	0	0	23	77	0	0
	6	0	1	28	59	12	0
	7	0	8	49	42	1	0
		2	3	4	5	6	7
$Q_{True} \setminus Q_{ILvb}$	3	0	100	0	0	0	0
	4	0	0	100	0	0	0
	5	0	0	0	99	1	0
	6	0	0	4	23	73	0
	7	0	2	14	44	27	13

TABLE 1

Nombre de classes sélectionnés par les critères *VBMOD*, *ICL* et *ILvb* pour des réseaux d'affiliation. 100 graphes ont été générés pour $Q_{True} \in \{3, \dots, 7\}$, avec $\lambda = 0.9$ et $\epsilon = 0.1$.

Cela revient à considérer un modèle d'affiliation à $Q - 1$ classes auquel on ajoute une classe de sommets ayant une forte probabilité d'être connectés à tout le monde, c'est-à-dire de hubs.

La Table 1 présente les résultats obtenus. On y observe que *VBMOD* s'adapte mal à cette complexification puisque la classe de hubs est intégrée aux autres classes dans environ 9 cas sur 10. Les critères *ICL* et *ILvb* retrouvent parfaitement le vrai nombre de classes quand celui-ci est suffisamment faible mais le sous-estiment lorsque le nombre de sommets par classe devient trop faible. Cependant, ce phénomène de sous-estimation est nettement plus faible pour *ILvb*.

3.4.3. Réseau métabolique d'*Escherichia coli*

Pour finir, nous avons considéré le réseau métabolique d'*Escherichia coli* [73] dont les sommets représentent les réactions et où deux réactions sont connectées si un produit de l'une est un substrat de l'autre. Ce graphe de 605 sommets et 1782 arêtes avait déjà été étudié à l'aide du critère *ICL* [33]. Nous lui avons appliqué à nouveau les critères *VBMOD*, *ICL* et *ILvb*, en ayant estimé les paramètres par les trois algorithmes associés pour $Q \in \{1 \dots 40\}$. Les critères atteignent leurs maxima respectifs pour $Q_{VBMOD} = 14$, $Q_{ICL} = 21$ et $Q_{ILvb} = 22$.

L'estimation plus basse de *VBMOD* s'explique par le fait que le modèle d'affiliation ne permet pas de retrouver certains types de structures. En effet, considérons la matrice d'adjacence du réseau réordonné suivant les classes estimées par notre procédure d'estimation, présentée en Figure 4. On y voit que la structure dominante est la structure d'affiliation avec des arêtes nombreuses à l'intérieur des classes et rares entre elles. Cependant, comme observé par Daudin *et al.* [33], certains couples de classes, par

		2	3	4	5	6	7
$Q_{True} \setminus Q_{VBMOD}$	3	95	0	3	0	0	2
	4	1	95	4	0	0	0
	5	0	0	94	6	0	0
	6	0	0	1	83	16	0
	7	0	0	2	15	78	5
		2	3	4	5	6	7
$Q_{True} \setminus Q_{ICL}$	3	0	100	0	0	0	0
	4	0	0	100	0	0	0
	5	0	0	12	88	0	0
	6	0	0	19	59	22	0
	7	0	3	29	56	12	0
		2	3	4	5	6	7
$Q_{True} \setminus Q_{ILvb}$	3	0	100	0	0	0	0
	4	0	0	100	0	0	0
	5	0	0	2	98	0	0
	6	0	0	1	29	70	0
	7	0	0	3	34	45	18

TABLE 2

Nombre de classes sélectionnées par les critères *VBMOD*, *ICL* et *ILvb* pour des réseaux d'affiliation avec hubs. 100 graphes ont été générés pour $Q_{True} \in \{3, \dots, 7\}$, avec $\lambda = 0.9$ et $\epsilon = 0.1$.

exemple les classes 1 et 17, présentent une connectivité plus forte entre elles. Ces classes sont distinctes au sens du modèle SBM car leurs relations aux autres classes sont différentes mais sont regroupées dans le modèle d'affiliation en raison de leur grand nombre de liens.

Les résultats obtenus en utilisant les critères *ICL* et *ILvb* sont comparables, aussi bien au niveau du nombre de classes qu'au niveau des sommets qu'elles contiennent.

4. Un modèle de mélange à classes chevauchantes

4.1. Contexte

La classification issue du modèle de mélange considéré au chapitre précédent possède une lacune vis-à-vis des réseaux observés, à savoir l'impossibilité pour un sommet d'être classé simultanément dans plusieurs classes. Il existe cependant des protéines, appelées *moonlighting proteins*, qui participent à plusieurs fonctions au sein de la cellule [60]. De même, une classification pertinente en sciences sociales doit pouvoir permettre à des sujets d'appartenir à plusieurs groupes d'intérêt simultanément [90]. Il est donc naturel de chercher à développer un modèle similaire à SBM autorisant de telles classifications.

Une approche algorithmique de ce problème a été proposée par Palla *et al.* [90] et est disponible dans le logiciel CFinder [91]. Elle consiste à fixer un entier k correspondant à la granularité de l'algorithme. Toutes les cliques de taille k du réseau sont ensuite énumérées et deux d'entre elles sont reliées si elles partagent $k-1$ sommets. Les composantes connexes de ce graphe des cliques définissent alors des communautés. Une granularité trop forte entraîne des communautés très homogènes mais petites et couvrant une faible partie

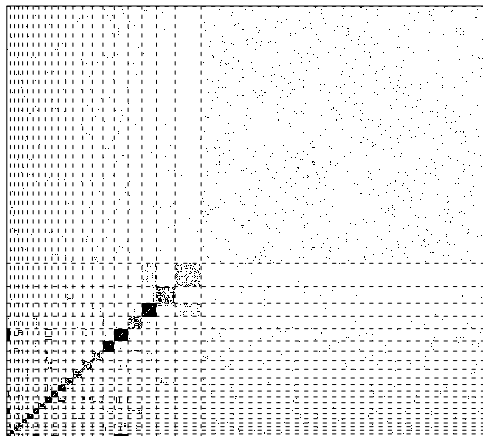


FIGURE 4. Dot-plot de la matrice d'adjacence du réseau métabolique d'*Escherichia coli*, une fois reordonnée suivant les 22 classes sélectionnées par *ILvb*.

du graphe alors qu'une granularité trop faible renvoie des communautés correspondant essentiellement aux composantes connexes du réseau. La granularité conseillée par les auteurs est alors la plus petite ne faisant pas apparaître de communauté géante.

Airoldi *et al.* [3] ont proposé le modèle *Mixed Membership Stochastic Block model* (MMSB) et l'ont appliqué aux réseaux d'interactions entre protéines [2]. Ce modèle est semblable à *SBM* mis à part que la classe de chaque sommet change suivant le sommet avec lequel on cherche à le relier. Plus précisément, un vecteur de poids π_i est tiré suivant une loi de Dirichlet pour chaque sommet i . Pour tout i et tout j , un vecteur $\mathbf{Z}_{i \rightarrow j}$ est alors tiré suivant une loi multinomiale $\mathcal{M}(1, \pi_i)$. Il représente la classe à laquelle appartient i dans le cadre de sa relation avec j . La probabilité de l'arête entre i et j est alors de la forme $\mathbf{Z}_{i \rightarrow j}^\top \mathbf{B} \mathbf{Z}_{i \rightarrow j}$, où \mathbf{B} est une matrice de connectivité semblable à la matrice $\mathbf{\Pi}$ du modèle *SBM*. Le vecteur de classification du sommet i est alors le vecteur π_i , ce qui permet à i d'interagir au sein de différentes classes via la présence de plusieurs coefficients non nuls au sein de ce vecteur. Cependant, la probabilité d'apparition de l'arête (i, j) est indépendante de l'interaction de ces sommets avec le reste du graphe. De plus, les vecteurs π_i que nous avons observé en pratique après convergence de l'algorithme d'estimation étaient composés de 0 et d'un 1, ce qui élimine de fait l'appartenance multiple.

Fu et Banerjee[47] ont proposé un modèle à classes chevauchantes adaptées à des matrices de données dont les lignes et les colonnes sont indépendantes. Il ne peut donc pas s'appliquer directement aux données issues d'un réseau. Cependant, les coordonnées des vecteurs latents \mathbf{Z}_i sont tirées suivant des lois de Bernoulli indépendantes plutôt que suivant une multinomiale, ce qui permet d'obtenir des chevauchements. Nous avons repris cette idée pour construire un modèle de graphes aléatoires à classes chevauchantes nommé *Overlapping Stochastic Block Model* (OSBM). Ce modèle est décrit et étudié dans la suite de cette section. Les démonstrations des propriétés citées sont détaillées dans [B10].

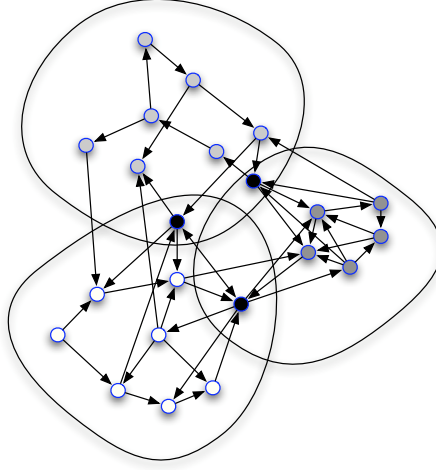


FIGURE 5. Exemple de réseau à classes chevauchantes.

4.2. Le modèle OSBM

Nous nous plaçons pour décrire le modèle OSBM dans le cadre des graphes orientés et sans auto-arête. La théorie décrite ci-dessous pourrait cependant être réécrite dans le cadre non-orienté et si besoin avec auto-arêtes.

Soit Q le nombre de classes envisagé. Un vecteur latent \mathbf{Z}_i composé de Q variables booléennes est associé à tout sommet i , comme dans les cas du modèle SBM. Cependant, afin de permettre l'appartenance multiple à des classes, il est tiré suivant une loi de Bernoulli multivariée :

$$\mathbf{Z}_i \sim \prod_{q=1}^Q \mathcal{B}(Z_{iq}; \alpha_q) = \prod_{q=1}^Q \alpha_q^{Z_{iq}} (1 - \alpha_q)^{1-Z_{iq}}. \quad (5)$$

Il est important de noter que sous ce modèle, le vecteur \mathbf{Z}_i peut être nul, ce qui veut dire que le sommet i peut n'appartenir à aucune classe. Ceci représente un avantage au niveau de l'interprétation car, dans la plupart des cas, une estimation de classes d'un modèle SBM sur un réseau réel crée une classe de sommets n'ayant d'autre point commun que leur faible connexion au reste du réseau. Ce phénomène est illustré par la plus grande classe de la figure 4.

Les lois des arêtes connaissant les classes sont alors données par

$$X_{ij} | \mathbf{Z}_i, \mathbf{Z}_j \sim \mathcal{B}(X_{ij}; g(a_{\mathbf{Z}_i, \mathbf{Z}_j})) = e^{X_{ij} a_{\mathbf{Z}_i, \mathbf{Z}_j}} g(-a_{\mathbf{Z}_i, \mathbf{Z}_j}),$$

avec g la fonction logistique définie par $g(x) = (1 + e^{-x})^{-1}$ et

$$a_{\mathbf{Z}_i, \mathbf{Z}_j} = \mathbf{Z}_i^\top \mathbf{W} \mathbf{Z}_j + \mathbf{Z}_i^\top \mathbf{U} + \mathbf{V}^\top \mathbf{Z}_j + W^*, \quad (6)$$

La matrice \mathbf{W} décrit les interactions des sommets suivant leurs classes, au même titre que la matrice $\mathbf{\Pi}$ du modèle SBM. Le vecteur U modélise la capacité générale des classes à émettre des arêtes, le vecteur V celle à en recevoir. Enfin, le terme W^* permet d'adapter le nombre moyen d'arêtes, et sert en pratique à modéliser la faible densité des réseaux réels.

Afin de simplifier les notations, considérons le vecteur $\tilde{\mathbf{Z}}_i = \begin{pmatrix} \mathbf{z}_i \\ 1 \end{pmatrix}$ et la matrice

$$\tilde{\mathbf{W}} = \begin{pmatrix} \mathbf{W} & \mathbf{U} \\ \mathbf{V}^\top & W^* \end{pmatrix}.$$

Le coefficient $a_{\mathbf{z}_i, \mathbf{z}_j}$ peut alors s'écrire

$$a_{\mathbf{z}_i, \mathbf{z}_j} = \tilde{\mathbf{Z}}_i^\top \tilde{\mathbf{W}} \tilde{\mathbf{Z}}_j. \quad (7)$$

De plus, les $\tilde{\mathbf{Z}}_i$ sont toujours tirés suivant des lois de Bernoulli multivariées en posant $\alpha_{Q+1} = 1$.

4.3. Identifiabilité

Avant de chercher à estimer les paramètres du modèle, il faut se poser la question de l'identifiabilité, c'est-à-dire de l'unicité du jeu de paramètres générant une distribution de graphes aléatoires données. Cette question a été résolue dans le cas de SBM par Allman *et al.* [5], qui ont montré que mis à part sur un ensemble de mesure nulle de l'espace des paramètres, SBM est un modèle identifiable aux permutations près. SBM est alors dit *génériquement identifiable* aux permutations près.

Nous avons démontré un théorème du même ordre pour le modèle OSBM, un autre type d'opération laissant la distribution invariante devant cependant être introduit.

4.3.1. Correspondance avec SBM

Considérons deux sommets i et j tels que $\mathbf{z}_i = \mathbf{z}_j$. Alors i et j ont la même probabilité de se connecter à tout autre sommet k . Il existe donc une fonction associant naturellement une distribution de type SBM à 2^Q classes à toute distribution de type OSBM.

Formellement, soit Θ_{OSBM} l'espace des paramètres du modèle OSBM à Q classes :

$$\Theta_{OSBM} = \{(\boldsymbol{\alpha}, \tilde{\mathbf{W}}) \in [0, 1]^Q \times \mathbb{R}^{(Q+1)^2}\},$$

et Θ_{SBM} celui du modèle SBM à 2^Q classes :

$$\Theta_{SBM} = \{(\boldsymbol{\gamma}, \mathbf{\Pi}) \in [0, 1]^{2^Q} \times [0, 1]^{2^{2Q}}, \sum_{\mathbf{c} \in \mathcal{C}} \gamma_{\mathbf{c}} = 1\}.$$

Considérons la fonction

$$\phi : \begin{array}{l} \Theta_{OSBM} \rightarrow \Theta_{SBM} \\ (\boldsymbol{\alpha}, \tilde{\mathbf{W}}) \rightarrow (\boldsymbol{\gamma}, \mathbf{\Pi}) \end{array},$$

avec

$$\gamma_{\mathbf{C}} = \prod_{q=1}^Q \alpha_q^{C_q} (1 - \alpha_q)^{1-C_q}, \forall \mathbf{C} \in \{0, 1\}^Q,$$

et

$$\Pi_{\mathbf{C}, \mathbf{D}} = g(\mathbf{C}^\top \mathbf{W} \mathbf{D} + \mathbf{C}^\top \mathbf{U} + \mathbf{V}^\top \mathbf{D} + W^*), \forall (\mathbf{C}, \mathbf{D}) \in \{0, 1\}^Q \times \{0, 1\}^Q.$$

Un paramètre $\theta \in \Theta_{OSBM}$ génère alors la même distribution sur l'ensemble des graphes à n sommets que le paramètre $\phi(\theta) \in \Theta_{SBM}$. Par conséquent, θ et θ' engendrent la même distribution si et seulement si $\phi(\theta)$ et $\phi(\theta')$ engendrent la même distribution. Or, Allman *et al.* [5] ont démontré le théorème suivant

Théorème 1. *Il existe un sous-espace $\Theta_{SBM}^{bad} \subset \Theta_{SBM}$ de mesure nulle tel que tous $(\gamma, \mathbf{\Pi})$ et $(\gamma', \mathbf{\Pi}')$ n'appartenant pas à Θ_{SBM}^{bad} génèrent la même loi de graphes aléatoires si et seulement si il existe P_ν tel que $(\gamma', \mathbf{\Pi}') = P_\nu((\gamma, \mathbf{\Pi}))$, avec :*

- ν est une permutation de $\{0, 1\}^Q$,
- $\gamma'_{\mathbf{C}} = \gamma_{\nu(\mathbf{C})}$, $\forall \mathbf{C} \in \{0, 1\}^Q$,
- $\mathbf{\Pi}'_{\mathbf{C}, \mathbf{D}} = \Pi_{\nu(\mathbf{C}), \nu(\mathbf{D})}$, $\forall (\mathbf{C}, \mathbf{D}) \in \{0, 1\}^Q \times \{0, 1\}^Q$.

Par conséquent, il suffit pour étudier l'identifiabilité du modèle OSBM de caractériser les paramètres pour lesquels $\phi(\theta') = P_\nu(\phi(\theta))$ pour une permutation ν de $\{0, 1\}^Q$.

4.3.2. Permutations et inversions

Comme dans le cas de SBM, une permutation des classes du modèle OSBM laisse la distribution induite inchangée. En effet, soit σ une permutation de $\{1, \dots, Q\}$. On définit $(\alpha', \tilde{\mathbf{W}}') = P_\sigma(\alpha, \tilde{\mathbf{W}})$ par :

$$\alpha'_q = \alpha_{\sigma(q)}, \forall q \in \{1, \dots, Q\},$$

et

$$\tilde{\mathbf{W}}'_{q,l} = \tilde{\mathbf{W}}_{\sigma(q), \sigma(l)}, \forall (q, l) \in \{1, \dots, Q+1\}^2.$$

Soit ν la permutation de $\{0, 1\}^Q$ définie par

$$\nu(\mathbf{C}) = (C_{\sigma(1)}, \dots, C_{\sigma(Q)}), \forall \mathbf{C} \in \{0, 1\}^Q.$$

Alors, pour tout $\theta \in \Theta_{OSBM}$, $\phi(P_\sigma(\theta)) = P_\nu(\phi(\theta))$.

Il existe une autre opération qui ne modifie pas la distribution. En effet, chaque coordonnée des vecteurs \mathbf{Z}_i partage les sommets en deux sous-ensembles suivant qu'elle vaut 0 ou 1. Remplacer tous les 0 par des 1 et inversement ne change pas cette information et il est possible de reparamétriser la matrice $\tilde{\mathbf{W}}$ afin que les probabilités de connexion ne changent pas.

Soit $\mathbf{A} \in \{0, 1\}^Q$ et $\mathbf{M}_{\mathbf{A}}$ la matrice définie par

$$\mathbf{M}_{\mathbf{A}} = \begin{pmatrix} I - 2diag(\mathbf{A}) & \mathbf{A} \\ 0 \dots 0 & 1 \end{pmatrix},$$

où $\text{diag}(\mathbf{A})$ désigne la matrice diagonale de taille $Q \times Q$ dont la diagonale est le vecteur \mathbf{A} .

L'inversion suivant A , notée $I_{\mathbf{A}}$, est le reparamétrage défini par

$$I_{\mathbf{A}} : \begin{array}{l} \Theta_{OSBM} \rightarrow \Theta_{OSBM} \\ (\boldsymbol{\alpha}, \tilde{\mathbf{W}}) \rightarrow (\boldsymbol{\alpha}', \tilde{\mathbf{W}}') \end{array},$$

où

$$\alpha'_j = \begin{cases} 1 - \alpha_j & \text{si } A_j = 1 \\ \alpha_j & \text{sinon} \end{cases}, \quad \forall j \in \{1, \dots, Q\},$$

et

$$\tilde{\mathbf{W}}' = \mathbf{M}_{\mathbf{A}}^{\top} \tilde{\mathbf{W}} \mathbf{M}_{\mathbf{A}}.$$

Soit ν la permutation de $\{0, 1\}^Q$ définie par

$$\forall \mathbf{C} \in \{0, 1\}^Q, \nu(\mathbf{C})_i = \begin{cases} 1 - C_i & \text{si } A_i = 1 \\ C_i & \text{sinon} \end{cases}.$$

En d'autres termes, ν inverse les coordonnées égales à 1 dans le vecteur \mathbf{A} . On peut alors montrer que, pour tout $\boldsymbol{\theta} \in \Theta_{OSBM}$, $\phi(I_{\mathbf{A}}(\boldsymbol{\theta})) = P_{\nu}(\phi(\boldsymbol{\theta}))$.

4.3.3. Identifiabilité générique

On définit dans l'ensemble Θ_{OSBM} la relation d'équivalence

$$\boldsymbol{\theta} \sim \boldsymbol{\theta}' \quad \text{si } \exists \sigma, \mathbf{A} \quad | \quad \boldsymbol{\theta}' = I_{\mathbf{A}}(P_{\sigma}(\boldsymbol{\theta})).$$

Il s'agit bien d'une relation d'équivalence car $I_{\mathbf{A}} \circ P_{\sigma} = P_{\sigma} \circ I_{\sigma^{-1}(\mathbf{A})}$.

D'après le paragraphe précédent, deux paramètres équivalents génèrent la même loi de graphes aléatoires. Nous avons démontré qu'à part sur un ensemble de mesure nulle, les classes d'équivalence sont en fait en bijection avec les lois générées.

Théorème 2. Pour tout $\boldsymbol{\alpha} \in]0, 1[^Q$, soit $\beta \in \mathbb{R}^Q$ le vecteur défini par $\beta_k = -\ln(\frac{\alpha_k}{1-\alpha_k})$, pour tout k .

Soit Θ_{OSBM}^{bad} l'ensemble des paramètres $(\boldsymbol{\alpha}, \tilde{\mathbf{W}})$ vérifiant au moins une des conditions suivantes :

- il existe $1 \leq k \leq Q$ tel que $\alpha_k = 0$ ou $\alpha_k = 1$ ou $\alpha_k = \frac{1}{2}$,
- il existe $1 \leq k, l \leq Q$ tels que $\alpha_k = \alpha_l$,
- il existe $\mathbf{C}, \mathbf{D} \in \{0, 1\}^Q \times \{0, 1\}^Q$ tels que $\sum_k \beta_k C_k = \sum_k \beta_k D_k$.

Alors Θ_{OSBM}^{bad} est un sous-ensemble de Θ_{OSBM} de mesure nulle et

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in (\Theta_{OSBM} \setminus \Theta_{OSBM}^{bad})^2, \quad \phi(\boldsymbol{\theta}) = \phi(\boldsymbol{\theta}') \Leftrightarrow \boldsymbol{\theta} \sim \boldsymbol{\theta}'.$$

En d'autres termes, le modèle OSBM est génériquement identifiable aux permutations et inversions près.

De plus, chaque classe d'équivalence contient un paramètre $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \tilde{\mathbf{W}})$ tel que $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_Q \leq \frac{1}{2}$. Si toutes les coordonnées du vecteur $\boldsymbol{\alpha}$ de ce paramètre sont distinctes et strictement inférieures à $\frac{1}{2}$, il est de plus unique. Ce paramètre particulier est celui qui sera rendu par les algorithmes d'estimation car il est celui qui s'interprète le plus facilement.

4.4. Estimation des paramètres

Comme dans le cas du modèle SBM, les probabilités $p(\mathbf{Z} | \boldsymbol{\alpha})$ et $p(\mathbf{X} | \mathbf{Z}, \tilde{\mathbf{W}})$ s'écrivent facilement, ce qui permet d'écrire la vraisemblance sous la forme

$$\begin{aligned} p(\mathbf{X} | \tilde{\mathbf{W}}, \boldsymbol{\alpha}) &= \sum_{\mathbf{Z}} p(\mathbf{X} | \mathbf{Z}, \tilde{\mathbf{W}}) p(\mathbf{Z} | \boldsymbol{\alpha}) \\ &= \sum_{\mathbf{Z}} \prod_{i \neq j}^N e^{X_{ij} a_{\mathbf{z}_i, \mathbf{z}_j}} g(-a_{\mathbf{z}_i, \mathbf{z}_j}) \prod_{i=1}^N \prod_{q=1}^Q \alpha_q^{Z_{iq}} (1 - \alpha_q)^{1 - Z_{iq}}. \end{aligned}$$

Cette somme contenant 2^{nQ} ne peut être maximisée directement. La démarche adoptée est alors la même que dans le cas de SBM, c'est-à-dire un algorithme de type EM appliqué à une approximation variationnelle. Pour ce faire, la log-vraisemblance est décomposée en

$$\ln p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}) + \text{KL}(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})), \quad (8)$$

avec

$$\mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\alpha}, \tilde{\mathbf{W}})}{q(\mathbf{Z})} \right\}, \quad (9)$$

et

$$\text{KL}(q(\mathbf{Z}) || p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\alpha}, \tilde{\mathbf{W}})}{q(\mathbf{Z})} \right\}. \quad (10)$$

La borne inférieure \mathcal{L} est alors maximisée par rapport à q , $\boldsymbol{\alpha}$ et $\tilde{\mathbf{W}}$, en se restreignant aux distributions factorisées de la forme

$$q(\mathbf{Z}) = \prod_{i=1}^n q(\mathbf{Z}_i) \quad (11)$$

avec

$$q(\mathbf{Z}_i) = \prod_{q=1}^Q \mathcal{B}(Z_{iq}; \tau_{iq}) = \prod_{q=1}^Q \tau_{iq}^{Z_{iq}} (1 - \tau_{iq})^{1 - Z_{iq}}. \quad (12)$$

L'interprétation des τ_{iq} reste la même que dans le cas de SBM, à savoir qu'il s'agit de la probabilité à posteriori pour le sommet i d'appartenir à la classe q .

On peut alors montrer que

$$\begin{aligned} \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}) &= \sum_{i \neq j}^N \left\{ X_{ij} \tilde{\boldsymbol{\tau}}_i^\top \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j} [\ln g(-a_{ij})] \right\} \\ &\quad + \sum_{i=1}^N \sum_{q=1}^Q \{ \tau_{iq} \ln \alpha_q + (1 - \tau_{iq}) \ln(1 - \alpha_q) \} \\ &\quad - \sum_{i=1}^N \sum_{q=1}^Q \{ \tau_{iq} \ln \tau_{iq} + (1 - \tau_{iq}) \ln(1 - \tau_{iq}) \}. \end{aligned} \quad (13)$$

Malheureusement, trouver les paramètres maximisant cette expression reste difficile dans la mesure où le terme $\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j} [\ln g(-a_{ij})]$ n'admet pas d'écriture plus simple. Nous introduisons alors un second niveau d'approximation en utilisant une inégalité introduite par Jaakkola et Jordan [58] qui ont montré que, pour tout réel ξ_{ij} ,

$$\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j} [\ln g(-a_{ij})] \geq \ln g(\xi_{ij}) - \frac{(\tilde{\boldsymbol{\tau}}_i^\top \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \xi_{ij})}{2} - \lambda(\xi_{ij}) \left(\mathbb{E}_{\mathbf{Z}_i, \mathbf{Z}_j} [(\tilde{\mathbf{Z}}_i^\top \tilde{\mathbf{W}} \tilde{\mathbf{Z}}_j)^2] - \xi_{ij}^2 \right). \quad (14)$$

Ceci permet de déterminer une borne inférieure de la première borne inférieure :

$$\ln p(\mathbf{X} | \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \geq \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}) \geq \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \boldsymbol{\xi}). \quad (15)$$

où

$$\begin{aligned} \mathcal{L}(q; \boldsymbol{\alpha}, \tilde{\mathbf{W}}, \boldsymbol{\xi}) = & \sum_{i \neq j}^N \left\{ \left(X_{ij} - \frac{1}{2} \right) \tilde{\boldsymbol{\tau}}_i^\top \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j + \ln g(\xi_{ij}) - \frac{\xi_{ij}}{2} \right. \\ & \left. - \lambda(\xi_{ij}) \left(\text{Tr} \left(\tilde{\mathbf{W}}^\top \tilde{\mathbf{E}}_i \tilde{\mathbf{W}} \boldsymbol{\Sigma}_j \right) + \tilde{\boldsymbol{\tau}}_j^\top \tilde{\mathbf{W}}^\top \tilde{\mathbf{E}}_i \tilde{\mathbf{W}} \tilde{\boldsymbol{\tau}}_j - \xi_{ij}^2 \right) \right\} \\ & + \sum_{i=1}^N \sum_{q=1}^Q \{ \tau_{iq} \ln \alpha_q + (1 - \tau_{iq}) \ln(1 - \alpha_q) \} \\ & - \sum_{i=1}^N \sum_{q=1}^Q \{ \tau_{iq} \ln \tau_{iq} + (1 - \tau_{iq}) \ln(1 - \tau_{iq}) \}. \end{aligned}$$

Cette dernière borne est alors optimisée par un algorithme de type EM consistant à optimiser successivement le paramètre $\boldsymbol{\xi}$, le couple $(\boldsymbol{\alpha}, \tilde{\mathbf{W}})$ et les paramètres $\boldsymbol{\tau}_i$ définissant la distribution q (cf Algorithme 1).

La complexité de cet algorithme est en $\mathcal{O}(N^2 Q^3)$, ce qui le rend comparable aux algorithmes variationnels pour SBM ([33] et [B9]) dont la complexité est en $\mathcal{O}(N^2 Q^2)$.

4.5. Expériences comparatives

Nous avons mené des expériences à partir de données simulées et réelles afin d'évaluer les performances du modèle OSBM en terme de classification des sommets dans les différents groupes. Les classifications obtenues ont été comparées à celles estimées pour SBM par la procédure variationnelle bayésienne (section 3), pour MMSB [3] et par l'algorithme CFinder [91].

CFinder renvoie directement une classification des sommets. Les trois autres procédures renvoient des probabilités à posteriori d'appartenance aux classes. Pour SBM, les vecteurs \mathbf{Z}_i sont obtenus en mettant à 1 la coordonnée la plus grande de $\boldsymbol{\tau}_i$ et à 0 toutes les autres. Pour OSBM et MMSB, les coordonnées de \mathbf{Z}_i fixées à 1 correspondent aux coordonnées de $\boldsymbol{\tau}_i$ supérieures à un certain seuil. Ce seuil est fixé à 0.5 pour OSBM, le choix du seuil n'influant pas en pratique car toutes les coordonnées des $\boldsymbol{\tau}_i$ sont très proches soit de 0, soit de 1. Pour MMSB en revanche, le seuil est fixé à $1/Q$ car pour un choix plus grand, tous les chevauchements ont tendance à être éliminés.

```

// INITIALIZATION;
Initialize  $\tau$  with an Ascendant Hierarchical Classification algorithm;
Sample  $\tilde{\mathbf{W}}$  from a zero mean  $\sigma^2$  spherical Gaussian distribution;

// OPTIMIZATION;
repeat
  //  $\xi$ -transformation;
  for  $(i, j) \in V$  do
     $\xi_{ij} \leftarrow \sqrt{\text{Tr}(\tilde{\mathbf{W}}^\top \tilde{\mathbf{E}}_i \tilde{\mathbf{W}} \Sigma_j) + \tau_j^\top \tilde{\mathbf{W}}^\top \tilde{\mathbf{E}}_i \tilde{\mathbf{W}} \tau_j}$ ;
  end
  // M-step;
  for  $q=1 : Q$  do
     $\alpha_q \leftarrow \frac{\sum_{i=1}^N \tau_{iq}}{N}$ ;
  end
  Optimize  $\mathcal{L}(q; \alpha, \tilde{\mathbf{W}}, \xi)$  with respect to  $\tilde{\mathbf{W}}$ , with a gradient based optimization algorithm ;
  // E-step;
  repeat
    for  $i=1 : N$  do
      Optimize  $\mathcal{L}(q; \alpha, \tilde{\mathbf{W}}, \xi)$  with respect to  $\tau_i$ , with a box constrained ( $\tau_{iq} \in [0, 1]$ ) gradient based
      optimization algorithm ;
    end
  until  $\tau$  converges;
until  $\mathcal{L}(q; \alpha, \tilde{\mathbf{W}}, \xi)$  converges;

```

Algorithm 1: Estimation des paramètres du modèle OSBM par approche variationnelle

4.5.1. Simulations

Dans un premier temps, nous avons généré 100 graphes d'affiliation en considérant la matrice \mathbf{W} définie par

$$\mathbf{W} = \begin{pmatrix} \lambda & -\epsilon & \dots & -\epsilon \\ -\epsilon & \lambda & & \vdots \\ \vdots & & \ddots & -\epsilon \\ -\epsilon & \dots & -\epsilon & \lambda \end{pmatrix},$$

les vecteurs \mathbf{U} et \mathbf{V} définis par

$$\mathbf{U} = \mathbf{V} = (\epsilon \quad \dots \quad \epsilon),$$

et les valeurs $Q = 4$, $\lambda = 4$, $\epsilon = 1$, et $W^* = -5.5$.

Nous comparons les classifications $\hat{\mathbf{Z}}$ estimées avec le vrai vecteur de classe \mathbf{Z} avec un critère proche de celui de [54] et [55], à savoir la distance $\|\mathbf{Z}\mathbf{Z}^\top, \hat{\mathbf{Z}}\hat{\mathbf{Z}}^\top\|_2$. En effet, ces deux matrices sont invariantes par permutation des classes et leurs coefficients indiquent le nombre de classes communes pour chaque paire de sommets. Leur distance L_2 est alors une bonne façon de mesurer à quel point la structure en classes du modèle initial a été retrouvée.

Les résultats sont présentés à la Figure 6(c).

L'expérience a ensuite été répétée en prenant une matrice \mathbf{W} de la forme

$$\mathbf{W} = \begin{pmatrix} \lambda & \lambda & -\epsilon & \dots & \dots & \dots & -\epsilon \\ -\epsilon & -\lambda & -\epsilon & \dots & \dots & \dots & \vdots \\ \vdots & -\epsilon & \lambda & \lambda & -\epsilon & \dots & \vdots \\ \vdots & \vdots & -\epsilon & -\lambda & -\epsilon & \dots & \vdots \\ \vdots & \vdots & \vdots & -\epsilon & \ddots & -\epsilon & -\epsilon \\ \vdots & \vdots & \vdots & \vdots & -\epsilon & \lambda & \lambda \\ -\epsilon & \dots & \dots & \dots & \dots & -\epsilon & -\epsilon \end{pmatrix}.$$

Si une classe i a une forte connectivité interne sous ce modèle, ses sommets sont également fortement connectés à ceux de la classe $i + 1$, qui a elle-même une connectivité interne faible (Figure 6(b)). Les résultats sont présentés à la Figure 6(d).

On observe dans les deux cas qu'OSBM est l'algorithme retrouvant le mieux la structure de classes initiale. Les mauvais résultats de SBM s'expliquent par le fait qu'il ne modélise pas les chevauchements. La dégradation des résultats de Cfinder dans la deuxième figure s'explique également par le type de situation modélisée car cet algorithme ne permet de retrouver que les classes ayant une forte connectivité interne.

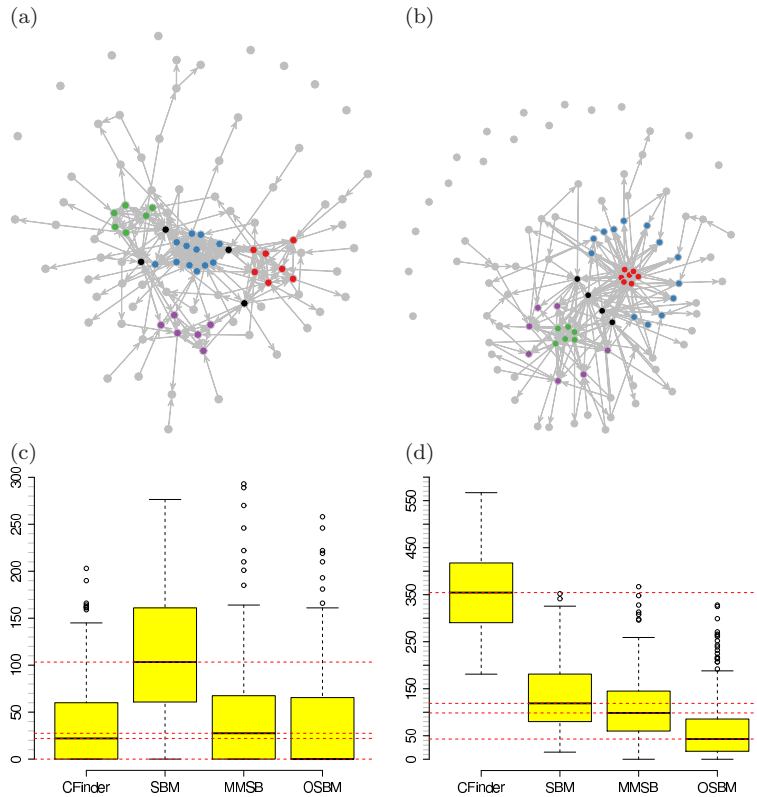


FIGURE 6. (a) Exemple de réseau d'affiliation. Les sommets noirs représentent les chevauchements. (b) Exemple de réseau d'affiliation avec des étoiles. (c-d) Comparaison des algorithmes de classification pour chacun de ces deux modèles.

4.5.2. Blogosphère politique française

Nous avons ensuite considéré un réseau de 196 blogs français dédiés à la politique extraits le 14 octobre 2006 par le projet *Observatoire Présidentielle* [112]. Les noeuds y sont les adresses des blogs et une arête correspond à un hyperlien d'un blog vers un autre. Le jeu de données contient également pour chaque blog le nom du parti politique auquel appartient son auteur ou la mention *Analyste*.

Les résultats des classements des différents noeuds par OSBM sont indiqués à la Figure 7. Une case de la forme $a + b$ indique que a sommets du parti indiqués sont classés uniquement dans cette classe alors que b sommets correspondent à des chevauchements.

Les cinq principaux clusters sont quasiment identiques à ceux retrouvés par MMSB et SBM. OSBM retrouve cependant plus de chevauchements que MMSB. De plus, ces chevauchements semblent pertinents dans la mesure où ils concernent principalement des blogs appartenant à la fois aux groupes UMP et UDF ou à des analystes politiques, l'un de ces derniers apparaissant même dans trois des clusters.

L'algorithme Cfinder a également été utilisé pour ce réseau mais ses résultats sont difficilement comparables car les communautés renvoyées sont beaucoup plus petites en raison de la condition de densité

	UMP	UDF	liberal	PS	analysts	others
cluster 1	30 + 3	0 + 1	0	0	0 + 1	0
cluster 2	2 + 3	29 + 1	0	0	1 + 3	0
cluster 3	0	0	24	0	1 + 1	0
cluster 4	0	0 + 2	0	40	0 + 4	1
outliers	5	1	1	17	5	30

FIGURE 7. Classement des blogs par OSBM

interne très forte. Ainsi, 95 blogs ne sont pas classés, mais les communautés trouvées sont des sous-communautés de celles trouvées par les trois autres algorithmes.

4.5.3. Réseau de régulation de la levure

Enfin, nous avons comparé les quatre algorithmes en utilisant un sous-réseau du réseau de régulation de la levure composé de 197 sommets pour 303 arêtes. Ce réseau est composé notamment de trois couples de régulateurs et de leurs gènes co-régulés.

Nous avons par conséquent OSBM avec $Q = 6$ classes et SBM et MMSB avec $Q = 7$ classes. Les trois algorithmes ont isolés les trois couples de régulateurs et les trois groupes correspondant aux gènes co-régulés pour chacun des couples. Cependant, OSBM est le seul à avoir repéré 5 gènes correspondant à des chevauchements parmi les groupes de gènes co-régulés.

L'algorithme Cfinder ne trouve aucune structure d'intérêt car ce réseau est si peu dense que sa plus grande clique est un triangle.

4.6. Un critère de choix de modèle

L'algorithme présenté au paragraphe précédent permet d'estimer les paramètres du modèle pour une valeur de Q fixée. Se pose alors la question d'un critère permettant de choisir le nombre de classes optimal. Le traitement de cette question est actuellement en cours de finition, toujours en collaboration avec Pierre Latouche et Christophe Ambroise. L'idée est similaire à celle développée pour le critère *ILvb* dans le cas non-chevauchant (section 3.3), à savoir adapter l'estimation variationnelle des paramètres à un cadre bayésien et maximiser dans ce cadre la log-vraisemblance marginale $\log p(\mathbf{X} | Q)$.

cluster	size	operons
1	2	STE12 TEC1
2	33	YBR070C MID2 YEL033W SRD1 TSL1 RTS2 PRM5 YNL051W PST1 YJL142C SSA4 YGR149W SPO12 YNL159C SFP1 YHR156C YPS1 YPL114W HTB2 MPT5 SRL1 DHH1 TKL2 PGU1 YHL021C RTA1 WSC2 GAT4 YJL017W TOS11 YLR414C BNI5 YDL222C
3	2	MSN4 MSN2
4	32	CPH1 TKL2 HSP12 SPS100 MDJ1 GRX1 SSA3 ALD2 GDH3 GRE3 HOR2 ALD3 SOD2 ARA1 HSP42 YNL077W HSP78 GLK1 DOG2 HXK1 RAS2 CTT1 HSP26 TPS1 TTR1 HSP104 GLO1 SSA4 PNC1 MTC2 YGR086C PGM2
5	2	YAP1 SKN7
6	19	YMR318C CTT1 TSA1 CYS3 ZWF1 HSP82 TRX2 GRE2 SOD1 AHP1 YNL134C HSP78 CCP1 TAL1 DAK1 YDR453C TRR1 LYS20 PGM2

TABLE 3

Classification des opérons en $Q = 6$ clusters. Les opérons indiqués en gras appartiennent à plusieurs clusters.

5. Un modèle basé sur les graphes bipartis

5.1. Contexte

Certains types de réseaux présentent une structure particulière dans la mesure où leurs sommets sont de deux types différents et où toute arête relie des sommets de types différents. On parle alors de structure bipartite. Cette structure peut être présente de façon explicite, par exemple dans le cas des réseaux métaboliques. Ceux-ci sont formés de réactions et de métabolites, les arêtes représentant la consommation ou la production d'un métabolite par une réaction (cf Figure 5.1).

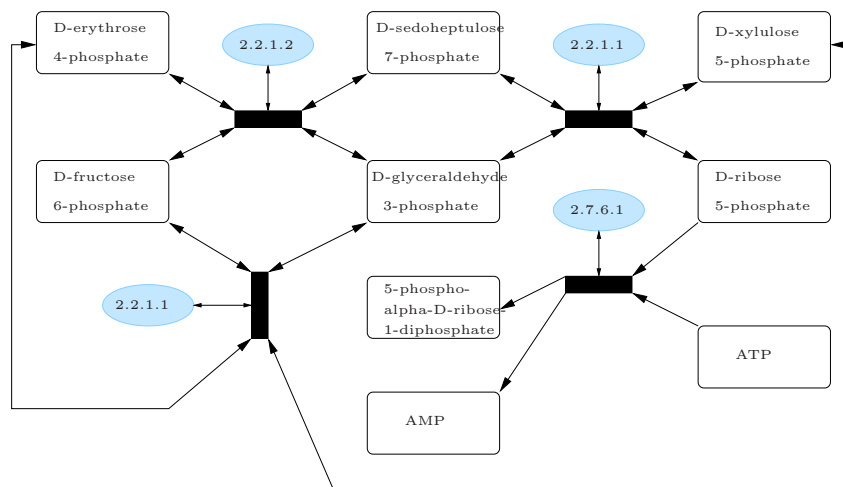


FIGURE 8. Partie du réseau métabolique de la levure tiré de [63]. Chaque réaction (rectangles noirs) est caractérisée par la classification de l'enzyme qui la gouverne (ovales bleus) et est relié aux métabolites qu'elle consomme et produit.

Cependant, la structure bipartite peut également être présente de façon sous-jacente. Il peut ainsi être intéressant d'étudier le réseau des enzymes régulant des réactions du métabolisme, deux enzymes étant

reliées si l'une régule la production d'un métabolite dont l'autre régule la consommation [72]. Dans ce cas, on oublie le métabolite intermédiaire afin d'obtenir un réseau de sommets homogènes. Ceci est également le cas dans le cas des réseaux de régulation de gènes où le facteur de transcription est assimilé au gène qui l'encode.

Cette structure permet dans certains cas de comprendre l'origine des caractéristiques structurelles des réseaux. Par exemple, un métabolite participant à de nombreuses réactions assurera à chacune d'elles un fort degré dans le graphe des réactions. Elle a par conséquent été utilisée par Guillaume et Latapy [50] ou Newman [84] pour construire des modèles de génération de graphes aléatoires par simulation. Je propose d'en tirer parti pour décrire un modèle stationnaire générant des graphes ayant des propriétés topologiques satisfaisantes [B6].

5.2. Le modèle

Un graphe biparti non orienté est un triplet $H = (\top, \perp, E)$ où \top et \perp sont des ensembles de sommets disjoints et $E \subseteq \top \times \perp$ est l'ensemble des arêtes. Chaque arête du graphe relie donc un sommet de \top et un sommet de \perp .

La \perp -projection de H est le graphe $G = (\perp, E')$ tel que (u, v) est une arête de E' si et seulement si u et v sont reliés à un même sommet de \top dans H (cf Figure 9).

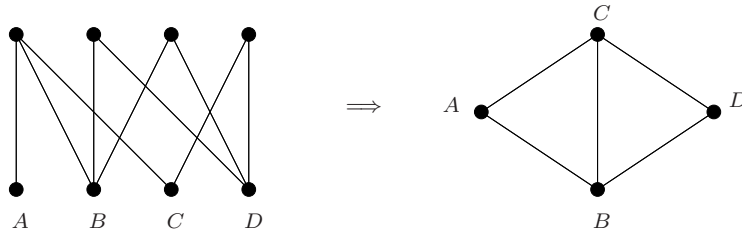


FIGURE 9. Un graphe biparti et sa \perp -projection

Soit n et m deux entiers et $\alpha > 1$. Considérons le modèle de graphes bipartis suivant, que nous notons $BG_{n,m,\alpha}$:

- (i) soit deux ensembles de sommets, notés \perp et \top , et de tailles respectives n et m ;
- (ii) pour tout $1 \leq i \leq n$, on tire de façon indépendante un entier $p_i \in \{1, \dots, m\}$ suivant une loi de puissance tronquée de paramètre α .

$$\mathbb{P}(p_i = k) = \begin{cases} \frac{C_\alpha(m)}{k^\alpha} & \text{si } 1 \leq k \leq m \\ 0 & \text{sinon} \end{cases},$$

où $C_\alpha(m)$ est une constante de normalisation.

- (iii) pour tout $(i, j) \in \perp \times \top$, on trace une arête entre i et j avec probabilité $\frac{p_i}{m}$. Conditionnellement aux valeurs des p_i , ces tirages sont indépendants.

Intuitivement, ce mode de génération revient à tirer un degré *a priori* p_i pour chacun des sommets de \perp puis à tirer les arêtes qui en sortent suivant une binomiale d'espérance p_i . Le degré effectif du sommet i sera donc en moyenne égal à son degré *a priori*.

Le modèle final de graphes aléatoires, noté $SFG_{n,m,\alpha}$ consiste alors à générer un graphe suivant $BG_{n,m,\alpha}$ et à en prendre la \perp -projection.

Il est à noter que ce modèle se généralise sans difficultés aux graphes orientés. Il suffit pour cela de générer un degré sortant et un degré entrant *a priori* pour chaque sommet de \perp , suivant des lois de puissance de paramètres α_{out} et α_{in} , puis de générer le graphe biparti aléatoire comme précédemment. La \perp -projection se fait en ajoutant une arête entre deux sommets u_1 et u_2 s'il existe un chemin de longueur 2 les reliant dans le graphe biparti.

5.3. Propriétés topologiques des graphes générés

Dans ce paragraphe, nous listons les caractéristiques topologiques des graphes générés sous $SFG_{n,m,\alpha}$ lorsque n et m tendent vers l'infini en restant du même ordre de grandeur. On suppose de plus que $2 < \alpha < 3$ et on note $\lambda_\alpha = \frac{C_\alpha}{C_{\alpha-1}}$ où $C_\alpha = \zeta^{-1}(\alpha)$ est la limite des $C_\alpha(m)$ lorsque m tend vers l'infini.

Les démonstrations sont détaillées dans [B6].

5.3.1. Distribution des degrés

Les auteurs de [50] étudient le problème consistant à partir d'un graphe G quelconque, à retrouver le plus petit graphe H possible tel que G est une \perp -projection de H . Ce problème se ramène à un problème de couverture minimale par des cliques et est par conséquent NP-difficile. Cependant, ils mettent au point une heuristique satisfaisante dans la mesure où les graphes H obtenus ont un nombre de sommets du même ordre dans \perp et \top . Les distributions de degrés qu'ils obtiennent en appliquant leur heuristique sur des graphes réels sont des distributions de Poisson pour les sommets de \top et des distributions en loi de puissance pour des sommets de \perp .

Ces distributions sont également celles retrouvées dans notre modèle.

Proposition 1. *Dans le cadre du modèle $BG_{n,m,\alpha}$,*

1. *soit*

$$p_\top = \sum_{p=1}^m \frac{p}{m} \frac{C_\alpha(m)}{p^\alpha} = \frac{C_\alpha(m)}{m C_{\alpha-1}(m)} \approx \frac{\lambda_\alpha}{m}.$$

Alors le degré des sommets de \top suit une loi binomiale de paramètres n et p_\top , qui peut être approchée par une loi de Poisson pour m et n grands.

2. *soit $H_\alpha(k) = \lim_{m \rightarrow \infty} \mathbb{P}(d_H(v) = k)$, $v \in \perp$. Alors $H_\alpha(k)$ existe pour tout k positif et*

$$H_\alpha(k) = \frac{C_\alpha}{k!} \sum_{p=1}^{+\infty} p^{k-\alpha} e^{-p} \underset{k \rightarrow +\infty}{\sim} \frac{C_\alpha}{k^\alpha}$$

La distribution des degrés est donc asymptotiquement une distribution de type scale-free.

Le modèle étant suffisamment simple conditionnellement aux valeurs des p_i , il est également possible de déterminer une formule exprimant la loi des degrés dans la \perp -projection en fonction de celle des sommets de \perp dans le graphe biparti. L'étude asymptotique de cette formule amène au résultat principal concernant le modèle $SFG_{n,m,\alpha}$, à savoir qu'il génère des graphes en loi de puissance de paramètre α .

Théorème 3. *Considérons le modèle $SFG_{n,m,\alpha}$ et soit $G_\alpha(k) = \lim_{m \rightarrow \infty} \mathbb{P}(d_G(v) = k)$. Alors $G_\alpha(k)$ existe pour tout $k \geq 1$ et*

$$G_\alpha(k) = \frac{1}{k!} \sum_{j=0}^{+\infty} e^{-\lambda_\alpha j} (\lambda_\alpha j)^k H_\alpha(j) \sim_{k \rightarrow +\infty} \frac{C_\alpha \lambda_\alpha^{\alpha-1}}{k^\alpha}.$$

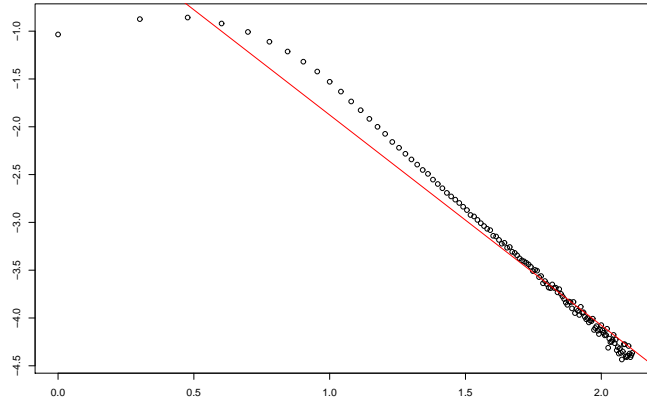


FIGURE 10. Tracé log-log de la distribution empirique des degrés déterminée à partir de 1000 graphes tirés sous le modèle $SFG_{200,200,2.2}$. La droite correspond à l'équivalent issu du Théorème 3.

5.3.2. Densités globale et locale

En plus de la distribution des degrés, le modèle $SFG_{n,m,\alpha}$ concorde avec la réalité en termes de densité. En effet, d'un point de vue global, l'espérance du nombre d'arêtes vérifie

$$\mathbb{E}(e(G)) \sim_{n \rightarrow \infty} \frac{\lambda_\alpha^2}{2} n,$$

et, pour tout $\epsilon < \frac{2-\alpha}{2}$,

$$\mathbb{P}(|e(G) - \mathbb{E}(e(G))| > \frac{\mathbb{E}(e(G))}{n^\epsilon}) \leq \frac{n^{2\epsilon} \text{Var}(e(G))}{\mathbb{E}(e(G))^2} \sim_{n \rightarrow \infty} \frac{4K}{\lambda_\alpha^4} n^{2-\alpha+2\epsilon}$$

En d'autres termes, le nombre d'arêtes est concentré autour de son espérance, qui elle-même croît linéairement en le nombre de sommets.

D'un point de vue local, on observe un phénomène d'attachement préférentiel. En effet, la probabilité pour de deux sommets d'être voisins sachant qu'ils ont un voisin commun varie en $O(\frac{1}{n^{\alpha-1}})$. Elle est donc plus grande que la probabilité générale pour deux sommets d'être voisins, qui est elle équivalente à $\frac{\lambda^2}{n}$. Ceci se traduit également par un coefficient de clustering ne tendant pas vers 0 puisque

$$\lim_{n \rightarrow \infty} \mathbb{P}(C(G) \leq \frac{e-2}{2e^2}) = 0.$$

5.3.3. Composante géante et diamètre

Pour finir, le modèle $SFG_{n,m,\alpha}$ est également fidèle à la réalité en termes de composante géante et de diamètre. Pour démontrer cela, il est à noter qu'un autre graphe aléatoire est obtenu à partir d'un graphe aléatoire biparti en projetant ce dernier sur les sommets de \mathbb{T} . Celui-ci a la particularité d'avoir des arêtes indépendantes deux à deux, c'est-à-dire d'être un graphe d'Erdős-Rényi de paramètre

$$\rho = 1 - \prod_{i=1}^m \left(1 - \frac{p_i^2}{m^2}\right)$$

Celui-ci est asymptotiquement presque sûrement supérieur à $\frac{1}{m}$, ce qui assure que le projeté sur les sommets de \mathbb{T} admet une composante géante et un diamètre en $O(\log(m))$ [20, 27]. Partant de là, on peut en déduire que le graphe biparti puis les projeté sur \perp vérifient les mêmes propriétés.

5.4. Conclusion

Le modèle $SFG_{n,m,\alpha}$ est un modèle à tirage simultané, dont le nombre de paramètres est particulièrement faible par rapport aux autres modèles à tirage simultané existants. Pourtant, il génère des graphes ayant les principales caractéristiques topologiques des réseaux biologiques, grâce à l'utilisation de la projection d'un graphe biparti.

Malheureusement, si l'écriture de la vraisemblance ou de l'espérance de comptages divers est envisageable pour le modèle biparti sous-jacent, la multiplicité des structures biparties ayant les mêmes projections rend impossible ces écritures dans le modèle final. De plus, si la structure bipartie sous-jacente est effectivement présente dans certains réseaux réels, les topologies de ces structures ne se retrouvent pas dans le modèle $BG_{n,m,\alpha}$. Un exemple en est la distribution des degrés de \mathbb{T} qui suit une loi de Poisson dans $BG_{n,m,\alpha}$, ce qui ne correspond pas à la distribution du nombre de substrats ou de produits dans les réseaux métaboliques.

6. Perspectives

Les perspectives principales en termes de développement de modèles aléatoires me semblent résider dans la spécialisation des modèles. En effet, les caractéristiques énumérées au chapitre 1 sont très générales

au sens où elles sont communes à tous les types de réseaux, qu'ils soient biologiques, sociaux ou informatiques. Or, les mécanismes régissant le développement des réseaux sont hétéroclites, ce qui induit des différences de structure. Ainsi, il a été observé, notamment par Aiello *et al.* [1], que la queue de la distribution des degrés est plus lourde pour les graphes biologiques que pour les autres. De même, les motifs apparaissent de façon groupée dans les graphes de régulation, contrairement au cas des graphes issus de l'électronique comme illustré au chapitre 9.

Un premier pas a été fait en ce sens par Chung et Lu [30] qui ont développé un modèle basé sur le phénomène de duplication, qui est le mécanisme principal d'évolution des réseaux de régulation et d'interactions entre protéines. Ils étudient la distribution des degrés correspondante et démontrent que sous leur modèle, l'exposant β de la loi de puissance obtenue vérifie $1 < \beta < 2$, ce qui correspond bien au plus grand nombre de hubs observés dans ce type de graphes. Cependant, si leur modèle est conforme aux mécanismes biologiques régissant la duplication, il est quasi-impossible à exploiter d'un point de vue statistique autrement que par simulations. En effet, les relations complexes de dépendance dans le réseau final rendent par exemple impossible l'écriture de la probabilité d'apparition d'une arête.

Le besoin de spécialisation des modèles se fait encore plus particulièrement ressentir dans le cadre des réseaux métaboliques. En effet, un tel réseau est composé de réactions et de métabolites, ces deux types de sommets jouant des rôles très différents. L'approche par graphes bipartis proposée au chapitre 5 ne permet pas de maîtriser suffisamment le degré des sommets correspondant aux réactions, celles-ci ne faisant jamais intervenir plus de cinq composants dans les réseaux réels. Une autre solution est de recourir à la notion d'hypergraphes [111], comme cela sera détaillé au chapitre 11. La taille des hyper-arêtes pourra alors être introduite comme un paramètre du modèle. Mithani *et al.* [81] ont proposé un modèle adapté à la structure d'hypergraphe. Un équivalent à tirage simultané permettant l'étude de motifs reste cependant à construire.

Deuxième partie

Recherche de motifs dans les réseaux biologiques

7. Introduction

7.1. Motivations

Une manière de synthétiser l'information structurelle d'un réseau est de s'intéresser au comptage des petits sous-graphes afin de voir si certains d'entre eux apparaissent en nombre plus important qu'attendu. Il est alors intéressant d'étudier ces sous-graphes, appelés motifs, afin de déterminer si leur sur-représentation est liée à une fonction biologique qui a favorisé leur apparition au cours de l'évolution.

Cette idée est apparue dans les années 70 en sciences sociales avec l'étude des relations entre groupes de trois personnes ou *triad censuses* [34, 106]. Elle a été reprise et popularisée en biologie par l'algorithme de recherche de motifs de Milo *et al.* [80]. Cet algorithme, basé sur le procédé de *stub-rewiring* (cf 2) et l'approximation normale de la loi des comptages, a depuis été amélioré et complété [67, 68, 108]. Picard *et al.* [92] ont montré que l'approximation devient meilleure en remplaçant le *stub-rewiring* par un modèle de mélange et la loi normale par une loi de Poisson composée. Berg et Lässig [14] ont proposé une approche différente basée sur l'alignement de graphes pour détecter les motifs. Finalement, Banks *et al.* [11] et Schbath *et al.* [100] ont étendu l'idée initiale à la notion de motifs colorés dans des réseaux dotés d'une classification de leurs sommets.

Parmi les articles offrant une étude biologique des motifs trouvés, on peut citer les travaux de Wuchty *et al.* [110] qui démontrent le lien entre motifs et structures conservées lors de l'évolution ou ceux de Alon [6] ou Kaplan *et al.* [65] qui étudient l'utilité d'un motif particulier appelé *feed-forward loop* (cf Figure 11) dans l'introduction de délais en terme de régulation. Enfin, de nombreux travaux n'ont pas les motifs comme sujet principal mais les utilisent comme outil de démonstration des phénomènes étudiés [10, 79, 93, 94].

Si l'utilisation des motifs comme élément d'analyse de la structure des réseaux est devenue relativement commune, plusieurs questions restent ouvertes d'un point de vue théorique sur leur détection. La première est d'ordre algorithmique et concerne la mise au point de méthodes de comptage systématiques et efficaces de tous les sous-graphes d'un réseau. Les algorithmes actuels se limitent en effet essentiellement aux motifs de taille au plus 8 dans un réseau d'au plus quelques milliers de noeuds. La seconde question est la loi du comptage d'un sous-graphe particulier dans un modèle aléatoire réaliste du type SBM. Les algorithmes de détection de motifs existants sont en effet basés sur une approximation normale non justifiée de cette loi. Enfin, il est nécessaire de pouvoir sélectionner uniquement les structures d'intérêt et non toutes les structures contenant une structure d'intérêt.

7.2. Schémas et occurrences

Par la suite, je désignerai par *schéma* un petit graphe donné dont on cherche à déterminer s'il est sur-représenté dans le réseau observé ou non. Le terme *motif* désigne un schéma qui est sur-représenté. En théorie, un schéma peut être de taille quelconque mais les contraintes en termes de temps de calcul ainsi que la volonté d'en avoir une interprétation biologique amènent à considérer uniquement des schémas de moins d'une dizaine de sommets.

Un *automorphisme* du schéma \mathbf{m} est une permutation ϕ de ses sommets telle que, pour toute paire (a, b) de sommets, $(\phi(a), \phi(b))$ (respectivement $\overrightarrow{\phi(a)\phi(b)}$ dans le cas orienté) est une arête si et seulement si (a, b) (respectivement \overrightarrow{ab}) est une arête. Soit \mathcal{R} la relation définie par $a\mathcal{R}b$ s'il existe un automorphisme de \mathbf{m} dont l'image de a est b . \mathcal{R} est une relation d'équivalence sur les sommets de \mathbf{m} , qui peuvent donc être partitionnés en classes d'équivalence. Par exemple, le motif *bi-fan* de la Figure 11 présente deux classes d'équivalence qui sont $\{a, b\}$ et $\{c, d\}$. Le nombre d'automorphismes de \mathbf{m} est désigné par $\text{aut}(\mathbf{m})$.

Soit (C_1, \dots, C_K) les classes d'équivalence de \mathbf{m} et (i_1, \dots, i_K) leurs tailles respectives. Une *position potentielle* U pour \mathbf{m} dans G est une liste (V_1, \dots, V_K) d'ensembles de sommets disjoints de G de tailles respectives (i_1, \dots, i_K) . Cette position est une *occurrence de \mathbf{m}* si le sous-graphe de G induit par U est isomorphe à \mathbf{m} .

La définition d'une occurrence comme un ensemble de sous-ensembles correspondant aux classes d'équivalence induit deux possibilités pour l'énumération des occurrences. Ce choix se pose à la fois dans les écritures mathématiques et dans la mise au point des algorithmes de comptage (cf 7.3). On peut ainsi

- soit parcourir l'ensemble des positions potentielles et compter une unique fois chaque occurrence du schéma d'intérêt ;
- soit parcourir l'ensemble des listes de k sommets et diviser le résultat par $\text{aut}(\mathbf{m})$, chaque occurrence ayant été parcourue $\text{aut}(\mathbf{m})$ fois.

Une autre manière de définir une occurrence est également utilisée dans la littérature, à savoir que l'on considère comme une occurrence d'un schéma tout sous-graphe de G de taille k contenant le schéma d'intérêt. Cela revient dans la définition précédente à considérer qu'il suffit qu'il existe une injection du schéma dans $G[U]$, où en d'autres termes à considérer tous les sous-graphes de G plutôt que ses seuls sous-graphes induits. Dans la suite de ce travail, je privilégierai les occurrences telles que définies initialement, c'est-à-dire que je considérerai des sous-graphes induits. Ce choix est lié d'une part au fait qu'il me semble biologiquement plus pertinent de considérer des sous-graphes induits car, dans la mesure où les réseaux biologiques sont creux, la présence d'une arête supplémentaire est porteuse d'information ; d'autre part, dans la mesure où les schémas considérés sont petits, les relations linéaires reliant les comptages des sous-graphes et les comptages des sous-graphes induits peuvent être explicitées si besoin.

7.3. Le problème du comptage

Pour pouvoir étudier la sur-représentation d'un schéma, il faut commencer par compter ses occurrences. Il peut également être intéressant de les énumérer, c'est-à-dire de garder en mémoire leurs positions. Ce problème est crucial d'un point de vue des performances algorithmiques des procédures de détection de motifs. En effet, l'approche naïve consistant à parcourir toutes les positions potentielles induit une complexité en n^k qui la rend très rapidement inutilisable. Deux options se présentent pour améliorer les performances du comptage. Elles reposent toutes deux sur le principe du backtracking, qui permet de tirer profit du caractère creux des réseaux réels.

La première option est d'énumérer simultanément tous les sous-graphes de taille k en parcourant tous les ensembles connexes de taille k . L'utilisation du backtracking permet de traiter la condition de connexité en étendant la recherche uniquement aux voisins des sommets de l'ensemble courant. Ce procédé, associé à un ordre des sommets du réseau permettant d'éviter la prise en compte multiple d'une même occurrence, est implémenté dans l'algorithme FANMOD de Wernicke et Rasche [108]. L'avantage de cette méthode est qu'elle parcourt une fois et une seule fois tout sous-graphe induit de taille k tout en ne faisant aucune recherche inutile. Elle implique cependant une phase de mise sous forme canonique et de tri des occurrences ainsi obtenues afin de regrouper celles qui correspondent à un même schéma.

Afin d'éviter la phase finale de l'algorithme FANMOD, qui se révèle assez coûteuse en temps, il est possible de travailler pour un schéma donné. Grochow et Kellis [49] proposent alors un algorithme d'énumération basé sur un backtracking qui permet d'obtenir directement les sommets de l'occurrence dans un ordre correspondant à une forme canonique. Cette méthode peut encore être accélérée dans le cas du comptage seul en ne parcourant pas les sommets de degré 1 du schéma et en optimisant l'ordre de parcours des autres sommets. Cependant, établir la liste de tous les schémas possibles de taille k n'est faisable que pour un k petit (il y a par exemple plus de 11 millions de graphes connexes à dix sommets). Cette méthode ne peut donc être utilisée que pour des valeurs petites de k ou pour un schéma donné de taille plus importante. L'algorithme MODA de Omid *et al.* [89] basé sur la réutilisation des occurrences des schémas plus petits est confronté au même problème.

La méthode de Grochow et Kellis [49] et son amélioration limitée aux comptages seront disponibles dans le package R *NeMo* en cours d'achèvement. Des évaluations de leurs performances respectives comparées à FANMOD sont disponibles dans [B19].

Il est également à noter que pour pouvoir traiter de grands réseaux, il est également possible d'avoir recours à de l'échantillonnage plutôt qu'au comptage exact des occurrences [7, 68, 89, 108]

8. Motifs globaux dans le modèle EDD

La notion de sur-représentation dépend d'une hypothèse nulle, c'est-à-dire d'un modèle de graphes aléatoires. En 2006, les méthodes existantes de recherche de motifs étaient toutes basées sur le principe du *stub rewiring*. Or, comme indiqué en 2, ce principe est discutable du point de vue des graphes créés et ne permet pas de calcul analytique.

Une première étape dans l'introduction de modèles simultanés a été de considérer le modèle *EDD* décrit en 2 et de déterminer les deux premiers moments du comptage des petits schémas. Ce travail a été fait pour les schémas non-orientés de taille 3 et 4, et a donné lieu à la publication [B2]. Le principe est développé ci-dessous uniquement pour le schéma \mathbf{m} correspondant à un triangle, les autres schémas se traitant de manière tout à fait similaire. Pour rappel, les probabilités de connexion dans le cadre du modèle EDD non dirigé et sans auto-arêtes sont données par :

$$P(X_{ij} = 1) = \pi_{ij} = \frac{d_i d_j}{C} \quad \text{et} \quad P(X_{ii} = 1) = \pi_{ii} = 0$$

Soit I_k l'ensemble des k -uplets de sommets et, pour tout ensemble J , notons $I_k(J)$ les éléments de I_k n'intersectant pas J . I_3 est alors l'ensemble des positions possibles pour un triangle, ce qui implique que le comptage du nombre $N(\mathbf{m})$ de triangles s'écrit

$$N(\mathbf{m}) = \sum_{\alpha \in I_3} Y_\alpha(\mathbf{m})$$

où $Y_\alpha(\mathbf{m})$ est la variable indicatrice de la présence d'un triangle en position α .

L'espérance du comptage est donc donnée par

$$\begin{aligned} \mathbb{E}(N(\mathbf{m})) &= \sum_{\alpha \in I_3} \mathbb{E}(Y_\alpha(\mathbf{m})) \\ &= \sum_{\{i,j,k\} \in I_3} \pi_{ij} \pi_{jk} \pi_{ik} \\ &= \sum_{\{i,j,k\} \in I_3} \frac{d_i^2 d_j^2 d_k^2}{C^3}. \end{aligned}$$

Le moment d'ordre 2 peut se décomposer en

$$\begin{aligned} \mathbb{E}(N^2(\mathbf{m})) &= \sum_{\alpha \in I_3} \sum_{\beta \in I_3} \mathbb{E}(Y_\alpha(\mathbf{m}) Y_\beta(\mathbf{m})) \\ &= \sum_{|\alpha \cap \beta| = 0} \mathbb{E}(Y_\alpha(\mathbf{m})) \mathbb{E}(Y_\beta(\mathbf{m})) + \sum_{|\alpha \cap \beta| = 1} \mathbb{E}(Y_\alpha(\mathbf{m})) \mathbb{E}(Y_\beta(\mathbf{m})) \\ &\quad + \sum_{|\alpha \cap \beta| = 2} \mathbb{E}(Y_\alpha(\mathbf{m}) Y_\beta(\mathbf{m})) + \sum_{\alpha \in I_3} \mathbb{E}(Y_\alpha^2(\mathbf{m})) \\ &= \sum_{\{i,j,k\} \in I_3} \sum_{\{\ell,u,v\} \in I_3(ijk)} \mathbb{E}(Y_{i,j,k}(\mathbf{m})) \mathbb{E}(Y_{\ell,u,v}(\mathbf{m})) \\ &\quad + \sum_{1 \leq i \leq n} \sum_{\{j,k\} \in I_2(i)} \sum_{\{\ell,u\} \in I_2(ijk)} \mathbb{E}(Y_{i,j,k}(\mathbf{m})) \mathbb{E}(Y_{i,\ell,u}(\mathbf{m})) \\ &\quad + \sum_{\{i,j\} \in I_2} \sum_{k \in I_1\{ij\}} \sum_{\ell \in I_1(ijk)} \mathbb{E}(Y_{i,j,k}(\mathbf{m}) Y_{i,j,\ell}(\mathbf{m})) + \sum_{\{i,j,k\} \in I_3} \mathbb{E}(Y_{i,j,k}^2(\mathbf{m})). \end{aligned}$$

L'espérance du produit a pu être remplacée par le produit des espérances dans les deux premières sommes car l'indépendance entre arêtes assure l'indépendance entre les occurrences du triangle en deux positions à moins que celles-ci partagent au moins deux sommets.

Le produit $Y_{i,j,k}(\mathbf{m})Y_{i,j,\ell}(\mathbf{m})$ est égal à l'indicatrice de la présence simultanée des arêtes ij , ik , il , jk et jl donc

$$\mathbb{E}(Y_{i,j,k}(\mathbf{m})Y_{i,j,\ell}(\mathbf{m})) = \pi_{ij}\pi_{jk}\pi_{ik}\pi_{j\ell}\pi_{i\ell} = \frac{d_i^3 d_j^3 d_k^2 d_\ell^2}{C^5}. \quad (16)$$

Enfin, $Y_{i,j,k}^2(\mathbf{m}) = Y_{i,j,k}(\mathbf{m})$ car il s'agit d'une variable indicatrice. Finalement, en comptant le nombre de répétitions de chaque terme,

$$\begin{aligned} \mathbb{E}(N^2(\mathbf{m})) &= 20 \sum_{\{i,j,k,\ell,u,v\} \in I_6} \frac{d_i^2 d_j^2 d_k^2 d_\ell^2 d_u^2 d_v^2}{C^6} \\ &+ 6 \sum_{i \in I_1} \sum_{\{j,k,\ell,u\} \in I_4(i)} \frac{d_i^4 d_j^2 d_k^2 d_\ell^2 d_u^2}{C^6} \\ &+ 2 \sum_{\{i,j\} \in I_2} \sum_{\{k,\ell\} \in I_2(ij)} \frac{d_i^3 d_j^3 d_k^2 d_\ell^2}{C^5} \\ &+ \sum_{\{i,j,k\} \in I_3} \frac{d_i^2 d_j^2 d_k^2}{C^3}. \end{aligned}$$

Il est à noter que dans [B2], les comptages ne sont pas les comptages des sous-graphes induits mais des sous-graphes en général (ce qui revient au même dans le cas du triangle). Cependant, l'écriture directe en termes de comptage de sous-graphes induits est possible modulo l'apparition supplémentaire de termes en $(1 - \pi_{ik})$ à chaque non arête. De même, l'extension à des graphes dirigés ne poserait pas de problèmes particuliers.

Le triangle étant le schéma pour lequel les calculs sont les plus simples, il est aisé de voir que cette approche est vouée à l'échec pour des schémas de taille plus grande. En effet, outre la complexité de l'écriture, la programmation de ces formules nécessite de parcourir toutes les façons de choisir k sommets dans le réseau. Cette opération peut être légèrement simplifiée en regroupant les sommets de même degrés mais le nombre de degrés différents présents dans un réseau est suffisamment grand pour que cette amélioration reste marginale.

Par la suite, Picard *et al.* [92] ont introduit le même genre d'approche sur le modèle *SBM*, qui a l'avantage d'être stationnaire. Le calcul de l'espérance se ramène alors au calcul de la probabilité d'apparition en une position donnée, que l'on multiplie par le nombre de positions possibles. Cette amélioration permet de traiter des schémas plus grand et de l'appliquer à des réseaux de plusieurs centaines de noeuds.

9. Recherche de motifs locaux

Les méthodes précédemment développées considèrent toutes les motifs de façon globale, à savoir qu'elles s'intéressent au comptage total du nombre d'occurrences d'un schéma dans un réseau. Je propose de considérer une autre définition des motifs, basée sur la notion de sur-représentation locale.

La première raison de ce changement de définition est la constatation par Dobrin *et al.* [38] du fait que les occurrences des motifs détectés par la méthode globale dans le réseau de régulation de la levure sont en fait fortement concentrées en certaines régions du réseau. Cette constatation est reprise par Alon [6] qui définit des *modules de régulation*, correspondant aux deux principales formes de regroupement des motifs de régulation.

Ce phénomène est également mis en lumière par l'étude plus biologique de Zhang *et al.* [114], dont les auteurs introduisent la notion de *thème de motifs*, qu'il définissent par «*recurring higher-order interconnection patterns that encompass multiple occurrences of network motifs*». En d'autres termes, les thèmes correspondent à plusieurs occurrences d'un même motif partageant certains de leurs sommets. Les auteurs recherchent les thèmes dans un réseau dont les sommets sont les gènes de la levure et les arêtes représentent des interactions entre gènes. Ces interactions peuvent être de plusieurs types (interactions génétiques, co-expression, homologie de séquence, régulation, interaction entre les protéines codées), ce qui induit un graphe coloré. Les motifs globaux sont détectés de façon automatique puis les thèmes sont trouvés manuellement et analysés d'un point de vue biologique.

Il en ressort que la notion de thèmes peut être un outil de prédiction, du point de vue des interactions ou des fonctions de gènes. Par exemple, un des thèmes décrits correspond à la corégulation par les gènes *Mcm1* et *Swi4* d'un grand nombre de gènes. La majorité d'entre eux étant liés au contrôle ou à l'exécution du cycle cellulaire, il est probable que ceux d'entre eux ayant une fonction inconnue soient également liés au cycle cellulaire. Cette prédiction a de plus l'avantage de reposer sur la présence simultanée d'un grand nombre d'arêtes, diminuant ainsi le risque lié aux arêtes correspondant à de faux positifs.

La seconde raison du changement de définition est la prise en compte des sous-schémas d'un schéma. En effet, comme cela a été relevé dès les premiers articles concernant les motifs [80], un schéma peut être sur-représenté uniquement en raison de la sur-représentation de l'un de ses sous-schémas. Celui-ci est alors la structure ayant biologiquement un sens, le schéma initial ne méritant d'être étudié que s'il apparaît lui-même en nombre important par rapport au nombre de sous-schémas. Ce conditionnement est pris en compte dans le cas des motifs globaux pour l'étude des motifs de taille $k = 3$ ou $k = 4$ [11, 80] mais ne l'est plus pour des motifs plus grands. En effet, l'étude des motifs de taille k à l'aide de modèles pas à pas nécessite la génération de graphes ayant le même nombre de sous-graphes que le réseau étudié, pour tout sous-graphe de taille 2 à $k - 1$. Pour $k = 3$, cela revient à fixer les degrés, ce qui est fait dans le cas du *stub-rewiring*. Pour $k = 4$, le recours à un algorithme de type recuit-simulé permet de générer des graphes avec le bon nombre de sous-graphes de taille 3 mais la méthode ne semble pas généralisable à des tailles plus importantes.

Ce chapitre est dévolu à la définition de la notion de motif local et à la recherche de motifs locaux dans les réseaux. Il a fait l'objet de [B14] et [B17]. Les démonstrations des propriétés mathématiques sont détaillées dans les annexes du second article.

9.1. Motifs locaux

Tout au long de ce chapitre, G désigne un réseau orienté à n sommets, ayant éventuellement des boucles ou une arête dans chaque sens entre certaines paires de sommets. Tous les résultats présentés peuvent être étendus sans difficulté aux graphes non orientés.

Dans ce contexte, les schémas considérés sont également orientés, avec possibilité de boucles ou d'arêtes opposées. Par convention, les lettres (a, b, \dots) désignent les sommets du schéma \mathbf{m} considéré, alors que les lettres (u, v, \dots) désignent les sommets du réseau G .

Un *sous-schéma* de \mathbf{m} désigne une paire (C, \mathbf{m}') , où C est une classe d'équivalence de \mathbf{m} et \mathbf{m}' le schéma à $k - 1$ sommets obtenu en supprimant l'un des sommets de C (cf Figure 11).

La relation d'équivalence entre les sommets de C assure que le schéma \mathbf{m}' est bien le même quelque soit le sommet de C qui est supprimé. Cependant, l'inverse n'est pas vrai : les sous-schémas obtenus en supprimant des sommets a ou b peuvent être isomorphes sans que a et b appartiennent à la même classe d'équivalence. Le schéma *FFL* de la Figure 11 en est un exemple puisque la suppression de chaque sommet amène à une arête simple alors que les sommets ne sont pas topologiquement équivalents.

Afin de représenter en un seul dessin un schéma et l'un de ses sous-schémas, j'adopte la convention consistant à représenter par un carré plein l'un des sommets de la classe d'équivalence considérée et à mettre en pointillés les arêtes qui lui sont incidentes. La suppression de ce sommet et des arêtes pointillées est alors une représentation de \mathbf{m}' .

Soit \mathbf{m} un schéma et (C, \mathbf{m}') l'un de ses sous-schémas. Une occurrence V de \mathbf{m} est une *extension* d'une occurrence U de \mathbf{m}' si les sommets de U sont un sous-ensemble des sommets de V . Un (\mathbf{m}, C) -*thème en* U est alors le sous-graphe induit par l'occurrence de \mathbf{m}' en U et de ses extensions. Le nombre de ces extensions est appelé l'*ordre* du thème (cf Figure 12).

La sur-représentation locale, c'est-à-dire la définition de motif local, peut alors être définie par l'apparition de thèmes d'ordres plus grands qu'attendus. Cela nécessite des tests statistiques, avec une correction liée aux multiples positions testées. Plus précisément, afin de détecter les motifs locaux de taille inférieure ou égale à k , je propose la procédure suivante, dont les parties seront détaillées dans les sections suivantes :

1. Inférence des paramètres d'un modèle de graphes aléatoires à tirage simultané qui servira d'hypothèse nulle ;
2. Enumération et calcul de l'ordre des thèmes de tout couple (schéma ; sous-schéma) à l'aide d'un algorithme de comptage (cf Section 7.3) ;
3. Calcul d'une p -valeur pour tout couple (schéma ; sous-schéma). Ce calcul se fait à l'aide d'une approximation de Poisson pour la loi de l'ordre des thèmes et d'une correction pour tests multiples ;
4. L'application d'une procédure de filtrage qui permet d'assurer que tout nouveau motif local détecté apporte une information par rapport aux motifs locaux plus petits.

Il est à noter que cette définition est réellement différente de celle des motifs globaux, un motif pouvant être local sans être global et réciproquement.

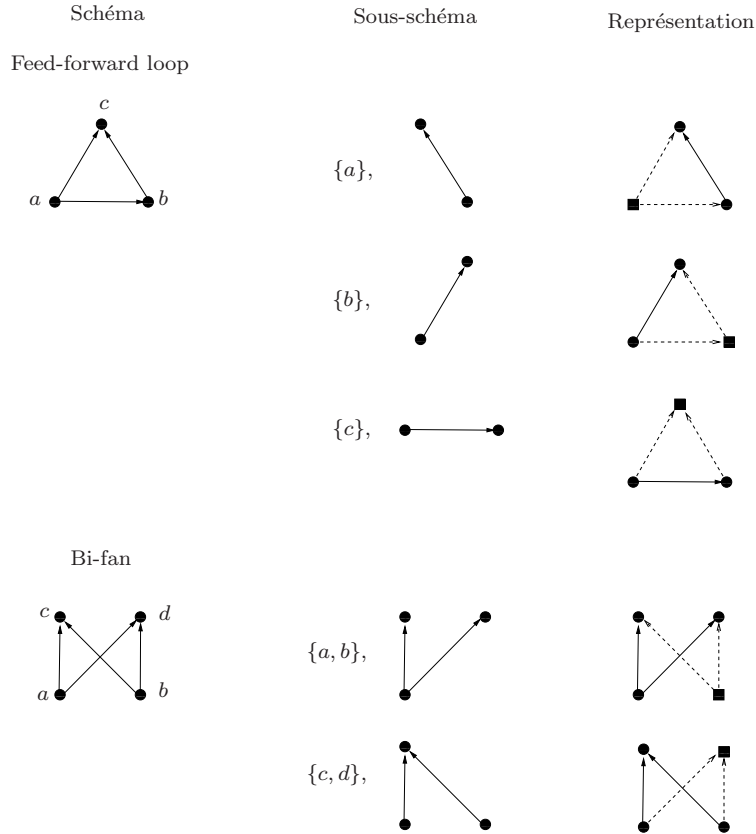


FIGURE 11. Les schémas FFL et bi-fan et la liste de leur sous-schémas. La dernière colonne montre le schéma et le sous-schéma considéré en un seul dessin.

9.2. Le modèle aléatoire

Le modèle aléatoire adopté est proche des modèles de mélange ou à classes empiétantes traités en partie I, à savoir que la probabilité de lien entre deux sommets dépend de leurs classes respectives. Cependant, pour des raisons de développement mathématique, les classes sont supposées connues et non pas latentes. Il dépend donc d'un quadruplet de données $(n, Q, \mathbf{Z}, \mathbf{\Pi})$ où

- n est le nombre de sommets,
- Q est le nombre de classes du modèle,
- $\mathbf{Z} \in \{1, \dots, Q\}^n$ est un vecteur donnant la classe de chaque sommet,
- $\mathbf{\Pi}$ est une matrice de connectivité de taille $Q \times Q$.

Sous ce modèle, les arêtes sont tirées de manière indépendante suivant des lois de Bernoulli, c'est-à-dire

$$X_{uv} \sim \mathcal{B}(\mathbf{\Pi}_{Z(u), Z(v)}).$$

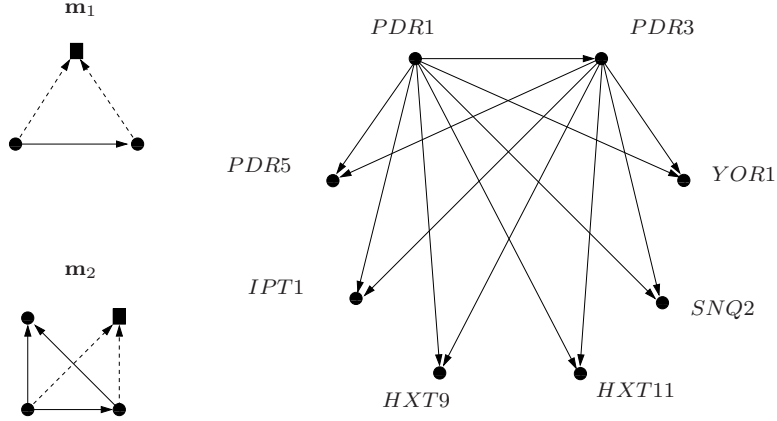


FIGURE 12. Sous-graphe du réseau de régulation de la levure, qui est un (\mathbf{m}_1, C_1) -thème d'ordre 6 en position $(\{PDR1, PDR3\})$ et un (\mathbf{m}_2, C_2) -thème d'ordre 5 en position $(\{PDR1, PDR3\}, \{PDR5\})$, où C_1 et C_2 sont les classes d'équivalences respectives des sommets carrés.

L'estimation des paramètres est faite en utilisant l'une des méthodes décrites en partie I, et en assignant à chaque sommet sa classe de probabilité à postériori la plus grande.

9.3. Approximation de Poisson pour la p-valeur

Considérons un schéma \mathbf{m} de taille k , un sous-schéma (C, \mathbf{m}') de \mathbf{m} et une position potentielle U de \mathbf{m}' dans G .

Soit $N_U(\mathbf{m}, \mathbf{m}')$ l'ordre du (\mathbf{m}, C) -thème situé en U (cf Figure 13). Si U ne correspond pas à une occurrence de \mathbf{m}' , $N_U(\mathbf{m}, \mathbf{m}') = 0$.

Posons $\lambda_U(\mathbf{m}, \mathbf{m}') = \mathbb{E}(N_U(\mathbf{m}, \mathbf{m}') | G[U] \sim \mathbf{m}')$ et

$$\Delta_U(\mathbf{m}, \mathbf{m}') = \frac{N_U(\mathbf{m}, \mathbf{m}') - \lambda_U(\mathbf{m}, \mathbf{m}')}{\lambda_U(\mathbf{m}, \mathbf{m}')}.$$

Les thèmes dont les ordres sont significativement plus grands qu'attendu correspondent aux positions pour lesquelles $\Delta_U(\mathbf{m}, \mathbf{m}')$ est significativement plus grand que 1.

Afin d'alléger les notations, les arguments \mathbf{m} et \mathbf{m}' des quantités N_U , λ_U et Δ_U ne seront écrits que si nécessaire. De même, le cas où la nullité de λ_U implique un problème de définition pour Δ_U n'est pas abordé car les méthodes utilisées pour l'inférence des paramètres du modèle assurent que λ_U est non nul dès lors que N_U est non nul.

9.3.1. Borne locale

Soit U une occurrence de \mathbf{m}' . Pour tout sommet $v \notin U$, soit I_U^v la variable indicatrice du fait que $U \cup \{v\}$ est une extension de \mathbf{m}' et $p_U^v = \mathbb{P}(I_U^v = 1)$. Alors $N_U = \sum_{v \notin U} I_U^v$ et $\lambda_U = \sum_{v \notin U} p_U^v$. Comme

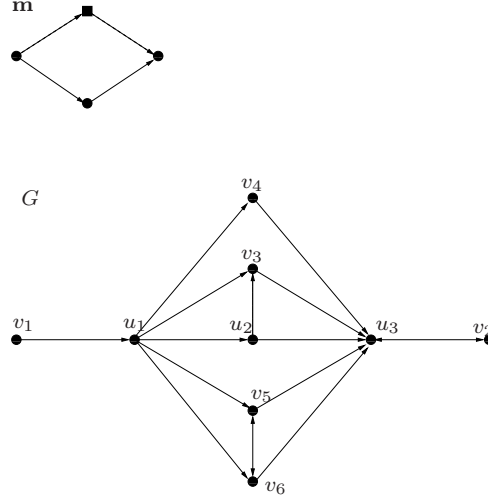


FIGURE 13. Pour $U = \{u_1, u_2, u_3\}$, $N_U(\mathbf{m}, \mathbf{m}') = 3$, car les sommets menant à des extensions sont v_4 , v_5 et v_6 . Dans un modèle d'Erdős où chaque arête potentielle est présente avec probabilité p , chaque sommet n'appartenant pas à U donne une extension avec une probabilité $p_U^v = p^2(1-p)^4$. Par conséquent, $\lambda_U(\mathbf{m}, \mathbf{m}') = 6p^2(1-p)^4$.

les variables $(I_U^v)_{v \notin U}$ sont indépendantes, la loi de N_U peut être approchée par une loi de Poisson de paramètre λ_U en utilisant la méthode de Chen-Stein [13, 25]. Plus précisément, en notant d_{TV} la distance en variation totale entre deux distributions,

$$d_{TV}(\mathcal{L}(N_U), \mathcal{L}(Po(\lambda_U))) \leq \min(1, \lambda_U^{-1}) \sum_{v \notin U} (p_U^v)^2.$$

Cette approximation peut encore être améliorée pour la queue de la distribution, en appliquant toujours la méthode de Chen-Stein. Ceci a été fait par Barbour *et al.* [13] qui démontrent un résultat (Theorem 2.R, p44) impliquant

$$\forall K > 2\lambda_U, \quad \mathbb{P}(N_U \geq K | G[U] \sim \mathbf{m}') \leq \frac{K - \lambda_U}{K - 2\lambda_U} Po(\lambda_U)([K, +\infty)), \quad (17)$$

Pour $K = \lceil \lambda_U(1+t) \rceil$ avec $t > 1$, cela implique

$$\forall t > 1, \mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}') \frac{\sqrt{t+1}}{\sqrt{2\pi\lambda_U}(t-1)} e^{-\lambda_U((1+t)\log(1+t)-t)} \quad (18)$$

Une autre approche consiste à appliquer une inégalité de concentration. Il existe en effet un grand nombre d'inégalités permettant d'établir que des variables de la forme $f(Z_1, \dots, Z_n)$ sont concentrées autour de leur moyenne pour certaines formes de fonctions f et des variables aléatoires Z_1, \dots, Z_n indépendantes [9, 22, 77, 104].

Dans le cas présent, N_U est une somme de variables aléatoires de Bernoulli indépendantes. On peut alors utiliser le Théorème 2.3 de MacDiarmid [77] qui implique

$$\forall t > 0, \quad \mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}') e^{-\lambda_U((1+t)\log(1+t)-t)}. \quad (19)$$

Une comparaison des bornes (18) et (19) montre que la seconde est meilleure pour les petites valeurs de t mais que l'approche par approximation de Poisson se révèle plus précise quand le nombre d'extensions observé est très supérieur au nombre attendu. Le choix de la meilleure des deux bornes suivant la valeur de t implique le théorème suivant :

Théorème 4. *Soit, pour tous $\lambda > 0$ et $t > 0$,*

$$t_\lambda = 1 + \frac{1}{4\pi\lambda}(1 + \sqrt{1 + 16\pi\lambda})$$

et

$$h(\lambda, t) = \begin{cases} 1 & \text{si } t \leq t_\lambda, \\ \frac{\sqrt{t+1}}{\sqrt{2\pi\lambda(t-1)}} & \text{si } t > t_\lambda. \end{cases}$$

Alors, pour tout schéma \mathbf{m} , tout sous-schéma (C, \mathbf{m}') , toute position U et tout $t > 0$,

$$\mathbb{P}(\Delta_U \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}')h(\lambda_U, t)e^{-\lambda_U((1+t)\log(1+t)-t)}.$$

On obtient ainsi une décroissance exponentielle pour la queue de distribution de l'ordre d'un thème en une position donnée.

9.3.2. Borne globale

Le théorème précédent permet de mettre en place un test de sur-représentation locale en une position donnée. Cependant, le nombre de positions potentielles pour \mathbf{m}' est proportionnel à n^{k-1} , ce qui induit un problème de tests multiples. Plutôt que de recourir aux corrections standard pour tests multiples, je propose de déterminer une borne pour la probabilité d'avoir au moins une sur-représentation locale en une position quelconque du graphe.

Pour cela, considérons la fonction $g : (]0, +\infty])^2 \rightarrow \mathbb{R}$ définie par

$$g(\lambda, t) = \lambda((1+t)\log(1+t) - t) - \log(h(\lambda, t)). \quad (20)$$

Pour une valeur fixée de λ , la fonction $g(\lambda, \cdot)$ est une bijection croissante de $]0, +\infty[$ sur lui-même. Un Δ_U très grand, c'est-à-dire une sur-représentation locale en position U , se traduit alors par le fait que $g(\lambda_U, \Delta_U)$ est très grand.

Plus précisément, en appliquant le Théorème 4 à y tel que $g(\lambda_U, y) = t$, la fonction g permet de réécrire la borne du Théorème 4 en supprimant la dépendance en U dans le terme exponentiel.

$$\forall t > 0, \quad \mathbb{P}(g(\lambda_U, \Delta_U) \geq t) \leq \mathbb{P}(G[U] \sim \mathbf{m}')e^{-t}.$$

Il devient alors possible, en notant que $E^t = \{\max_U(g(\lambda_U, \Delta_U)) \geq t\}$ est l'union sur toutes les positions U des événements $E_U^t = \{g(\lambda_U, \Delta_U) \geq t\}$, d'obtenir une borne pour la probabilité de l'existence d'un grand thème n'importe où dans le graphe.

Théorème 5. Soit g la fonction définie par l'équation (20) et $N(\mathbf{m}')$ le nombre d'occurrences de \mathbf{m}' dans G . Alors, pour tout $t > 0$,

$$\mathbb{P}\left(\max_U(g(\lambda_U, \Delta_U)) \geq t\right) \leq \mathbb{E}N(\mathbf{m}')e^{-t} \quad (21)$$

On obtient ainsi une borne supérieure pour la p -valeur du test consistant à chercher s'il existe quelque part dans le graphe un thème d'ordre beaucoup plus grand qu'attendu.

9.3.3. Procédure de filtrage

Le Théorème 5 permet de sélectionner les schémas sur-représentés localement dans le réseau.

Cependant, considérons les deux schémas et sous-schémas respectifs de la Figure 12 et soit C_1 et C_2 les classes d'équivalence dont un sommet est supprimé. La figure permet de voir que tout (\mathbf{m}_2, C_2) -thème d'ordre k est également un (\mathbf{m}_1, C_1) -thème d'ordre $k + 1$. Dans ce cas, l'existence d'un (\mathbf{m}_2, C_2) -thème de taille significative est donc redondante avec la même information concernant les (\mathbf{m}_1, C_1) -thèmes.

Afin d'éviter un telle redondance dans la liste finale des motifs, nous considérons qu'un schéma \mathbf{m} est un motif local par rapport au sous-schéma (C, \mathbf{m}') si les deux conditions suivantes sont remplies :

1. \mathbf{m} est localement sur-représenté par rapport à (C, \mathbf{m}') , c'est-à-dire que la p -valeur donnée par le Théorème 5 est plus petite qu'un seuil fixé;
2. Soit a un sommet de \mathbf{m} appartenant à C . Il n'existe pas d'ensemble \mathcal{A} de sommets de \mathbf{m} tel que
 - il n'y a pas d'arête entre a et \mathcal{A} ,
 - $\mathbf{m} \setminus \mathcal{A}$ est localement sur-représenté par rapport à $(D, \mathbf{m} \setminus (\mathcal{A} \cup \{a\}))$, où D est la classe d'équivalence de $\{a\}$ dans $\mathbf{m} \setminus \mathcal{A}$.

S'il existe un ensemble \mathcal{A} satisfaisant les deux conditions, la sur-représentation locale de \mathbf{m} par rapport à $(C, \mathbf{m} \setminus \{a\})$ est une conséquence de la sur-représentation de $\mathbf{m} \setminus \mathcal{A}$ par rapport à $(D, \mathbf{m} \setminus (\mathcal{A} \cup \{a\}))$. Par conséquent, la paire (\mathbf{m}, C) est éliminée par la procédure de filtrage.

La Figure 14 illustre cette procédure : le premier schéma n'est pas considéré comme un motif local car les conditions de filtrage sont réalisées pour $\mathcal{A} = \{b\}$. Par contre, le second schéma est conservé à cause de l'arête entre b et a . En effet, celle-ci oblige la présence de k arêtes supplémentaires dans le réseau pour qu'il y ait un thème d'ordre k . Or, les probabilités de présence des arêtes étant faibles dans les réseaux réels, la présence de ces k arêtes est porteuse d'information.

La procédure de filtrage est plus simple à voir sur la représentation graphique des schémas et de leurs sous-schémas. En effet, soit (\mathbf{m}_1, C_1) et (\mathbf{m}_2, C_2) deux paires sélectionnées par l'application du Théorème 5, \mathbf{m}_1 ayant moins de sommets que \mathbf{m}_2 . Si \mathbf{m}_1 est un sous-graphe induit de \mathbf{m}_2 en se faisant correspondre les deux sommets carrés et si les sommets carrés ont même degré dans (\mathbf{m}_1, C_1) et (\mathbf{m}_2, C_2) , alors (\mathbf{m}_2, C_2) est filtré par la procédure.

9.3.4. Aspects algorithmiques

D'un point de vue algorithmique, l'application de cette procédure à un réseau donné et une taille de schéma k nécessite une partie énumérative des couples (schéma, sous-schéma) et de leur positions, une

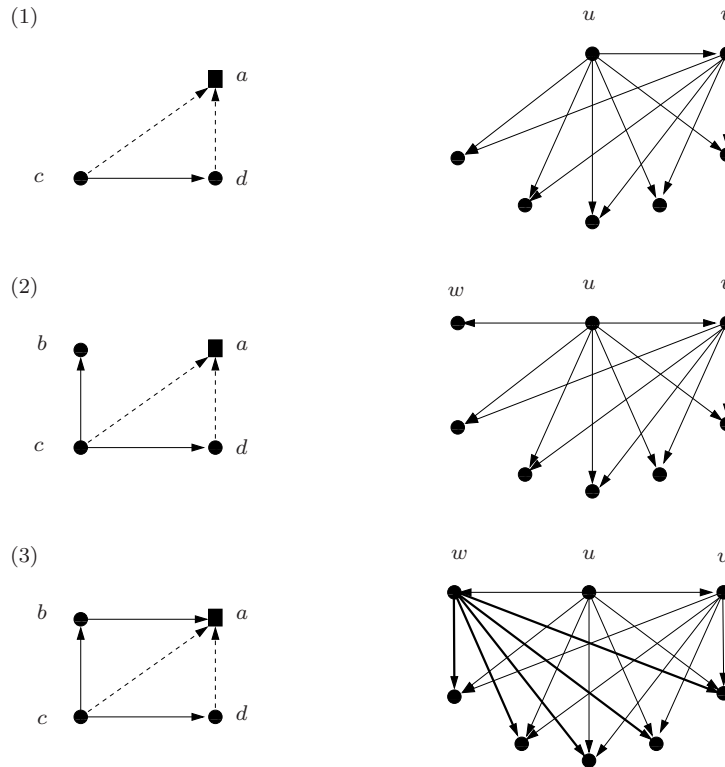


FIGURE 14. *Illustration de la procédure de filtrage. (1) On suppose que le schéma FFL est localement sur-représenté par rapport à la suppression de a . Un thème associé est situé en position $\{u, v\}$. (2) Le schéma représenté est localement sur-représenté par rapport à la suppression de a . Il n'est cependant pas ajouté à la liste des motifs car $\mathcal{A} = \{b\}$ satisfait les conditions pour le filtrage par rapport au motif FFL. (3) Le schéma représenté est également localement sur-représenté par rapport à la suppression de a . Il est considéré comme un motif local à cause de l'arête entre b et a . En effet, elle implique la présence de 5 arêtes dans le thème, ce qui est porteur d'informations dans un réseau creux.*

partie d'estimation du modèle aléatoire et une partie de calcul des p -valeurs.

L'estimation des paramètres du modèle peut être faite indépendamment du reste comme indiqué au paragraphe 9.2.

La partie énumérative peut être traitée par tout algorithme d'énumération cité en 7.3. Cependant, en raison du caractère creux des réseaux biologiques, il est plus rentable d'un point de vue du temps d'énumérer toutes les occurrences des schémas et d'en déduire ensuite les positions des sous-schémas que l'inverse. En effet, la démarche inverse implique la recherche d'un nombre potentiellement grand d'occurrences du sous-schéma qui n'ont pas d'extension.

Enfin, le calcul de la p -valeur se divise en deux parties. La première consiste à calculer, pour tout couple (schéma, sous-schéma) et toute position U listée dans la partie énumérative, la valeur λ_U correspondante. Ce coefficient correspondant à une somme simple, son calcul est rapide, de même que celui des Δ_U , g_U et $\max_U g_U$. La seconde est le calcul du terme $\mathbb{E}(N(\mathbf{m}'))$, qui est l'étape la plus pénalisante de l'algorithme. En effet, elle peut uniquement être simplifiée en regroupant les occurrences potentielles partageant la

même répartition des sommets dans les classe de degrés, ce qui implique une complexité de l'ordre de Q^{k-1} , où Q est le nombre de degrés distincts de G .

L'algorithme résultant de cette démarche est en cours d'implémentation en collaboration avec Gilles Grasseau et donnera lieu à la diffusion du package R *paloma*.

9.4. Borne inférieure de la p -valeur

La borne supérieure de la queue de distribution donné par le Théorème 5 permet de contrôler la proportion de faux-positifs. Cependant, la précision de cette borne mérite d'être étudiée pour savoir ce qu'il en est des faux-négatifs.

Pour des déviations modérées, il est possible d'évaluer la précision de la borne supérieure donnée par le Théorème 4 pour la queue de distribution en une position donnée :

Proposition 2. *Soit \mathbf{m} un schéma et (C, \mathbf{m}') un sous-schéma de \mathbf{m} . Soit U une position correspondant à une occurrence de \mathbf{m}' dans G , $\lambda_{2,U} = \sum_{v \notin U} (p_U^v)^2$ et $LB_U(t)$ la borne supérieure de $\mathbb{P}(\Delta_U \geq t)$ donnée par le Théorème 4.*

Supposons que $\lambda_{2,U} < \frac{1}{4}$. Alors, pour tout t vérifiant $1 < t < \frac{1}{8\sqrt{\lambda_{2,U}}} - 1$,

$$\frac{\mathbb{P}(\Delta_U \geq t)}{LB_U(t)} \geq \left(1 - 52 \frac{\lambda_{2,U}}{\lambda_U} (1+t)\right) \left(1 - \frac{2}{1+t}\right) \left(1 - \frac{1}{10\lambda_U(1+t)}\right)$$

Ce théorème implique que pour une suite de nombres $t^{(n)}$, de graphes $G^{(n)}$ et de positions $U^{(n)}$ tels que $t^{(n)}$ et $(t\lambda_U)^{(n)}$ tendent vers l'infini et $(t\lambda_{2,U}/\lambda_U)^{(n)}$ tend vers 0, la borne supérieure de $\mathbb{P}(\Delta_U^{(n)} \geq t^{(n)})$ est asymptotiquement atteinte.

Considérons par exemple le modèle d'Erdős-Rényi de paramètre $p^{(n)} = \frac{c}{n}$. Soit k et n les tailles respectives de \mathbf{m} et G et soit r le nombre d'arêtes présentes dans \mathbf{m} mais non dans \mathbf{m}' . Alors, pour toute position U ,

$$\lambda_U^{(n)} = (n - k + 1)p^r \sim_{n \rightarrow +\infty} \frac{c^r}{n^{r-1}} \quad (22)$$

$$\lambda_{2,U}^{(n)} = (n - k + 1)p^{2r} \sim_{n \rightarrow +\infty} \frac{c^{2r}}{n^{2r-1}} \quad (23)$$

Ainsi, pour $r \geq 2$, la suite $t^{(n)} \sim n^\alpha$ avec $r - 1 < \alpha < r$ est une suite de seuils pour lesquels la borne supérieure du théorème 4 est asymptotiquement optimale.

L'optimalité de la borne globale du Théorème 5 est plus compliquée à étudier en raison de la maximisation sur un grand nombre de positions. En notant $E^t = \{\max_U (g(\lambda_U, \Delta_U)) > t\}$ et $E_U^t = \{g(\lambda_U, \Delta_U) > t\}$, une borne inférieure pour la probabilité de E^t peut être obtenue en utilisant que

$$\mathbb{P}(E^t) \geq \sum_U \mathbb{P}(E_U^t) - \frac{1}{2} \sum_{U \neq V} \mathbb{P}(E_U^t \cap E_V^t).$$

Le terme $\sum_U \mathbb{P}(E_U^t)$ correspond à la borne supérieure donnée par le théorème. Il suffit par conséquent de borner supérieurement les probabilités des intersections $E_U^t \cap E_V^t$. Cependant, pour certains schémas, les

ordres des thèmes en deux positions chevauchantes U et V peuvent être fortement corrélés. Le schéma \mathbf{m}_2 de la Figure 12 en est un exemple : le nombre d'extensions de l'occurrence de \mathbf{m}'_2 en position $U = (\text{PDR1}, \text{PDR3}, \text{PDR5})$ est égal au nombre d'extensions de l'occurrence de \mathbf{m}'_2 en position $V = (\text{PDR1}, \text{PDR3}, \text{IPT1})$. Par conséquent, la probabilité de l'intersection $E_U^t \cap E_V^t$ n'est pas négligeable devant celle de E_U^t et E_V^t .

Cependant, cette approche simple permet de montrer que la borne supérieure est asymptotiquement optimale pour certains couples (schéma ; sous-schéma) pour un modèle d'Erdős-Rényi générant des graphes suffisamment creux.

Proposition 3. *Soit \mathbf{m} un schéma de taille k ayant un sommet a relié à tous les autres sommets de \mathbf{m} . Soit (C, \mathbf{m}') le sous-schéma où C est la classe d'équivalence de a .*

Considérons le modèle d'Erdős-Rényi de paramètre ρ avec $\rho = \mathcal{O}(n^{-\frac{1}{2}-\epsilon})$, $\epsilon > \frac{1}{2k-1}$. Soit λ et λ_2 les valeurs de λ_U et $\lambda_{2,U}$, qui ne dépendent pas de U sous ce modèle.

Soit $0 < t < n^\epsilon$ et $GB(t)$ la borne globale donnée par le Théorème 5. Alors, si $\lambda_2 < \frac{1}{4}$, et si $y > 0$ tel que $g(\lambda, y) = t$ vérifie $1 < y < \frac{1}{8\sqrt{\lambda_2}} - 1$,

$$\frac{\mathbb{P}(E^t)}{GB(t)} \geq (1 - \eta) \left(1 - 52 \frac{\lambda_2}{\lambda} (1 + y)\right) \left(1 - \frac{2}{1 + y}\right) \left(1 - \frac{1}{10\lambda(1 + y)}\right) - kn^{2k} e^{-n^\epsilon + t}$$

où $\eta = \mathcal{O}(n^{-\epsilon})$.

La condition sur ρ permet une croissance du nombre d'arêtes en $\mathcal{O}(n^{\frac{3}{2}-\epsilon})$, ce qui comprend la croissance linéaire observée sur les données réelles [29].

La combinaison des Propositions 2 et 3 implique l'optimalité asymptotique de la borne de la p -valeur dans certains cas : considérons à nouveau le modèle d'Erdős-Rényi de paramètre $p^{(n)} = \frac{c}{n}$ et un couple (schéma, sous-schéma) vérifiant les hypothèses de la Proposition 3. Pour tout $0 < \delta < \frac{1}{2}$, le choix $t^{(n)} \sim n^\delta \log(n^{k-1+\delta})$ correspond à $y^{(n)} \sim n^{k-1+\delta}$. Le premier terme du second membre de la Proposition 3 tend alors vers 1 alors que le second tend vers 0 puisque $\delta < \epsilon = \frac{1}{2}$.

9.5. Simulations et application

9.5.1. Données simulées

J'ai généré 500,000 graphes dirigés de 90 sommets, que j'appelle les *graphes de référence*. Les sommets y sont répartis en trois classes de 30 sommets et, pour toute paire de sommets (u, v) , l'arête de u vers v est présente avec probabilité 0.04 si u et v sont dans la même classe et 0.01 s'ils sont dans des classes différentes. Le degré sortant moyen et le degré entrant moyen dans ce modèle sont tous deux de 1.76.

Ma méthode est illustrée avec les schéma *FFL* et *bi-fan* (cf Figure 11). Les sous-schémas considérés sont obtenus en supprimant le sommet de degré entrant 2 dans le cas de *FFL* et l'un des sommets de degré entrant 2 pour *BIFAN*. Ce choix ne joue aucun rôle ici en raison de la symétrie du modèle.

Les parties (a) et (b) de la Figure 15 montrent les p -valeurs empiriques ainsi que leurs bornes supérieures données par le Théorème 5, en fonction de la valeur de t .

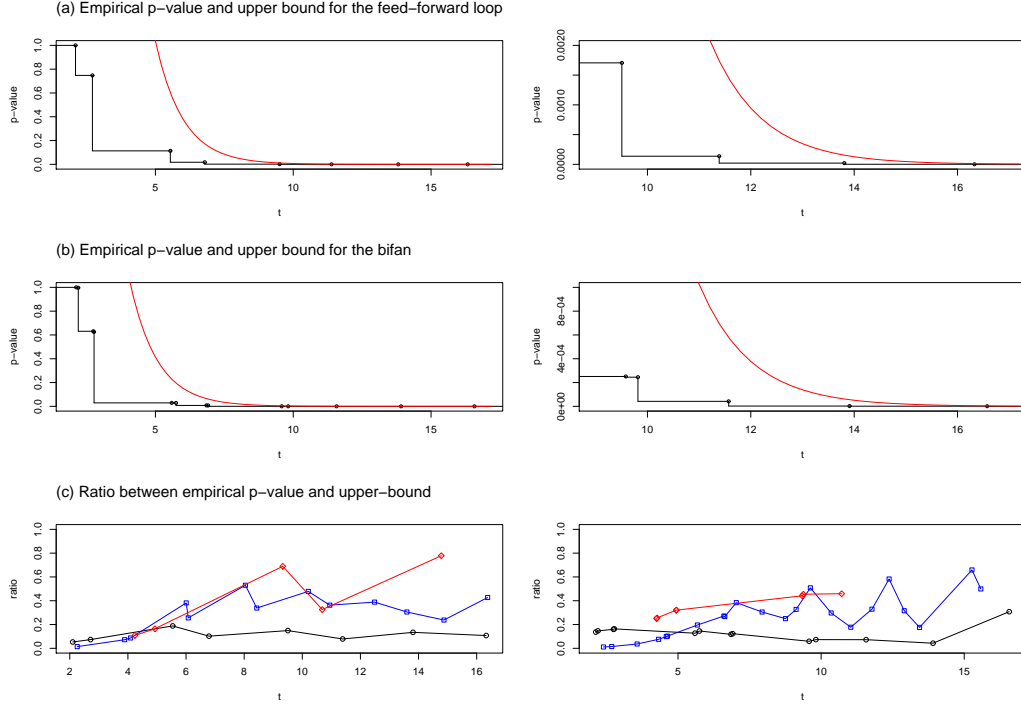


FIGURE 15. (a) p -valeurs empiriques et bornes supérieures correspondantes pour le schéma FFL dans les 500.000 graphes de référence; zoom sur la queue de distribution. (b) Idem pour le schéma BIFAN. (c) Ratio entre la p -valeur empirique et la borne supérieure pour les graphes de référence (points noirs), les graphes denses (carrés bleus) et les grands graphes (losanges rouges).

Afin d'estimer la sur-évaluation de la p -valeur, la partie (c) de la Figure 15 montre la valeur du ratio entre les deux grandeurs précédentes. Ce ratio a également été tracé pour des graphes plus denses (30.000 graphes tirés en multipliant les probabilités de connection par 5) et pour des graphes plus grands mais de mêmes degrés moyens (30.000 graphes formés de trois classes de 120 sommets chacune avec des probabilités de connection de 0.01 et 0.0025). On voit que, dans les graphes de référence, le rapport reste relativement stable en fonction de t et ce pour les deux schémas. Sa valeur est environ de 1/10. De plus, il devient meilleur quand le graphe grandit ou se densifie.

9.5.2. Influence du modèle choisi

Les p -valeurs obtenues dépendent des coefficients de la matrice $\mathbf{\Pi}$. Afin d'étudier l'influence de la méthode utilisée pour inférer cette matrice, les motifs locaux de taille 3 et 4 ont été cherchés sur un même jeu de données avec différentes méthodes d'estimation des coefficients de $\mathbf{\Pi}$:

- l'estimation correspondant au modèle d'Erdős-Rényi;
- celle correspondant au modèle *EED* décrit au chapitre 2.1. Pour rappel, la probabilité d'une arête

- allant de u vers v y est proportionnelle au produit du degré sortant de u et du degré entrant de v ;
- l’estimation bayésienne du modèle de mélange SBM proposée dans [B9] et décrite au chapitre 3 ;
 - l’estimation du modèle de mélange par l’algorithme *BLOCKS* [87] disponible au sein du logiciel *STOCNET* (<http://stat.gamma.rug.nl/stocnet/>).








	Erdős	EDD	SBM	BLOCKS
	4.3 e-46		5.7 e-7	1.3 e-6
	5.2 e-14	5.8 e-4	8.2 e-6	3.9 e-5
	1.3 e-30			
	3.2 e-14	2.9 e-5	3.1 e-8	9.6 e-5
	(4.6 e-45)	9.6 e-4	(4.1 e-9)	(1.2 e-6)
	(1.9 e-30)		3.5 e-2	3.8 e-2
	1.6 e-4			

TABLE 4

Motifs locaux trouvés suivant différentes procédures d’estimation de Π . Les valeurs indiquées sont celles données par l’inégalité (21). Celles entre parenthèses désignent les motifs potentiels ultérieurement rejetés par la procédure de filtrage.

BLOCKS ne pouvant pas traiter des graphes de plus de 200 sommets, l’application a été menée sur un sous-réseau connexe de 194 sommets du réseau de régulation de la levure. La Table 4 montre tous les motifs locaux trouvés avec une p -valeur inférieure à 0,05 pour au moins l’une des méthodes d’inférence. On peut voir que la méthode se révèle relativement stable vis-à-vis de la procédure d’estimation choisie.

La seule méthode qui se démarque des autres, tant au niveau du nombre de motifs trouvés que des p -valeurs, est celle correspondant au modèle d’Erdős-Rényi. Cela est dû à l’absence de densité locale sous ce modèle, ce qui explique la sous-évaluation des p -valeurs calculées.

On constate d’autre part que la méthode sélectionnant le moins de motifs locaux est celle associée au modèle EDD. Le premier schéma de la Table 4 n’est en particulier pas sélectionné, alors qu’il l’est pour SBM et BLOCKS. Cependant, un thème d’ordre k pour ce motif correspond à un thème d’ordre $k - 1$ pour le motif *bi-fan* (5^{ème} ligne de la Table 4), qui est sélectionné comme un motif local pour le modèle EDD. Les thèmes considérés comme significatifs par les trois méthodes se correspondent donc.

Finalement, les deux méthodes qui tiennent compte à la fois de la distribution des degrés et de la modularité du réseau trouvent les mêmes motifs locaux avec des p -valeurs proches.

9.5.3. Réseaux réels

Afin d'étudier le caractère local ou non des motifs globaux, j'ai appliqué la procédure de recherche des motifs locaux au réseau de régulation de la levure et à deux réseaux électroniques, tous étudiés au niveau des motifs globaux dans [80] et disponibles à l'adresse <http://weizmann.ac.il/mcb/UriALon>. La recherche a été effectuée pour les motifs locaux de taille 3 à 5.

La méthode d'inférence choisie est la méthode SBM bayésienne. Cependant, cette approche a un défaut qui est une tendance à créer une classe de *hubs* qui est inhomogène en termes de degrés et fausse l'interprétation. Ainsi, dans le cas du réseau de régulation de la levure, une classe est créée avec les deux sommets de degrés sortants maximaux, à savoir 71 et 44. Cette classe est alors considérée par le modèle comme générant des sommets de degré sortant moyen égal à 57,5. En raison de la présence du sommet de degré 71, tout motif en étoile sera alors sélectionné comme un motif local. Afin de pallier à cet artefact, j'ai lancé la procédure avec SBM puis je l'ai relancé avec EDD lorsque le motif sélectionné était une étoile dont le thème correspondant était centré sur le sommet de degré sortant 71.

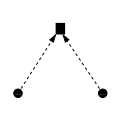
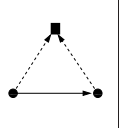
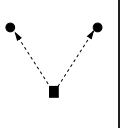
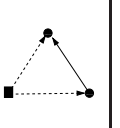
Local motif				
p -value bound	2.0 e-16	2.3 e-9	4.6 e-4	8.6 e-4
N_U^*	38	15	5	3

TABLE 5
Motifs locaux de taille 3 pour le réseau de régulation de la levure.

La Table 5 indique les motifs locaux de taille 3 trouvés pour le réseau de la levure avec un seuil de 0,001, la ligne N_U^* représentant l'ordre du thème en la position pour laquelle Δ_U est maximal.

Deux motifs apparaissent avec des p -valeurs largement inférieures au seuil. Le premier correspond à deux régulateurs co-régulant un grand nombre de gènes. Les différents algorithmes de recherche de motifs globaux ayant servi à analyser ce réseau [14, 80, 108] n'ont pas détecté ce schéma, mais ont tous sélectionné le schéma *bi-fan*. Or, tout thème d'ordre k correspondant au motif local trouvé implique l'existence de $\frac{k(k-1)}{2}$ occurrences du *bi-fan*. Les trois plus grands thèmes, qui sont respectivement d'ordre 38, 32 et 18, représentent ainsi à eux seuls 73% des occurrences *bi-fan* dans le réseau. Ceci indique que la sur-représentation globale du *bi-fan* est une conséquence de la sur-représentation locale du motif indiqué.

Le second motif local correspond au schéma *FFL*, qui est également trouvé par tous les algorithmes de recherche globaux. Il correspond d'un point de vue biologique à un régulateur initial X régulant un gène Y , X et Y co-régulant ensuite un troisième gène Z . Ce mécanisme permet d'induire différentes dynamiques dans le processus de régulation suivant le caractère inhibiteur ou activateur des régulations [6]. Cependant, ma méthode permet de mettre à jour à l'aide uniquement de méthodes statistiques l'assymétrie dans les rôles des trois gènes impliqués. Le schéma *FFL* est en effet fortement sur-représenté par rapport à la suppression de son sommet de degré entrant égal à 2. Un grand thème de ce motif correspond alors à un même couple (X, Y) co-régulant un grand nombre de gènes cibles Z_1, \dots, Z_k . Ce

phénomène a été décrit par Alon [6] sous le nom de *Multi-output feed-forward loops*.

Le schéma *FFL* apparaît encore comme localement sur-représenté par rapport à la suppression du régulateur initial X mais la p -valeur est beaucoup plus faible et le plus grand thème n'est que de taille 3. Enfin, ce schéma n'apparaît pas du tout comme sur-représenté par rapport à la suppression du gène intermédiaire Y , indiquant qu'il n'existe nulle part dans le réseau un régulateur initial X utilisant un grand nombre de gènes intermédiaires Y afin de réguler la dynamique d'un gène Z particulier.

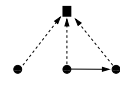
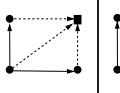
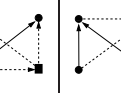
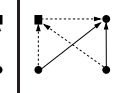

Local motif					
p -value bound	6.5 e-15	3.4 e-6	1.4 e-4	5.6 e-4	9.2 e-4
$N_{\mathcal{M}}^*$	7	2	2	1	1

TABLE 6

Motifs locaux de taille 4 pour le réseau de régulation de la levure.

La Table 6 indique les motifs de taille 4 sélectionnés. Le premier d'entre eux correspond à l'occurrence d'une *multi-output feed-forward loop* avec un troisième gène régulant 7 des sorties. Ce motif est également sélectionné par les méthodes globales mais seulement en sixième position parmi les motifs de taille 4. Les quatre autres motifs sélectionnés semblent peu pertinents dans la mesure où la taille du plus grand thème associé est de 1 ou 2. La raison de la sélection de ces motifs est donc liée à leur très faible probabilité d'apparition plutôt qu'à leur sur-représentation locale. Il en est de même pour tous les motifs locaux de taille 5 détectés. Ceci semble indiquer que les motifs locaux pertinents sont de petite taille, phénomène récemment illustré également dans le cas des motifs globaux [59].

J'ai ensuite appliqué la recherche des motifs locaux aux réseaux électroniques cités plus haut. Aucun motif local n'y est présent, pour k compris entre 2 et 5 et pour un seuil pourtant large de 0,01. Milo *et al.* [80] ont pourtant trouvé trois motifs globaux dans ces circuits. Cette constatation indique une différence de structure entre les deux types de réseaux encore plus forte que celle suggérée dans [80], dont les auteurs ne comparent que la topologie des motifs trouvés : leur répartition dans le réseau est également différente.

9.6. Perspectives

L'introduction des motifs locaux ouvre plusieurs perspectives. La première est d'ordre mathématique et consiste à adapter le travail exposé ci-dessus aux modèles de mélanges présentés au chapitre I, c'est-à-dire un modèle où les classes des sommets sont des variables latentes. Ceci rendrait l'algorithme de recherche plus efficace car la stationnarité de ces modèles permet d'écrire l'espérance du nombre d'occurrences de \mathbf{m}' comme le produit de sa probabilité d'apparition en une position donnée par le nombre de positions possibles. Cependant, l'approximation de Poisson devient plus compliquée à établir car les arêtes cessent d'être indépendantes.

La seconde perspective est celle de la comparaison de réseaux, et en particulier de leur alignement.

L'alignement des graphes d'interaction entre protéines peut en effet être un complément à la comparaison de séquence pour l'annotation fonctionnelle de gènes par similarité avec leurs orthologues [70, 113]. Le problème de l'isomorphisme de graphes étant NP-complet, il est nécessaire de passer par des heuristiques d'alignement. De grands thèmes communs entre organismes pourraient être des points de repère permettant de jouer le rôle d'ancres pour l'alignement de réseaux.

Enfin, les motifs étudiés jusqu'ici, qu'ils soient locaux ou globaux, ne sont définis que par leur topologie, c'est-à-dire par leurs arêtes. Aucune condition n'y est énoncée pour les sommets, ce qui représente une perte d'information qui peut se révéler pertinente. En effet, la question peut se poser de savoir comment sont répartis les schémas reliant uniquement certains types de noeuds. Ce problème a été abordé dans le cas des motifs globaux si l'on demande aux noeuds d'une occurrence d'appartenir à des familles données. C'est notamment le cas de l'étude de [11] qui étudient un réseau PPI dont les protéines sont étiquetées par leur annotation *GO* ou de celle de [100] qui classent les réactions d'un graphe métabolique en fonction des classifications E.C. de leurs enzymes.

La distinction entre types de sommets peut également avoir un lien étroit avec les motifs locaux. Ainsi, parmi les thèmes découverts par Zhang [114] dans les différents réseaux de la levure, certains correspondent à des couples de gènes dupliqués ayant de nombreux voisins dans le réseau.

Or, un couple de gènes partageant de nombreux voisins revient à l'existence d'un thème de taille significative pour l'un des motifs de la Figure 16. L'utilisation des scores associées aux thèmes de ces motifs pourrait servir à confirmer des paires de dupliqués inférés dans l'optique de la reconstruction des familles de dupliqués. Elle aurait en effet l'avantage de reposer sur une information biologique (le réseau d'interactions entre protéines) d'une autre nature que l'alignement de séquences habituellement utilisé pour inférer ces familles.

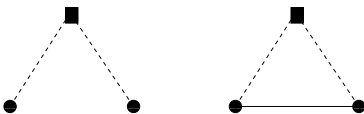


FIGURE 16. Motifs dont un thème sur-représenté dans un réseau d'interaction entre protéines correspond à deux sommets ayant de nombreux voisins communs et donc possiblement dupliqués.

D'autre part, l'analyse des thèmes impliquant des paires de dupliqués pourrait être une méthode d'étude de différents types d'évolution suite à la duplication. En effet, il est admis que durant la phase de divergence qui fait suite au phénomène de duplication, a lieu l'un des trois scénarii que sont la redondance fonctionnelle, la spécialisation ou la création d'une nouvelle fonction [88]. L'importance relative de ces trois phénomènes est cependant encore discutée [48, 53]. Une des manières de l'étudier est la comparaison des voisinages des dupliqués dans le réseau d'interaction entre protéines, qui évolue différemment suivant le scénario suivi comme le montre la Figure 17. L'utilisation de scores liés aux thèmes pourrait se révéler plus riche en informations que la distance de Czekanowski-Dice utilisée classiquement pour l'étude comparative de voisinages [23, 53].

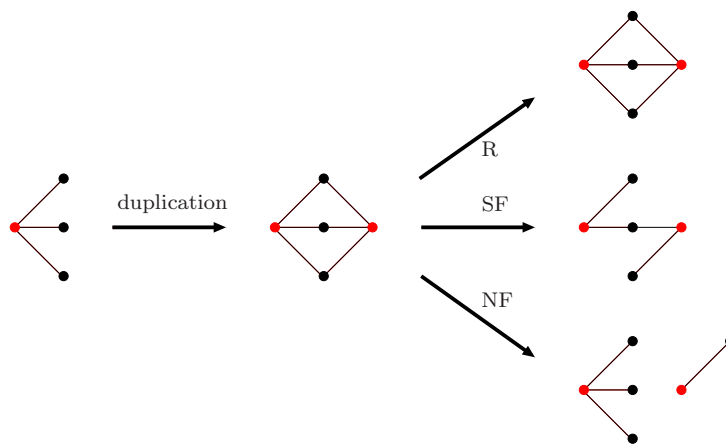


FIGURE 17. Scénarii post-duplication. *R* : redondance, *SF* : sous-fonctionalisation, *NF* : néo-fonctionalisation

Troisième partie

Cycles dans les réseaux

Le dernier chapitre de ce mémoire est consacré à l'étude de la répartition d'un autre type de sous-graphes, à savoir les cycles. Ces derniers confèrent leur complexité topologique aux graphes puisque les graphes acycliques ne sont autres que les unions disjointes d'arbres. De nombreux problèmes complexes devenant simples dans le cas des arbres, il est par conséquent souvent utile de voir dans quel mesure la structure d'un graphe quelconque peut être ramenée à une structure d'arbre afin d'adapter les algorithmes correspondants.

Je présente ici deux problèmes à priori très différents auxquels j'ai travaillé. Le premier consiste en la recherche de bornes pour la largeur d'arborescence de certaines familles de graphes non dirigés [B1,B4,B7,B11]. Le second concerne l'énumération des organisations chimiques d'un réseau métabolique [B15]. Ils reposent cependant sur une approche commune qui est la suivante :

1. Identifier la famille \mathcal{C} de cycles qui donnent au problème sa difficulté ;
2. Trouver un transversal H de \mathcal{C} , c'est-à-dire un ensemble de sommets interceptant tous les cycles de \mathcal{C} ;
3. Résoudre le problème sur le graphe privé de H puis étendre la solution au graphe entier.

10. Bornes pour la largeur d'arborescence

10.1. Largeur d'arborescence

La largeur d'arborescence a été introduite par Robertson et Seymour dans une série d'articles [96–99]. Elle repose sur la notion de décomposition arborescente, appelée suivant les domaines d'application *tree-decomposition* ou *junction tree*.

Définition 1. Une décomposition arborescente d'un graphe non orienté G est un arbre T dont les sommets sont indexés par des sous-ensembles $(W_t)_{t \in V(T)}$ de sommets de G de telle façon que :

- (C1) pour tout sommet u de G , l'ensemble des sommets de T tels que u est contenu dans W_t forme un sous-arbre T_u de T ;
- (C2) pour toute arête (u, v) de G , T_u et T_v ont une intersection non vide.

La figure 10.1 donne un exemple de décomposition arborescente.

D'un point de vue algorithmique, la structure d'arbre est intéressante dans la mesure où une solution partielle au problème donné peut être cherchée sur un des W_t jouant le rôle de racine puis étendue de proche en proche le long de l'arbre. A chaque étape, on choisit une arête (t, t') de T telle que W_t a déjà été traité mais pas $W_{t'}$ et on étend la solution partielle de W_t à $W_{t'}$. L'absence de cycle dans l'arbre T ainsi que les conditions (C1) et (C2) assurent qu'à aucune étape l'extension requise ne contredit les décisions prises précédemment.

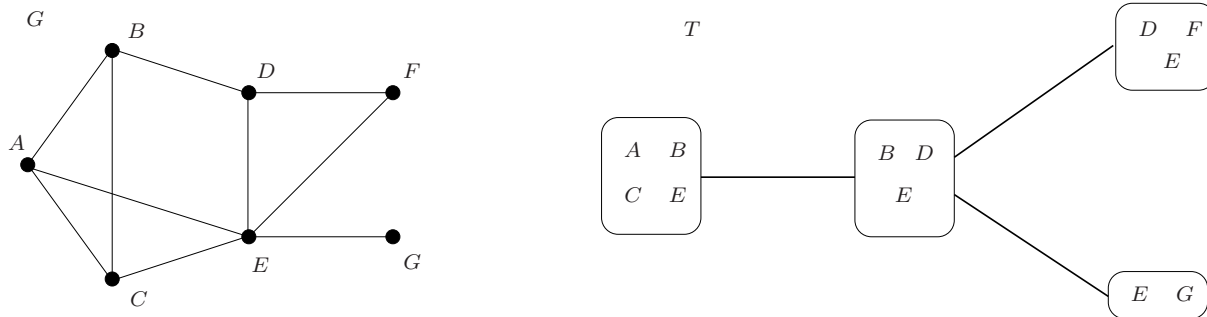


FIGURE 18. Exemple de décomposition arborescente.

Cependant, la complexité de tels algorithmes dépend de la taille des ensembles W_t sur lesquels il faut propager l'information pas à pas. Ils seront donc d'autant plus efficaces que les W_t seront petits, et leur complexité dépend de la taille maximale des W_t . Ceci explique l'introduction de la largeur d'arborescence.

Définition 2. La largeur d'une décomposition arborescente est l'entier $\max_{t \in V(T)} (|W_t| - 1)$.

La largeur d'arborescence d'un graphe G est la plus petite des largeurs de ses décompositions arborescentes. Elle est notée $TW(G)$.

Cette largeur est bien toujours définie dans la mesure où tout graphe admet au moins la décomposition triviale constituée d'un noeud indexé par l'ensemble des sommets du graphe. Le terme -1 a été ajouté dans la définition de la largeur afin que les arbres soient de largeur d'arborescence égale à 1 mais il ne joue pas d'autre rôle.

Les algorithmes pour lesquels une méthode basée sur la propagation est possible deviennent de complexité polynomiale quand on les restreint à des familles de graphes de largeur d'arborescence bornée. Ceci concerne de nombreux problèmes d'algorithmique classiques comme la recherche du nombre chromatique ou de la plus grande clique d'un graphe. Courcelle et Mosbah [32] ont montré que cette complexité devient même linéaire pour tous les problèmes pouvant s'écrire à l'aide d'une formule utilisant uniquement les opérations logiques ($\wedge, \vee, \neg, \Rightarrow$) et les quantificateurs (\exists, \forall) appliqués à des ensembles de sommets et d'arêtes ou à des tests d'appartenance ou d'adjacence.

Les décompositions arborescentes sont également utilisées pour la résolution pratique de problèmes tels que la factorisation des matrices de Cholesky, la construction d'arbres d'évolution des espèces ou la recherche de la distribution de probabilité des réseaux bayésiens [17, 62]. On se reportera à [18] ou [95] pour une bibliographie plus fournie.

10.2. Mineurs interdits

Afin de caractériser les familles de graphes ayant une largeur d'arborescence bornée, il faut introduire la notion de mineur et de graphe planaire :

Définition 3. Un graphe est planaire s'il peut être représenté dans le plan sans croisement d'arêtes.

La contraction d'une arête (u, v) consiste à supprimer u et v pour les remplacer par un nouveau sommet w dont le voisinage est l'union des voisinages de u et v .

Un graphe H est un mineur d'un graphe G s'il peut être obtenu à partir de G par suppression de sommets ou d'arêtes et/ou contraction d'arêtes (Figure 10.2).

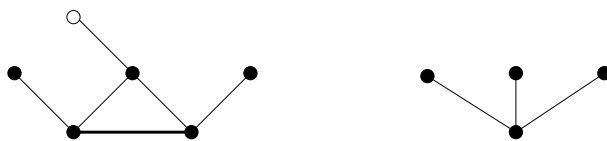


FIGURE 19. Le graphe de droite est un mineur du graphe de gauche obtenu en supprimant le sommet blanc en contractant l'arête en gras.

Robertson et Seymour [99] puis Diestel *et al.* [36] ont démontré qu'une famille de graphes \mathcal{F} est de largeur d'arborescence bornée si et seulement si il existe une famille de graphes planaires \mathcal{H} telle que tout graphe de \mathcal{F} ne contient aucun graphe de \mathcal{H} comme mineur. Les bornes démontrées dans ces articles sont cependant exponentielles en la taille des mineurs interdits. Dans l'optique d'une utilisation pratique, la question se pose de savoir s'il existe des bornes polynômiales en fonction de ces tailles.

Les articles [B1,B4,B7,B11] développent cette question. Plus précisément, ils établissent des bornes polynômiales en la taille des mineurs interdits pour les mineurs suivants :

1. le cycle de longueur k , $k \geq 3$;
2. l'union disjointe de p cycles de longueur k , $p \geq 1$, $k \geq 3$;
3. le prisme de taille l , $l \geq 3$, qui correspond à deux cycles disjoints de sommets respectifs u_1, \dots, u_l et v_1, \dots, v_l reliés par les l arêtes de type $u_i v_i$ (Figure 10.2(a));
4. la grille 3×3 (Figure 10.2(b)).

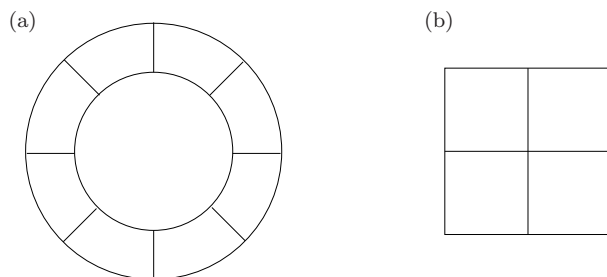


FIGURE 20. Le prisme de taille 8 et la grille 3×3

Ces travaux ont fait l'objet de ma thèse, où ils sont complétés par des bornes pour d'autres familles de cycles interdits ou des bornes plus précises quand on se restreint à certains types de graphes.

10.3. Borne polynômiale pour la largeur d'arborescence des graphes sans longs cycles

Soit k et p deux entiers, k étant au moins égal à 3. Le traitement du second cas de la liste précédente, c'est-à-dire l'interdiction du mineur composé de p cycles disjoints de longueur k , est une bonne illustration de la méthode décrite dans l'introduction de ce chapitre.

Un k -long cycle est un cycle de longueur supérieure ou égale à k . Alors, un graphe G n'admet pas l'union disjointe de p cycles de longueur k comme mineur si et seulement si il ne contient pas p k -longs cycles disjoints.

Notons \mathcal{C} la famille de tous les k -longs cycles de G . Pour tout transversal X de \mathcal{C} , on observe que

$$TW(G) \leq TW(G \setminus X) + |X| \quad (24)$$

En effet, toute décomposition arborescente de $G \setminus X$ peut être étendue en une décomposition arborescente de G en gardant le même arbre et en ajoutant tous les éléments de X à chacun des ensembles W_t .

De plus, la largeur d'arborescence de $G \setminus X$ est bornée d'après le théorème suivant [B1].

Théorème 6. *Soit $k \geq 3$ et G un graphe ne contenant aucun k -long cycle. Alors $TW(G) \leq k - 2$.*

Démonstration. Considérons un graphe G sans k -long cycle et soit T un arbre de parcours en profondeur de G . Indexons chaque sommet t de T par la liste W_t de $k - 1$ sommets composée de t et de ses $k - 2$ ancêtres directs (ou t et tous ses ancêtres si t est à distance inférieure à $k - 2$ de la racine).

On obtient ainsi une décomposition arborescente de G de largeur $k - 2$. En effet,

- (C1) T_v est composé de v et de l'ensemble de ses descendants sur $k - 2$ générations dans l'arbre T , c'est-à-dire un sous-arbre de T .
- (C2) Considérons une paire de sommets u et v reliés par une arête dans G . Alors, de par la structure des arbres en profondeur, on peut supposer que v est descendant de u . Comme G n'a pas de k -long cycle, il y a au plus $k - 1$ générations d'écart entre les deux sommets, ce qui implique que v appartient à T_u . T_u et T_v sont donc bien d'intersection non vide.

□

Il suffit donc de montrer qu'il existe un transversal X de taille polynômiale pour pouvoir conclure, ce qui est démontré dans [B4].

Théorème 7. *Soit G un graphe qui ne contient pas p k -longs cycles disjoints. Alors il existe un transversal des k -longs cycles de G de taille au plus $13k(p - 1)(p - 2) + (2k + 3)(p - 1)$.*

La démonstration de ce théorème n'est pas reproduite ici en raison de sa longueur et de son côté très technique. L'idée directrice de son raisonnement est l'utilisation du théorème de Menger [78] qui affirme que pour tous ensembles de sommets A et B et tout entier p , il existe soit p chemins disjoints entre A et B , soit un ensemble de $p - 1$ sommets interceptant tous les chemins entre A et B .

Afin de donner une illustration de l'utilisation de ce théorème, supposons le théorème 7 vrai pour $p = 2$ et démontrons qu'il est vérifié pour $p = 3$.

Soit G un graphe n'ayant pas 3 k -longs cycles disjoints. On peut supposer que G admet 2 k -longs cycles disjoints C_1 et C_2 , puisque dans le cas contraire le théorème 7 s'applique pour $p = 2$.

Supposons qu'il existe $26k$ chemins disjoints reliant C_1 et C_2 . Alors il en existe 26 dont les extrémités sont à distance au moins k le long de C_1 . De plus, Erdős et Szekeres [43] ont démontré que pour tous entiers p et q , toute suite de $(p-1)(q-1)+1$ termes contient soit une sous-suite de p termes croissants, soit une sous-suite de q termes décroissants. Il existe donc 6 chemins parmi les 26 précédents dont les extrémités sur C_1 et C_2 apparaissent dans le même ordre. L'union de C_1 , C_2 et de ces six chemins permet de construire trois k -longs cycles comme le montre la figure 10.3.

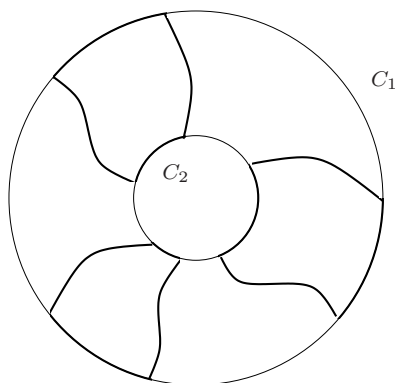


FIGURE 21. Exemple de décomposition arborescente.

Par conséquent, C_1 et C_2 ne sont pas reliés par $26k$ chemins disjoints ce qui implique, via le théorème de Menger, qu'il existe un ensemble X d'au plus $26k$ qui sépare C_1 et C_2 .

Soit A une composante connexe de $G \setminus X$. Si A contient un k -long cycle, elle n'en contient pas deux disjoints. En effet, C_1 ou C_2 ne rencontre pas A car X les sépare et on obtiendrait ainsi trois k -longs cycles disjoints. D'après le théorème 7 pour $p = 2$, il existe donc un transversal de taille $2k + 3$ pour les k -longs cycles de A .

Comme seules deux composantes de $G \setminus X$ peuvent contenir un k -long cycle à moins d'en construire trois disjoints, on obtient finalement un transversal de taille au plus $26k + 2(2k + 3)$ pour l'ensemble des k -longs cycles de G . Le théorème est donc vrai pour $p = 3$.

Les autres étapes de la récurrence sont techniquement plus compliquées à écrire mais reposent sur le même principe.

Finalement, les théorèmes 7 et 6 combinés avec l'inégalité 24 permettent de conclure que, pour tout graphe G ne contenant pas p k -longs cycles disjoints,

$$TW(G) \leq 13k(p-1)(p-2) + (2k+3)(p-1) + k - 2$$

Au moment de la construction de la décomposition arborescente, on peut s’apercevoir que certains des sommets du transversal construit dans la démonstration du théorème 7 sont superflus et affiner légèrement le résultat pour obtenir

Théorème 8. *Soit G un graphe ne contenant pas k p -longs cycles disjoints. Alors*

$$TW(G) \leq 13k(p-1)(p-2) + 3k + 1$$

10.4. Perspectives

Le but initial de ce travail, à savoir trouver une borne polynomiale en la taille du mineur interdit quelque soit la forme du mineur, semble extrêmement difficile à atteindre. De plus, de nombreuses heuristiques efficaces ont été mises au point pour déterminer des décompositions de largeur aussi faible que possible ainsi que des bornes inférieures à la largeur d’arborescence d’un graphe (<http://www.treewidth.com/> pour une bibliographie et une comparaison des différents algorithmes).

Il me semble par conséquent plus intéressant d’utiliser les décompositions arborescentes dans l’analyse des réseaux biologiques plutôt que de chercher à affiner les heuristiques existantes. Elles ont par exemple déjà été utilisées dans le cadre de la comparaison de réseaux de protéines par Dost *et al.* [40] ou d’alignement de réseaux métaboliques par Cheng *et al.* [26].

11. Cycles et réseaux métaboliques

Une collaboration récente, et toujours en cours, m’a amené à travailler sur un problème totalement différent mais où la difficulté du problème peut à nouveau être contournée à l’aide de transversaux d’un certains types de cycles. Ce travail a donné lieu à la publication [B15].

11.1. Réseaux métaboliques consistants

Un réseau métabolique est un ensemble de réactions chimiques. Il peut être modélisé par un graphe biparti comme au chapitre 5 mais également sous la forme d’un hypergraphe dirigé et pondéré $G = (M, R)$. L’ensemble M des sommets est constitué des métabolites présents dans la cellule et l’ensemble R des hyperarcs correspond aux réactions possibles. Une réaction $r \in R$ est une paire ordonnée d’ensemble de métabolites $r = (subs(r), prod(r))$, où $subs(r)$ désigne les substrats de r et $prod(r)$ ses produits.

A chaque réaction x correspond un ensemble de poids appelés coefficients stoechiométriques : tout substrat x de r se voit attribuer un coefficient négatif correspondant au nombre d’unités de x consommées par r , et tout produit a un coefficient positif correspondant au nombre d’unités produites. La *matrice stoechiométrique* S a $|M|$ lignes et $|R|$ colonnes et le coefficient $S_{x,r}$ correspond au coefficient stoechiométrique du métabolite x dans la réaction r . Un même métabolite pouvant être à la fois substrat et produit d’une même réaction, $S_{x,r}$ est alors la somme du coefficient (négatif) en tant que substrat et du coefficient en tant que produit. Il est à noter qu’en raison de cette possibilité, des hypergraphes différents peuvent avoir la même matrice stoechiométrique.

Un réseau métabolique possède des *entrées* et des *sorties*, c'est-à-dire des métabolites qu'il consomme ou produit uniquement et qui sont échangés avec le milieu. Pour des raisons de simplification des écritures mathématiques, chaque entrée est considérée comme produite par une réaction dont le substrat est l'ensemble vide. De même, chaque sortie initie une réaction donnant l'ensemble vide.

Le graphe sous-jacent de l'hyper-graphe G est le graphe dirigé dont les sommets sont les métabolites et ayant une arête \overrightarrow{xy} entre deux métabolites si il existe une réaction r tel que $x \in \text{subs}(r)$ et $y \in \text{prod}(r)$. Dans la représentation des graphes métaboliques en graphes bipartis, H est le projeté sur l'ensemble des métabolites. Par la suite, les termes *cycle* et *chemin* dans G désigneront un cycle orienté ou un chemin orienté dans H . Le fait qu'un métabolite puisse être à la fois substrat et produit implique qu'il peut y avoir des auto-arêtes dans le graphe sous-jacent. Celles-ci sont considérées comme des cycles.

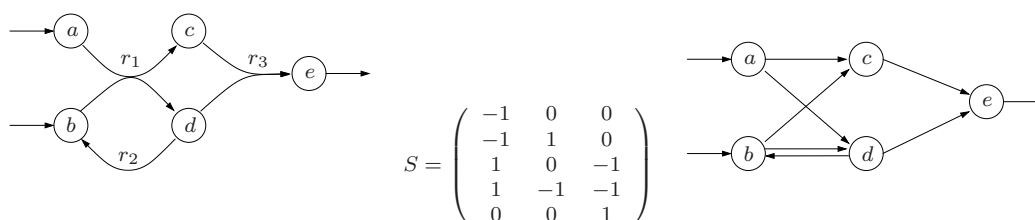


FIGURE 22. Exemple de réseau métabolique et sa matrice stoichiométrique.

Un flux dans l'hypergraphe G est un vecteur $v \in \mathbb{R}^{|R|}$ où chaque $v[i]$ représente le taux d'activation de la réaction i . Dans un réseau métabolique fonctionnant à l'équilibre, chaque métabolite est consommé autant qu'il est produit, ce qui se traduit par le fait que $Sv = 0$.

Afin de correspondre le mieux possible à la réalité, les hypergraphes considérés par la suite satisfont également deux conditions de consistance :

- ils sont consistants d'un point de vue de la masse, c'est-à-dire qu'il existe un vecteur de masse qui attribue une masse à chaque métabolite et qui est compatible avec l'équation de masse des réactions. Ceci correspond à un vecteur $m \in \mathbb{R}^{|M|}$ dont toutes les coordonnées sont strictement positives et tel que $m^T S = 0$.

L'exemple donné en Figure 11.1 est consistant en masse car un vecteur affectant une masse de 1 à a , b , c et d et une masse de 2 à e convient.

- ils sont consistants d'un point de vue des flux, c'est-à-dire qu'il existe un flux strictement positif correspondant à l'équilibre. Autrement dit, il existe un vecteur $v > 0$ tel que $Sv = 0$.

L'exemple donné en Figure 11.1 n'est pas consistant d'un point de vue du flux car, à l'équilibre, les flux à travers r_1 et r_3 doivent être égaux pour que le métabolite c soit consommé autant que produit. Or cela entraîne que le flux à travers r_2 doit être nul.

11.2. Organisations chimiques et complexité de leur énumération

Un réseau métabolique est une structure statique qui décrit les réactions chimiques ayant lieu dans une cellule, ce qui est un processus dynamique. En d'autres termes, la question se pose de savoir quels

sont les sous-ensembles de réactions pouvant constituer le réseau actif à un moment donné.

L'approche la plus courante de cette question est l'étude des *modes élémentaires*, c'est-à-dire des ensembles minimaux sur lesquels peut être défini un flux assurant un fonctionnement à l'équilibre [74, 101]. En effet, tout flux correspondant à un état d'équilibre peut s'écrire comme une combinaison linéaire des flux des modes élémentaires. Le fait qu'un réseau est consistant du point de vue du flux revient alors à dire qu'il existe pour toute réaction un mode élémentaire la contenant. Cependant, les modes élémentaires ne permettent pas d'étudier le passage d'un état d'équilibre à un autre, phase pendant laquelle une accumulation de métabolites est possible. Nous nous sommes donc intéressés au concept d'organisation chimique, introduit par Dittrich et di Fenizio [37] à partir des travaux de Fontana et Buss [45].

Définition 4. *Un ensemble $C \subseteq M$ est fermé si, pour toute réaction $r \in \mathcal{R}_C$, $\text{subs}(r) \subseteq C$ entraîne $\text{prod}(r) \subseteq C$.*

Un ensemble de métabolites C est auto-subsistant s'il existe un vecteur de flux v tel que :

1. *pour toute réaction $r \in \mathcal{R}_C$, $v[r] > 0$;*
2. *pour toute réaction $r \notin \mathcal{R}_C$, $v[r] = 0$;*
3. *pour tout métabolite $x \in C$, $(Sv)[x] > 0$.*

Un ensemble de métabolites C est une organisation si il est fermé et auto-subsistant.

Une organisation est réactive si tout métabolite participe à au moins une réaction en tant que substrat ou produit. Elle est connexe si le graphe orienté sous-jacent est connexe.

L'auto-subsistance empêche qu'il y ait un métabolite qui soit plus consommé que produit, mais permet par contre une production supérieure à la consommation. De même, elle entraîne que toute réaction dont tous les substrats sont présents a bien lieu. Les produits ainsi générés font bien partie de l'organisation de par son caractère fermé.

Il est à noter que les entrées du réseau étant produites à partir l'ensemble vide, tout ensemble fermé et donc toute organisation les contient.

Dittrich et di Fenizio [37] ont démontré que déterminer si un réseau contient une organisation est NP-complet. Cependant, le réseau utilisé pour réduire le problème à partir de 3-SAT n'est consistant ni du point de vue de la masse ni du point de vue du flux.

Centler *et al.* [24] ont développé deux algorithmes pour énumérer les organisations. Le premier énumère les ensembles fermés puis vérifie leur auto-suffisance, le second énumère les ensembles auto-suffisants puis vérifie leur caractère fermé.

Les réseaux métaboliques réels étant consistants, la question se pose de savoir si le problème devient polynomial quand on les restreint aux réseaux consistants et si un algorithme plus efficace peut être mis au point.

Pour tout ensemble $C \subseteq M$, il existe un plus petit ensemble fermé contenant C . Cet ensemble, noté Cl_C , est appelé *fermeture* de C . La fermeture de C peut être obtenue à partir de C par *propagation*, c'est-à-dire en ajoutant, tant que possible, les produits des réactions dont les substrats sont dans l'ensemble courant.

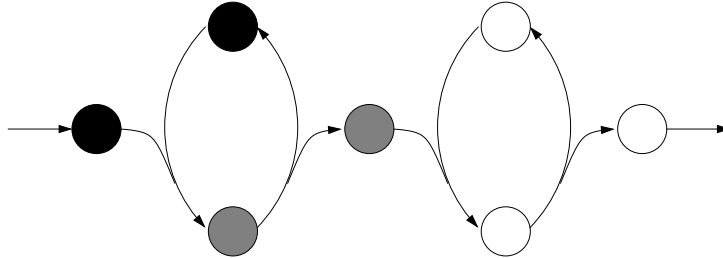


FIGURE 23. La propagation depuis l'ensemble des sommets noirs sélectionne les deux sommets gris. Elle est par contre stoppée par le second cycle.

La réponse à la question de savoir s'il existe une organisation dans un réseau consistant devient triviale car le réseau entier lui-même est une organisation. Plus précisément, il existe deux organisations, éventuellement confondues.

Proposition 4. *Soit G un réseau métabolique consistant d'un point de vue du flux. Alors le réseau entier et la fermeture de l'ensemble vide sont des organisations.*

Par la suite, ces deux organisations sont appelées les organisations triviales du réseau.

Démonstration. Le réseau entier est une organisation car il est forcément fermé et un flux garantissant l'auto-subsistance peut être obtenu en additionnant les flux correspondants à tous les modes élémentaires du réseau.

La fermeture de l'ensemble vide est également auto-subsistante car tout flux sur cet ensemble peut être augmenté autant que nécessaire puisque les entrées sont disponibles en quantité illimitée. \square

Le problème de l'existence d'une organisation non triviale et, par conséquent, de l'énumération, reste cependant non polynomial comme le montre le théorème suivant.

Théorème 9. *Décider si un réseau consistant du point de vue du flux et de la masse contient une organisation non triviale est NP-complet.*

La démonstration de ce résultat repose sur une réduction de 3-SAT. Plus précisément, étant donnée une formule logique sous forme CNF d'ordre 3, il est possible de construire explicitement un réseau consistant du point de vue du flux et de la masse dont les organisations sont en bijection avec les ensembles de littéraux satisfaisant toutes les clauses. Cette construction est détaillée dans l'annexe de [B15].

11.3. Cycles bloquants

Considérons un réseau acyclique, c'est-à-dire que le graphe orienté sous-jacent ne contient pas de cycle orienté. Alors la procédure de propagation à partir de l'ensemble vide génère tout le réseau et la seule organisation est le réseau entier. Par conséquent, la complexité de l'énumération mise en lumière par le Théorème 9 provient des cycles du réseau.

Cependant, tout cycle ne représente pas un obstacle à la propagation comme le montre la Figure 11.3(a).

Un *cycle potentiellement bloquant* est un cycle tel qu'il existe une réaction r de substrats $\{s_1, \dots, s_h\}$ et de produits $\{p_1, \dots, p_k\}$ satisfaisant les conditions suivantes :

1. il existe i et j tels que (s_i, p_j) est une arête du cycle ;
2. il existe l tel que s_l n'est pas dans le cycle.

Un tel cycle est dit *potentiellement bloquant* car il peut interrompre ou non la propagation suivant la configuration du réseau, comme le montre la Figure 11.3(b) et (c).

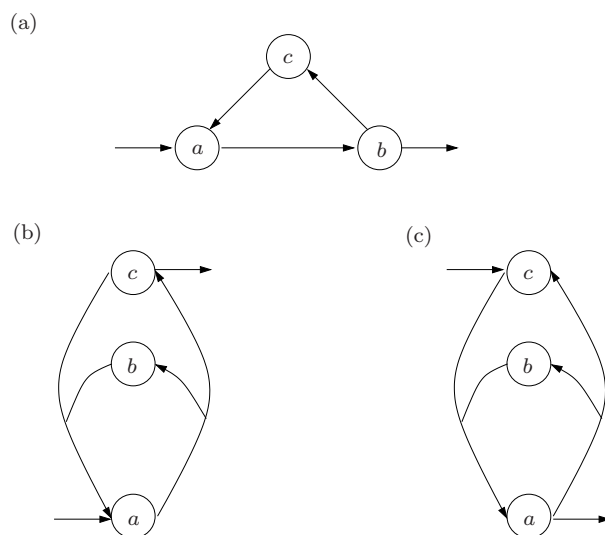


FIGURE 24. (a) le cycle $abca$ n'est pas potentiellement bloquant. (b) Le cycle potentiellement bloquant aca ne bloque pas la propagation. (c) Le cycle potentiellement bloquant aca bloque la propagation.

Théorème 10. Soit H un transversal de l'ensemble des cycles potentiellement bloquants d'un hypergraphe dirigé G . Alors l'ensemble \mathcal{O} des organisations réactives et connexes de G vérifie

$$\mathcal{O} \subset \bigcup_{C \subset H} \{Cl_C\}$$

Démonstration. Soit A une organisation réactive et connexe et $C = A \cap H$. Il suffit alors de montrer que $A = Cl_C$. A étant fermé, il est clair que $Cl_C \subseteq A$. Supposons que A contient des métabolites qui ne sont pas dans Cl_C . On les colore en blanc, et ceux de Cl_C en noir.

A étant une organisation connexe, il existe des métabolites blancs $\{x_i\}_{1 \leq i \leq k}$ qui sont les produits de réactions ayant au moins un substrat noir. Pour chaque métabolite x_i , soit r_i une telle réaction. Alors r_i a au moins un substrat blanc noté w_i , car sinon x_i serait noir par fermeture.

S'il existe un indice i tel que $x_i = w_i$, il y a une auto-arête sur le x_i qui forme un cycle potentiellement bloquant avec la réaction r_i . Comme x_i n'est pas contenu dans H , ceci contredit le fait que H est un transversal. On peut donc supposer chaque w_i distinct de x_i .

Pour tout i , on définit alors T_i comme l'union de w_i et de l'ensemble des métabolites blancs w tels qu'il existe un chemin blanc de w à w_i . On peut de plus supposer que T_1 est de cardinal minimal parmi les T_i . Comme T_1 est non vide et qu'il existe un flux strictement positif sur toutes les réactions de A , il existe un chemin reliant une des entrées du réseau à T_1 . Or les entrées sont toutes des métabolites noirs donc T_1 contient forcément l'un des x_i .

Si T_1 contient x_1 , il existe un chemin blanc de x_1 à w_1 . En y ajoutant l'arête $\overrightarrow{w_1x_1}$ et le fait que r_1 a un substrat noir, on obtient un cycle potentiellement bloquant composé uniquement de sommets blancs. Ceci contredit le fait que H est un transversal.

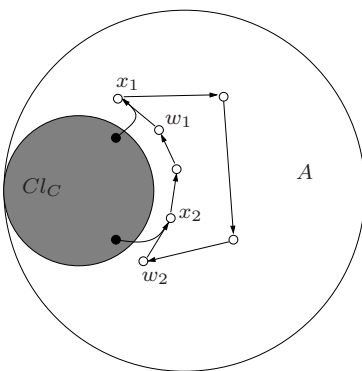


FIGURE 25. Construction d'un cycle blanc potentiellement bloquant qui contredit la transversalité de H .

Sinon, on peut supposer que T_1 contient x_2 . Il contient alors aussi w_2 puis T_2 dans son ensemble. Par minimalité du cardinal de T_1 , on obtient alors $T_1 = T_2$. Ainsi, il existe un chemin blanc de x_2 à w_1 car $x_2 \in T_1$ et un chemin blanc de w_1 à w_2 car $w_1 \in T_1 = T_2$. En d'autres termes, il existe un chemin blanc de x_2 à w_2 , qui peut être complété en un cycle potentiellement bloquant et entièrement blanc par l'ajout de $\overrightarrow{w_2x_2}$. On obtient donc à nouveau une contradiction avec le fait que H est un transversal des cycles potentiellement bloquants.

□

On obtient ainsi un ensemble d'au plus $2^{|H|}$ ensembles fermés dont il faut vérifier s'il sont auto-subsistants afin d'énumérer toutes les organisations. Cette borne de $2^{|H|}$ est optimale, même dans le cas des organisations réactives et connexes dans un réseau consistant, comme le montre l'exemple de la Figure 11.3.

11.4. Algorithme

L'approche par cycles potentiellement bloquants développé au paragraphe précédent permet de mettre au point un algorithme qui effectue un premier tri parmi les ensembles fermés du réseau. L'algorithme de programmation linéaire nécessaire pour vérifier l'auto-suffisance de l'ensemble est ainsi appliqué aux

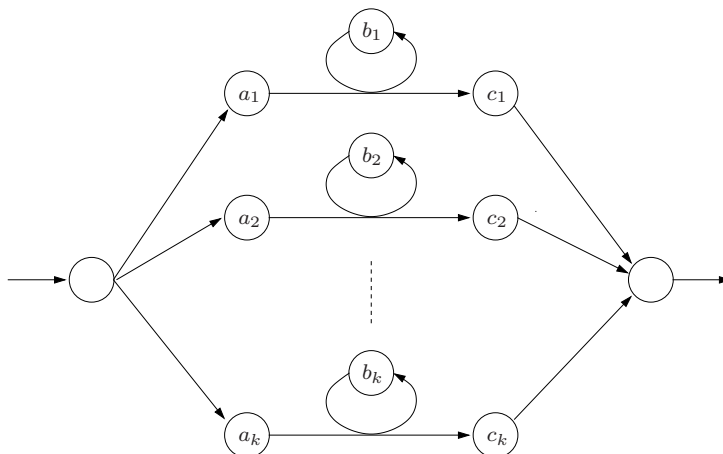


FIGURE 26. Le plus petit transversal pour les cycles de ce réseau est l'ensemble des $(b_i)_{1 \leq i \leq k}$ qui est de taille k . Par ailleurs, le réseau comprend 2^k organisations puisque y construire une organisation revient à décider pour chacun des couples $(b_i, c_i)_{1 \leq i \leq k}$ s'il est compris ou non.

ensembles décrits par le Théorème 10 plutôt qu'à tous les ensembles fermés. Cependant, trouver un ensemble transversal minimal pour les cycles d'un graphe dirigé est un problème NP-complet car équivalent au problème du *Feedback vertex set* [66]. Des algorithmes d'approximation tels que celui développé par Seymour [102] peuvent être utilisés pour réaliser cette étape.

Cependant, dans la mesure où tout cycle potentiellement bloquant ne se révèle pas forcément bloquant en pratique (cf Figure 11.3), il est plus efficace de s'intéresser à une version locale de notre approche. Celle-ci consiste à appliquer la procédure de propagation jusqu'au blocage, à chercher un transversal des cycles bloquant effectivement cette procédure puis à débloquent successivement tout sous-ensemble de ces cycles.

Pour cela, considérons un ensemble C de métabolites, obtenu en appliquant la procédure de propagation. Un *cycle C-bloquant* est un cycle dont les sommets sont extérieurs à C et qui contient une réaction dont l'un des substrats est dans C .

Considérons alors l'algorithme 11.4, consistant à ne chercher des transversaux que lorsque la propagation est réellement bloquée. Il permet de réduire les ensembles de cycles dont on cherche des transversaux puisque les cycles sont traités au fur et à mesure qu'il bloquent la propagation et que les cycles potentiellement bloquants traversés par la propagation sont ignorés.

On est cependant certain du fait que toute organisation réactive et connexe fait partie de l'ensemble d'organisations potentielles \mathcal{CO} retournée par l'algorithme grâce au théorème suivant :

Théorème 11. *Soit C un ensemble fermé et H un transversal des cycles C -bloquants. Soit A une organisation réactive et connexe contenant C . Alors $C = A$ ou il existe un sous-ensemble non-vide B de H tel que $C|_{C \cup B} \subseteq C$.*

Data: Un graphe métabolique sous forme d'hypergraphe dirigé
Result: Une liste \mathcal{CO} d'organisations potentielles
 $\mathcal{CO} \leftarrow Cl_{\{\}} ;$
for *tout élément de \mathcal{CO} non traité encore* **do**
 | Déterminer un transversal H des cycles C -bloquants ;
 for *tout $B \subseteq H$* **do**
 | Déterminer $Cl_{C \cup B}$ et l'ajouter à \mathcal{CO} s'il n'y est pas déjà;
 end
end

Ce théorème se démontre de la même façon que le Théorème 10. Pour toute organisation réactive et connexe A , son application à un élément maximal C de \mathcal{CO} inclus dans A entraîne que $C = A$ et donc que $A \in \mathcal{CO}$.

Cette méthode donnera de meilleurs résultats que l'approche d'énumération locale de tous les ensembles fermés développée par Centler *et al.* [24]. En effet, les ensembles potentiels retenus sont un sous-ensemble des ensembles potentiels de cet algorithme. L'implémentation de notre méthode et son application à des données réelles sont en cours.

11.5. Perspectives

Concernant les organisations, la question se pose de savoir s'il est possible, selon un schéma similaire, d'énumérer plutôt les ensembles auto-suffisants pour vérifier ensuite leur caractère fermé. Outre la simplicité de cette dernière vérification, cela permettrait d'étudier, tout en tenant compte de la stoechiométrie, le problème des précurseurs, c'est-à-dire les ensembles minimaux d'entrées nécessaires à produire un sous-ensemble donné de métabolites cibles [31].

D'un point de vue plus large, l'étude de la structure des cycles se révèle cruciale dans d'autres problèmes liés aux réseaux biologiques. Au sein du même groupe de travail, nous avons notamment commencé la recherche de ce que nous nommons des *histoires métaboliques*, c'est-à-dire des ensembles de réactions dont l'activation permet d'expliquer l'évolution des métabolites dont la concentration est mesurée. Dans ce cadre, les cycles, et plus précisément les ensembles transversaux d'arêtes, jouent un rôle déterminant.

Références

- [1] AIELLO, W., CHUNG, F. and LU, L. (2000). A random graph model for massive graphs. In *Proceedings of the Thirtysecond Annual ACM Symposium on Theory of Computing* 171-180.
- [2] AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2006). Mixed membership stochastic block models for relational data with application to protein–protein interactions. In *Proceedings of the International Biometrics Society Annual Meeting*.
- [3] AIROLDI, E., BLEI, D., FIENBERG, S. and XING, E. (2008). Mixed membership stochastic block-models. *Journal of Machine Learning Research* **9** 1981-2014.
- [4] ALBERT, R. and BARABÁSI, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* **74** 47-97.
- [5] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* **37** 3099-3132.
- [6] ALON, U. (2007). Network motifs : theory and experimental approaches. *Nature Reviews Genetics* **8** 450-461.
- [7] ALON, N., DAO, P., HAJIRASOULIHA, I., HORMOZDIARI, F. and SAHINALP, S. C. (2008). Biomolecular network motif counting and discovery by color coding. *Bioinformatics* **24** 241-249.
- [8] ARTZY-RANDRUP, Y., FLEISHMAN, S. J., BEN-TAL, N. and STONE, L. (2004). Comment on "Network Motifs : Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and Designed Networks". *Science* **305**.
- [9] AZUMA, K. (1967). Weighted sums of certain dependant random variables. *Tokoku Math. Journal* **19** 357-367.
- [10] BABU, M. M., LUSCOMBE, N. M., ARAVIND, L., GERSTEIN, M. and TEICHMANN, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology* **14** 283 - 291.
- [11] BANKS, E., NABIEVA, E., CHAZELLE, B. and SINGH, M. (2008). Organization of Physical Interactomes as Uncovered by Network Schemas. *PLoS Comput. Biol.* **4** e1000203.
- [12] BARABÁSI, A. L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509-512.
- [13] BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson approximation*. Oxford University Press.
- [14] BERG, J. and LÄSSIG, M. (2004). Local graph alignment and motif search in biological networks. *Proc. Nat. Acad. Sci.* **101** 14689-14694.
- [15] BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel* **7** 719-725.
- [16] BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2010). Exact and monte carlo calculations of integrated likelihoods for the Latent Class Model. *Journal of Statistical Planning and Inference* **140** 2991-3002.
- [17] BISHOP, C. M. (2006). *Pattern Recognition and MACHine Learning*. Springer.
- [18] BODLAENDER, H. L. (1993). A tourist guide through tree-width. *Acta Cybernetica* **11** 1-23.

- [19] BOER, P., HUISMAN, M., SNIJDERS, T. A. B., STEGLICH, C. E. G., WICHERS, L. H. Y. and ZEGGELINK, E. P. H. (2006). StOCNET : an open software system for the advanced statistical analysis of social networks Version 1.7.
- [20] BOLLOBÁS, B. (2001). *Random Graphs*. Cambridge University Press.
- [21] BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2007). The phase transition in inhomogeneous random graphs. *Random structures and algorithms* **31** 3-122.
- [22] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2003). Concentration inequalities using the entropy method. *Ann. Probab.* **31** 1583-1614.
- [23] BRUN, C., HERRMANN, C. and GUÉNOCHE, A. (2004). Clustering proteins from interaction networks for the prediction of cellular functions. *Proc. Nat. Acad. Sci.* **101** 14689-14694.
- [24] CENTLER, F., KALETA, C., DI FENIZIO, P. S. and DITTRICH, P. (2008). Computing chemical organizations in biological networks. *Bioinformatics* **24** 1611-1618.
- [25] CHEN, L. H. Y. (1975). Poisson approximation for dependant trials. *Ann. Probab.* **3** 534-545.
- [26] CHENG, Q., BERMAN, P., HARRISON, R. and ZELIKOVSKY, A. (2010). Efficient Alignments of Metabolic Networks with Bounded Treewidth. In *Proc. 10th IEEE Intl Conf on Data Mining (ICDM)* 687-694.
- [27] CHUNG, F. and LU, L. (2001). The diameter of random sparse graphs. *Adv. in Appl. Math.* **26** 257-279.
- [28] CHUNG, F. and LU, L. (2002). Connected components in random graphs with given degree sequence. *Annals of Combinatorics* **6** 125-145.
- [29] CHUNG, F. and LU, L. (2006). *Complex Graphs and Networks (CBMS Regional Conference Series in Mathematics)*. AMS.
- [30] CHUNG, F., LU, L., DEWEY, G., , and GALAS, D. J. (2003). Duplication models for biological networks. *Journal of Computational Biology* **10** 677-688.
- [31] COTTRET, L., MILREU, P. V., ACUNA, V., MARCHETTI-SPACCAMELA, A., MARTINEZ, F. V., SAGOT, M. F. and STOUGIE, L. (2008). Enumerating Precursor Sets of Target Metabolites in a Metabolic Network. In *Workshop on Algorithms in Bioinformatics (WABI'08)* (VOLUME 5251 OF LECTURE NOTES IN BIOINFORMATICS, ed.) 233-244. Springer Verlag Berlin.
- [32] COURCELLE, B. and MOSBAH, M. (1993). Monadic second-order evaluations on tree-decomposable graphs. *Theor. Comp. Sc.* **109** 49-82.
- [33] DAUDIN, J. J., PICARD, F. and ROBIN, S. (2008). Mixture model for random graphs. *Stat. Comput.* **18** 173-183.
- [34] DAVIS, J. A. and LEINHARDT, S. (1972). *Sociological Theories in Progress, Volume 2* The Structure of Positive Interpersonal Relations in Small Groups 218-251. Boston : Houghton Mifflin.
- [35] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **B39** 1-38.
- [36] DIESTEL, R., GORBUNOV, K. Y., JENSEN, T. R. and THOMASSEN, C. (1999). Highly connected sets and the excluded grid theorem. *J. Combin. Theory Ser. B* **75** 61-73.
- [37] DITTRICH, P. and DI FENIZIO, P. S. (2007). Chemical organization theory. *Bull. Math. Biol.* **69**

- 1199-1231.
- [38] DOBRIN, R., BEG, Q. K., BARABÁSI, A. L. and OLTVAI, Z. N. (2004). Aggregation of topological motifs in *Escherichia Coli* transcriptional regulatory network. *BMC Bioinformatics* **5** 10.
 - [39] DOGOROVTSEV, S. and MENDES, J. F. (2001). Effect of the accelerating growth of communications networks on their structure. *Phys. Rev. E* **63** 025101.
 - [40] DOST, B., SHLOMI, T., GUPTA, N., RUPPIN, E., BAFNA, V. and SHARAN, R. (2008). QNet : A Tool for Querying Protein Interaction Networks. *Journal of Computational Biology* **15** 913-925.
 - [41] ELATI, M., NEUVIAL, P., BOLOTIN-FUKUHARA, M., BARILLOT, E., RADVANYI, F. and ROUVEIROL, C. (2007). LICORN : learning co-operative regulation networks from expression data. *Bioinformatics* **23** 2407-2414.
 - [42] ERDŐS, P. and RÉNYI, A. (1959). On random graphs I. *Publ. Math. Debrecen* **6** 290-297.
 - [43] ERDŐS, P. and SZEKERES, G. (1935). A combinatorial problem in geometry. *Compositio Math.* **2** 463-470.
 - [44] FIENBERG, S. E. and WASSERMAN, S. (1981). Categorical data analysis of single sociometric relations. *Sociological Methodology* **12** 156-192.
 - [45] FONTANA, W. and BUSS, L. (1994). The arival of the fittest : towards a theory of biological organization. *J. Biol. Syst.* **2** 165-182.
 - [46] FRANK, O. and HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association* **77** 835-840.
 - [47] FU, Q. and BANERJEE, A. (2008). Multiplicative mixture models for overlapping clustering. In *Proceedings of the IEEE International Conference on Data Mining* 791-796.
 - [48] GIBSON, T. and GOLDBERG, D. S. (2009). Questioning the ubiquity of neofunctionalization. *PLOS Computational Biology* **5** e1000252.
 - [49] GROCHOW, J. A. and KELLIS, M. (2007). Network motif discovery using subgraph enumeration and symmetry-breaking. In *RECOMB 2007, Lecture Notes in Computer Science* (SPRINGER, ed.) **4453** 92-106.
 - [50] GUILLAUME, J. L. and LATAPY, M. (2004). Bipartite structure of *all* complex networks. *Information Processing Letters* **90** 215-221.
 - [51] GUIMERA, R., SALES-PARDO, M. and AMARAL, L. A. N. (2006). Classes of complex networks defined by role-to-role connectivity profiles. *Nature Physics* **3** 63-69.
 - [52] HATHAWAY, R. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters* **4** 53-56.
 - [53] HE, X. and ZHANG, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. **169** 1157-1164.
 - [54] HELLER, K. and GHAHRAMANI, Z. (2007). A nonparametric bayesian approach to modeling overlapping clusters. In *Proceedings of the 11th international conference on AI and statistics*.
 - [55] HELLER, K., WILLIAMSON, S. and GHAHRAMANI, Z. (2008). Statistical models for partial membership. In *Proceedings of the 25th International Conference on Machine Learning* 392-399.
 - [56] HOFMAN, J. and WIGGINS, C. (2008). Bayesian approach to network modularity. *Phys. Rev. Lett.*

100.

- [57] HOLLAND, P., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels : some first steps. *Social Networks* **5** 109-137.
- [58] JAAKKOLA, T. S. and JORDAN, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10** 25-37.
- [59] JAMAKOVIC, A., MAHADEVAN, P., VAHDAT, A., BOGUÑÁ, M. and KRIOUKOV, D. V. (2009). How small are building blocks of complex networks. *arXiv :0908.1143v1*.
- [60] JEFFERY, C. J. (1999). Moonlighting proteins. *Trends in Biochemical Sciences* **24** 8-11.
- [61] JEFFREYS, H. (1946). An invariant form for the prior probability in estimations problems. In *Proceedings of the Royal Society of London. Series A* **186** 453-461.
- [62] JENSEN, F. V., LAURITZEN, S. L. and OLESEN, K. G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly* **4** 92-114.
- [63] JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. and BARABASI, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407** 651-654.
- [64] JEONG, H., MASON, S. P., BARABÁSI, A. L. and OLTVAI, Z. (2001). Lethality and centrality in protein networks. *Nature* **411** 41-42.
- [65] KAPLAN, S., BREN, A., DEKEL, E. and ALON, U. (2008). The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Molecular Systems Biology* **4**.
- [66] KARP, R. M. (1972). *Complexity of Computer Computations* Reducibility among combinatorial problems 218-251. Plenum Press.
- [67] KASHANI, Z. R. M., AHRABIAN, H., ELAHI, E., NOWZARI-DALINI, A., ANSARI, E. S., ASADI, S., MOHAMMADI, S., SCHREIBER, F. and MASOUDI-NEJAD, A. (2009). Kavosh : a new algorithm for finding network motifs. *BMC Bioinformatics* **10**.
- [68] KASHTAN, N., ITZKOVITZ, S., MILO, R. and ALON, U. (2004). Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* **20-11** 1746.
- [69] KLEINBERG, J., KUMAR, S. R., RAPHAVAN, P., RAJAGOPALAN, S. and TOMKINS, A. (1999). The web as a graph : Measurements, models and methods. In *Proc. COCOON* (J. TOKYO, ed.) 1-17. Springer.
- [70] KOLÁR, M., LÄSSIG, M. and BERG, J. (2008). From protein interactions to functional annotation : graph alignment in *Herpes*. *BMC Systems Biology* **2**.
- [71] KRAPIVSKY, P. L. and REDNER, S. (2001). Organization in growing random networks. *Phys. Rev. E* **63** 066123.
- [72] LACROIX, V., FERNANDES, C. G. and SAGOT, M. F. (2005). Motif search in graphs : application to metabolic networks. **3** 360-368.
- [73] LACROIX, V., FERNANDES, C. G. and SAGOT, M. F. (2006). Motif search in graphs : application to metabolic networks. *Transactions in Computational Biology and Bioinformatics* **3** 360-368.
- [74] LACROIX, V., COTTRET, L., THÉBAUT, P. and SAGOT, M. F. (2008). An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5** 594-617.

- [75] LATOUCHE, P. (2010). *Modèles de graphes aléatoires à structure cachée pour l'analyse des réseaux*. Université d'Evry-Val-d'Essonne.
- [76] MARIADASSOU, M., ROBIN, S. and VACHER, C. (2010). Uncovering latent structure in valued graphs : a variational approach. *Annals of Applied Statistics* **4**.
- [77] MCDIARMID, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics* (J. R.-A. M. HABIB, C. MCDIARMID and B. REED, eds.) 195-248. Springer.
- [78] MENGER, K. (1927). Zur allgemeinen Kurventheorie. *Fundamenta Mathematicae* **10** 96-115.
- [79] MIDDENDORF, M., ZIV, E. and WIGGINS, C. H. (2005). Inferring network mechanisms : the *Drosophila megalonaster* protein interaction network. *PNAS* **102** 3192-3197.
- [80] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network Motifs : Simple Building Blocks of Complex Networks. *Science* **298** 824-827.
- [81] MITHANI, A., PRESTON, G. M. and HEIN, J. (2009). A stochastic model for the evolution of metabolic networks with neighbor dependence. *Bioinformatics* **25** 1528-1535.
- [82] MOLLOY, M. and REED, B. (1998). The size of the giant component of a random graph with given degree sequence. *Combin. Probab. Comput* **7** 295-305.
- [83] MOLLOY, M. and REED, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms* **6** 161-179.
- [84] NEWMAN, M. E. J. (2003). The structure and function of complex networks. *SIREV* **45** 167-256.
- [85] NEWMAN, M. and LEICHT, E. (2007). Mixture models and exploratory analysis in networks. In *Proceedings of the National Academy of Sciences* **104** 9564-9569.
- [86] NEWMAN, M. E. J., WATTS, D. J. and STROGATZ, S. H. (2001). Random graphs with arbitrary degree distributions and their applications. *Phys. Rev.* **64**.
- [87] NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *JASA* **96** 1077-87.
- [88] OHNO, S. (1970). *Gene duplication*. New York : Springer.
- [89] OMIDI, S., SCHREIBER, F. and MASOUDI-NEJAD, A. (2009). MODA : an efficient algorithm for network motif discovery in biological networks. *Genes Genet. Syst.* **84** 385-395.
- [90] PALLA, G., DERENYI, I., FARKAS, I. and VICSEK, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** 814-818.
- [91] PALLA, G., DERENYI, I., FARKAS, I. and VICSEK, T. (2006). CFinder, the community cluster finding program Version 2.0.1.
- [92] PICARD, F., DAUDIN, J. J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2008). Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15** 1-20.
- [93] PRESSER, A., ELOWITZ, M. B., KELLIS, M. and KISHONY, R. (2008). The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication. *PNAS* **105** 950-954.
- [94] PRZYTYCKA, T. M. (2006). An important connection between network motifs and parsimony models. *Lecture Notes in Computational Biology* 321-335.
- [95] REED, B. A. (2003). *Recent Advances in Algorithms and Combinatorics* Algorithmic aspects of tree-width 85-108. Springer Verlag.

- [96] ROBERTSON, N. and SEYMOUR, P. D. (1983). Graph Minors I : Excluding a forest. *J. Combin. Theory Ser. B* **35** 39-61.
- [97] ROBERTSON, N. and SEYMOUR, P. D. (1984). Graph Minors III : Planar tree-width. *J. Combin. Theory Ser. B* **36** 49-64.
- [98] ROBERTSON, N. and SEYMOUR, P. D. (1986a). Graph Minors II : Algorithmic aspects of tree-width. *J. Algorithms* **7** 309-322.
- [99] ROBERTSON, N. and SEYMOUR, P. D. (1986b). Graph Minors V : Excluding a planar graph. *J. Combin. Theory Ser. B* **41** 92-114.
- [100] SCHBATH, S., LACROIX, V. and SAGOT, M. F. (2009). Assessing the exceptionality of coloured motifs in networks. *EURASIP Journal on Bioinformatics and Systems Biology* ID 616234.
- [101] SCHUSTER, S. and HILGETAG, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **2** 165-182.
- [102] SEYMOUR, P. D. (1995). Packing directed circuits fractionally. *Combinatorica* **15** 281-288.
- [103] STUMPF, M. P. H., WIUF, C. and MAY, R. M. (2006). Subnets of scale-free networks are not scale-free. *PNAS* **103** 7566-7570.
- [104] TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. Institut des Hautes Etudes Scientifiques* **81** 73-205.
- [105] VAZQUEZ, A. (2009). Finding hypergraph communities : a Bayesian approach and variational solution. *Journal of Statistical Mechanics* P07006.
- [106] WASSERMAN, S. and FAUST, K. (1994). *Social network analysis : methods and applications*. Cambridge University Press.
- [107] WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of small-world networks. *Nature* **393** 440-442.
- [108] WERNICKE, S. and RASCHE, F. (2006). FANMOD : a tool for fast network motif detection. *Bioinformatics* **22** 1152-1153.
- [109] WHITE, H. C., BOORMAN, S. A. and BREIGER, R. L. (1976). Social structure from multiple networks I : Blockmodels of roles and positions. *American Journal of Sociology* **81** 730-779.
- [110] WUCHTY, S., OLTVAI, Z. N. and BARABÁSI, A. L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics* **35** 176-179.
- [111] YEUNG, M. and AL., (2007). Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* **8**.
- [112] ZANGHI, H., AMBROISE, C. and MIELE, V. (2008). Fast Online Graph Clustering via Erdős Renyi Mixture. *Pattern Recognition* **41** 3592-3599.
- [113] ZASLAVSKIY, M., BACH, F. and VERT, J. P. (2009). Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics* **25** 259-267.
- [114] ZHANG, L. V., KING, O. D., WONG, S. L., GOLDBERG, D. S., H., T. A., LESAGE, G., ANDREWS, B., BUSSEY, H., BOONE, C. and ROTH, F. P. (2005). Motifs, themes and thematic maps of an integrated *Saccharomyces Cerevisiae* interaction network. *J. Biol.* **4**.

Publications

Revues avec comité de lecture

- [B1] BIRMELÉ, ETIENNE (2003). Tree-width and circumference of graphs. *J. Graph Theory*, **43**, 1, 24–25.
- [B2] MATIAS, CATHERINE ET SCHBATH, SOPHIE ET BIRMELÉ, ETIENNE ET DAUDIN, JEAN-JACQUES ET ROBIN, STÉPHANE (2006). Networks motifs : mean and variance for the count. *Revstat*, **4**, 1, 31–51.
- [B3] BESSY, STÉPHANE ET BIRMELÉ, ETIENNE ET HAVET, FRÉDÉRIC (2006). Arc-chromatic number of digraphs in which every vertex has bounded outdegree or bounded indegree. *J. Graph Theory*, **53**, 4, 315–332.
- [B4] BIRMELÉ, ETIENNE ET BONDY, JOHN ADRIAN ET REED, BRUCE (2007). The Erdős-Posá property for long circuits. *Combinatorica*, **27**, 2, 135–145.
- [B5] BIRMELÉ, ETIENNE (2008). Every longest circuit of a 3-connected, $K_{3,3}$ -minor free graph has a chord. *J. Graph Theory*, **58**, 4, 293–298.
- [B6] BIRMELÉ, ETIENNE (2009). A scale-free graph model based on bipartite graphs. *Discrete Applied Mathematics*, **157**, 10, 2267–2284.
- [B7] BIRMELÉ, ETIENNE ET BONDY, JOHN ADRIAN ET REED, BRUCE (2009). Tree-width of graphs without a 3 by 3 grid minor. *Discrete Applied Mathematics*, **157**, 12, 2577–2596.
- [B8] BIRMELÉ, ETIENNE ET DELBOT, FRANÇOIS ET LAFOREST, CHRISTIAN (2009). Mean analysis of an online algorithm for the vertex cover problem. *Information Processing Letters*, **109**, 9, 436–439.
- [B9] LATOUCHE, PIERRE ET BIRMELÉ, ETIENNE ET AMBROISE, CHRISTOPHE (2011). Assessing a mixture model for graphs with a non asymptotic approximation of the marginal likelihood. *Statistical Modelling*, à paraître, <http://arxiv.org/abs/0912.2873v2>
- [B10] LATOUCHE, PIERRE ET BIRMELÉ, ETIENNE ET AMBROISE, CHRISTOPHE (2011). Overlapping stochastic block model with application to the french political blogosphere. *Annals of Applied Statistics*, **5**, 1, 309–336.

Actes de conférences avec comité de lecture

- [B11] BIRMELÉ, ETIENNE ET BONDY, JOHN ADRIAN ET REED, BRUCE (2007). Brambles, prisms and grids. *Graph Theory in Paris - Proceedings of a Conference in Memory of Claude Berge*, Birkhauser.
- [B12] BIRMELÉ, ETIENNE ET ELATI, MOHAMED ET ROUVEIROL, CÉLINE ET AMBROISE, CHRISTOPHE (2008). Identification of functional modules based on transcriptional regulation structure. *BMC Proceedings*, **2**, 4 :S4.
- [B13] LATOUCHE, PIERRE ET BIRMELÉ, ETIENNE ET AMBROISE, CHRISTOPHE (2009). Bayesian methods for graph clustering. *Advances in Data Analysis, Data Handling and Business Intelligence, GFKL Proceedings*, Springer.
- [B14] BIRMELÉ, ETIENNE (2009). Detecting network motifs by local concentration. *Proceedings des Journées Ouvertes en Biologie, Informatique et Mathématiques 2009*, 91–97.
- [B15] MILREU, PAULO ET ACUNA, VICENTE ET BIRMELÉ, ETIENNE ET CRESCENZI, PIERLUIGI ET MARCHETTI-SPACCAMELA, ALBERTO ET SAGOT, MARIE-FRANCE ET STOUGIE, LEEN ET LA-CROIX, VINCENT (2010). Enumerating chemical organisations in consistent metabolic networks : Complexity and algorithms. *WABI'2010*, **6293**, 226–237.
- [B16] LATOUCHE, PIERRE ET BIRMELÉ, ETIENNE ET AMBROISE, CHRISTOPHE (2010). Uncovering overlapping clusters in biological networks. *Proceedings des Journées Ouvertes en Biologie, Informatique et Mathématiques 2010*.

Travaux soumis

- [B17] BIRMELÉ, ETIENNE Detection of local network motifs. soumis à *Annals of Applied Statistics*, <http://arxiv.org/pdf/1007.1410>.
- [B18] BIRMELÉ, ETIENNE ET DELBOT, FRANÇOIS ET LAFOREST, CHRISTIAN. Average comparison of three algorithms on Erdos-Renyi graphs for the vertex cover problem. soumis à *COCOON 2011*.
- [B19] KOSKAS, MICHEL ET GRASSEAU, GILLES ET BIRMELÉ, ETIENNE ET SCHBATH, SOPHIE ET ROBIN, STÉPHANE. NeMo : Fast Count and Statistical Significance of Network Motifs. soumis à *JOBIM 2011*.

Liste des co-auteurs

Vicente Acuña, CNRS, Lyon.
Christophe Ambroise, Université d'Évry Val d'Essonne, Évry.
Stéphane Bessy, Université Montpellier 2, Montpellier.
Adrian Bondy, retraité.
Pierluigi Crescenzi, Università degli Studi di Firenze, Florence.
Jean-Jacques Daudin, Agro Paris Tech, Paris.
François Delbot, Université de Provence, Marseille.
Mohamed Elati, Université d'Évry Val d'Essonne, Évry.
Gilles Grasseau, Université d'Évry Val d'Essonne, Évry.
Frédéric Havet, INRIA, Sophia-Antipolis.
Michel Koskas, Agro Paris Tech, Paris.
Vincent Lacroix, Université Claude Bernard, Lyon.
Christian Laforest, Université Blaise Pascal, Clermont-Ferrand.
Pierre Latouche, Université d'Évry Val d'Essonne, Évry.
Alberto Marchetti-Spaccamela, Università de la Sapienza, Rome.
Catherine Matias, Université d'Évry Val d'Essonne, Évry.
Paulo Milreu, CNRS, Lyon.
Bruce Reed, Université Mc-Gill, Montréal.
Stéphane Robin, Agro Paris Tech, Paris.
Céline Rouveirol, Université Paris-Nord, Villetaneuse.
Marie-France Sagot, INRIA, Lyon.
Sophie Schbath, INRA, Jouy-en-Josas.
Leen Stougie, Vrije Universiteit, Amsterdam.